Edmondo Trentin
DIISM – Universita' di Siena
trentin@dii.unisi.it

Libero arbitrio e macchine intelligenti

Abstract – L'annosa diatriba tra sostenitori dell'intelligenza artificiale (IA) forte (Searle) e debole (Churchland) è tuttora lungi dal conoscere soluzione. È verosimile che un problema complesso come il provare o il confutare la possibilità di creare macchine dotate di una mente funzionalmente assimilabile a quella umana sia da affrontare con un approccio *divide et impera*, cioè scomponendolo in sotto-problemi meno ardui le cui eventuali soluzioni contribuiscano a formare la soluzione del problema iniziale. Il presente saggio cerca di dare un contributo in tale senso, investigando la questione del libero arbitrio nell'IA. Appurato che un tradizionale sistema a regole condizionali di tipo deterministico non può essere qui applicato con successo, è d'uopo rivolgersi a paradigmi aleatori dell'IA sub-simbolica, fondati sulla teoria bayesiana delle decisioni. Chiedersi se mai sia possibile concepire in tale ambito dei "liberi decisori" equivale, in termini formali, a interrogarsi sulla calcolabilità o meno di un insieme di funzioni che rappresentino ed esprimano il libero arbitrio. La ricerca di una risposta a questo interrogativo passa attraverso il ricorso a quel paradigma di calcolo universale che è la macchina di Turing, e ci porta a trarre conclusioni negative. A partire da queste ultime, proponiamo degli argomenti per dimostrare come esse si estendano, *ipso facto*, a paradigmi di calcolo di diversa natura (macchine non-deterministiche, quantistiche, neurali). Ne emerge la contradittorietà delle posizioni di quanti, ad un tempo, ritengano l'essere umano libero di decidere e la mente "umana" calcolabile. Assumendo l'assenza di trascendenza nel fenomeno "mente", proponiamo infine di contestualizzare i medesimi argomenti, pur per via qualitativa, al caso particolare di quella potente macchina calcolatrice biologica che è il cervello umano, ottenendone conclusioni altrettanto negative. Per dirla con Schopenhauer, un uomo può in definitiva fare ciò che vuole ma non può volere ciò che vuole. Questo ha, naturalmente, una serie di implicazioni ulteriori in ambito etico, giuridico e teologico. Non ultima, l'esistenza del trascendente che conseguirebbe dalla confutazione dell'assioma stesso su cui il nostro argomento si fonda.

1. Introduzione: intelligenze artificiali

L'intelligenza artificiale (IA) è la branca della *computer science* che si occupa dello studio di macchine intelligenti. Per "studio" qui si intende tanto la definizione di un corpus teorico di conoscenze fondamentalmente matematiche (quali la determinazione esatta di aspetti specifici legati al fenomeno intelligenza che siano o meno intrinsecamente realizzabili da una macchina, e la loro eventuale complessità di realizzazione), quanto lo sviluppo di effettivi paradigmi di calcolo (in particolare, algoritmi che specifichino *come* realizzare quei medesimi aspetti). Con "macchine", pur non ponendo limiti a possibili future interpretazioni estensive del termine (i quantum computer, di recente formalizzazione, costituiscono un buon esempio in tal senso), ci si riferisce nella sostanza ai cosiddetti calcolatori digitali o analogici, i primi essendo molto più diffusi e consolidati dei secondi. Assai meno semplice, invece, spiegare in modo univoco ed esaustivo cosa si intenda per "intelligente". La difficoltà non va ascritta tanto all'impaccio tipico degli informatici, quanto alla vaghezza (o, se si preferisce, all'elevato grado di polisemia) che il termine presenta, sia nella lingua parlata che nei gerghi specialistici coniati, di volta in volta, da antropologi, filosofi o psicologi. Senza pretesa di risolvere in questa sede una questione invero abbastanza spinosa, mi limiterò a condividere con il lettore la mia personale concezione in materia. Questo ci permetterà di procedere oltre, facendo saldamente leva su un sistema di coordinate di riferimento tutto sommato sufficientemente chiaro, oltre che comprensivo, *to the best of my knowledge*, di tutte le forme di intelligenze artificiali fino ad oggi sviluppate o osservate dall'uomo. Le coordinate ci vengono offerte dagli psicologi *behavioristi,*

secondo quanto da essi postulato nel saggio *L'intelligenza*[1]. Il bando a cui la scuola comportamentista mette l'introspezione come possibile strumento di indagine dei fenomeni della psiche, ivi inclusa l'intelligenza, torna invero assai utile al nostro caso, non potendosi basare l'analisi dell'eventuale "intelligenza" posseduta da un qualsivoglia artefatto se non sull'esclusiva osservazione, dall'esterno, dei comportamenti esibiti dall'artefatto stesso, essendoci infatti preclusa ogni forma di eterointrospezione. I comportamenti riconducibili alla presenza di una mente intelligente rientrano, secondo i behavioristi, nel novero delle cosiddette "attività intellettuali", tutte rigorosamente riscontrabili sul piano comportamentale, e nella fattispecie: (i) il ragionamento (sillogistico), ovvero la facoltà di trarre conseguenze logicamente giustificate dalla conoscenza di determinati fatti e di regole deduttive di derivazione; (ii) l'induzione, intimamente legata ai fenomeni dell'apprendimento e della generalizzazione (facoltà di maturare modelli generali di porzioni del mondo a partire dall'esperienza contingente); (iii) la sussunzione, attraverso la quale il soggetto riconosce le diverse esperienze sensoriali come appartenenti a categorie o modelli appresi in precedenza; (iv) la capacità di elaborare, in modo originale, delle sequenze di azioni volte alla soluzione di problemi che il soggetto si trovi ad affrontare per la prima volta. Potremo quindi, da qui in avanti, concordare di definire "intelligente" una macchina che esibisca un qualunque comportamento riconducibile, a buon diritto, ad una delle quattro categorie di attività intellettuali testé riassunte. In quest'ottica, le macchine intelligenti non sono dunque un mero oggetto di speculazione filosofica, né tantomeno il soggetto fantastico di qualche opera di fantascienza, bensì un qualcosa di reale, di appartenente (da tempo, invero) alla realtà tecnologica in cui siamo quotidianamente immersi.

Inverosimile, naturalmente, ipotizzare che - per esempio - un riconoscitore ottico di codici di avviamento postale (certamente intelligente, nell'accezione behaviorista) possieda una mente cosciente, a tutto tondo, simile magari a quella umana. Se a oggi la mente artificiale è ancora lungi dall'essere una realtà, la sua realizzabilità resta invece centrale tanto all'informatica quanto, più in generale, alle scienze cognitive. Lasciandoci per un attimo il terreno rassicurante della tecnologia alle spalle, ed avventurandoci un po' più in là, verso i più vasti territori della filosofia, l'IA diventa luogo di sfida aperta tra i sostenitori dell'*IA debole* e dell'*IA forte*. La distinzione è dovuta al filosofo del linguaggio e della mente John Searle[2]. Obiettivo dell'IA debole sarebbe, secondo il Searle, la messa a punto di macchine caratterizzate da comportamenti intelligenti ma prive di una vera e propria *mente* (intesa come capacità di comprensione, ovvero dotata di una semantica). Lo studioso di IA debole sarebbe, anzi, consapevole dell'impossibilità di distillare semantica (cioè, contenuti mentali) da un coagulo di eventi meramente sintattici (i programmi per computer). Nell'IA forte ci si proporrebbe, al contrario, di arrivare a costruire artefatti *intelligenti per davvero;* dotati, insomma, di una mente vera e propria, caratterizzata da contenuti semantici e, verosimilmente, cosciente. Una volta introdotta questa distinzione, Searle propose anche un argomento in forma pseudo-sillogistica (la celeberrima *stanza cinese*[3]) per dimostrare l'assurdità delle pretese dei fautori dell'IA forte. L'argomento fa leva sulla condivisibile "ragionevolezza" dell'esito di un esperimento concettuale, architettato per comprovare allo sperimentatore l'assoluto vuoto di semantica esibito da una macchina intelligente fondata sulla manipolazione di simboli (intesi come gli elementi di un alfabeto finito e discreto, secondo l'accezione invalsa nella teoria dei linguaggi formali) mediante l'applicazione di regole di natura meramente sintattica. L'argomento, ancorché intellettualmente piacevole, risulta però sostenibile (e proprio in questi termini il Seale lo espose) solo a condizione di concordare su di una tesi per nulla scontata, in particolare nell'evo moderno. La tesi è

1 P. Oleron, J. Piaget, B. Inhelder e P. Greco, *L'intelligenza*, trad. di L. Zonta, Einaudi, Torino, 1976.
2 J. Searle, *Minds, Brains, and Programs,* Behavioral and Brain Sciences, 3(3): 417-457, 1980.
3 J. Searle, *La mente e' un programma?*, Le Scienze, 259, 1990.

quella in base alla quale un calcolatore dovrebbe essere tacciato di possedere un'intelligenza nel momento in cui esso riuscisse a superare il test di Turing[4]. Come sappiamo, il test di Turing si propone di stabilire se una macchina sia o meno intelligente in base alla capacità di essa di sostenere conversazioni di qualità linguistiche tali da renderla, al termine di ogni colloquio con occasionali interlocutori (ignari della sua reale natura), indistinguibile da un ipotetico soggetto umano. In senso lato, il test è dunque di matrice behaviorista. Esso rinuncia deliberatamente allo sbirciare dentro la scatola nera (ché il tentativo si rivelerebbe invariabilmente fallace), per focalizzare invece l'attenzione su di un comportamento che – quantomeno in certa misura – sia osservabile dall'esterno.

Al test di Turing è legato, a corda doppia, l'argomento searliano. Di fatto, le conclusioni che si traggono dall'esperimento concettuale su cui questo si fonda sono legittimate solo nel momento in cui si assuma che, per sua stessa definizione, l'IA forte postuli il superamento del test di Turing come condizione sufficiente per comprovare la presenza di una mente raziocinante. Per contro, lo studioso di IA forte potrebbe rigettare questo assunto, ritenendolo arbitrario, e rilevare piuttosto che proprio l'esito dell'esperimento concettuale searliano starebbe a dimostrare l'intrinseca inadeguatezza dello stesso test di Turing quale criterio di verifica. In qualità di contraddittori delle tesi searliane si proposero i Churchland[5], che si cimentarono nel tentativo di smontare l'argomento della stanza cinese attraverso la costruzione di un altro pseudo-sillogismo (la *stanza luminosa*), apparentemente analogo a quello searliano ma fondato su di un diverso esperimento concettuale dall'esito ad esso assimilabile. In modo apparentemente paradossale, così come la stanza cinese sembra implicare l'impossibilità di una mente artificiale, la stanza luminosa sembra negare la fondatezza delle teorie di Maxwell sull'origine elettromagnetica della luce. Teorie che però la fisica moderna ha confermato, invalidando di conseguenza, retroattivamente e *ad absurdum,* l'impiantito pseudo-sillogistico del Searle. Ad un'analisi attenta non sfugge però che il modo di procedere churchlandiano non è affatto foriero dell'auspicata (dai Churchland) confutazione della tesi searliana. Esso fallisce proprio nel suo momento più delicato ed essenziale: infatti, mentre la stanza cinese porta il proprio sperimentatore a trarre le debite conclusioni nel rispetto pieno di un ben preciso e stringente vincolo (il superamento del test di Turing, giustappunto) che è in sé concettualmente garante di quelle stesse conclusioni, la stanza luminosa si fonda su di un esperimento concettuale libero da vincoli, il cui esito non può inficiare dunque - per sua stessa natura - il ragionamento searliano; del quale essa, in definitiva, non rappresenta affatto una controparte esatta e solo diversamente contestualizzata.

Ad ogni buon conto, i Churchland riconobbero a Searle l'avere puntualizzato in maniera incontestabile l'incapacità di fatto dell'IA classica - che fino ad allora era stata prevalentemente simbolica - di mettere a punto menti artificiali. Tale mea culpa churchlandiano s'accompagnò, nondimeno, all'individuazione di una strada da percorrere per superare questo limite passando attraverso modelli di IA sub-simbolica. In essi l'elaborazione dell'informazione può essere compiuta in modo distribuito e parallelo, ad opera di artefatti costruiti a immagine e somiglianza di quella peculiare macchina (biologica) capace di dare luogo ad una mente cosciente: il cervello umano. Le reti neurali artificiali ne costituiscono un'illustre famiglia di esemplari, dotati di meccanismi auto-adattativi di apprendimento blandamente ispirati alla plasticità delle sinapsi. Più recentemente, le neuroscienze computazionali si sono spinte verso la modellizzazione al calcolatore di intere porzioni della corteccia cerebrale, nel tentativo di replicare nel modo più esatto possibile l'attività e le mutue interazioni tra le cellule neurali[6]. Più in generale, è

4 A. Turing, *Computing Machinery and Intelligence*, Mind, 59: 433-460, 1950.
5 P. Churchland e P. Churchland, *Could a Machine Think?*, Scientific American, 262(1): 32-39, 1990.
6 T. Trappenberg, *Fundamentals of Computational Neuroscience,* Oxford University Press, 2010.

utile pensare all'IA sub-simbolica come a una disciplina in cui l'oggetto delle computazioni non siano più i simboli linguistici di un alfabeto finito e discreto, bensì i “segnali” grezzi, quali quelli generati dalle stimolazioni sensoriali del sistema nervoso periferico, rappresentati opportunamente in qualche forma numerica a valori reali. Laddove nell'IA classica i simboli vengono manipolati secondo determinate regole univoche di produzione che, a qualche livello di rappresentazione, associano stringhe di simboli ad altre stringhe di simboli, nell'IA sub-simbolica il segnale è invece trasformato (cioè, *elaborato*) tramite l'applicazione di funzioni matematiche, eventualmente di natura probabilistica anziché deterministica. Forte o debole che la si intenda, l'IA moderna è prevalentemente sub-simbolica, ed è in questo ambito che ha trovato il proprio humus la teoria bayesiana delle decisioni.

2. Macchine che prendono decisioni

Molti problemi dell'IA moderna possono essere espressi come problemi decisionali: la macchina, trovandosi in un dato stato interno S e imbattendosi in determinati eventi esterni $\mathbf{x}$, viene chiamata a *decidere* le prossime azioni da intraprendere. Mi avvarrò di alcuni esempi per esplicitare questo concetto generale.

Esempio 1: navigazione robotica autonoma. Si immagini una piattaforma mobile motorizzata e controllata da un dispositivo di calcolo, impegnata nella navigazione autonoma attraverso gli spazi di un edificio. Il robot deve percorrere corridoi, attraversare porte, evitare ostacoli fissi o in movimento. Ad ogni istante esso deve decidere quale accelerazione imprimere alle proprie ruote al fine di compiere i movimenti appropriati in ogni circostanza. Queste decisioni possono essere meramente conseguenti alle percezioni sensoriali $\mathbf{x}$ contingenti (i sonar o una coppia di telecamere stereoscopiche potrebbero fungere da sensori atti all'uopo), come nel caso di un ostacolo improvviso da evitare. Più in generale, la decisione verrà però presa, oltre che in funzione di $\mathbf{x}$, anche sulla base di una rappresentazione interna, in memoria, dell'edificio (o di quella parte di edificio fino a quel momento esplorata) e di una verosimile destinazione cui la piattaforma deve giungere per completare il proprio *task*.

Esempio 2: espansione di nodi nella ricerca nello spazio degli stati (problem solving). Un modo tradizionale di affrontare algoritmicamente la ricerca della soluzione ad un dato problema passa attraverso la riformulazione del problema stesso nei termini di quello, concettualmente equivalente, di realizzare una progressiva trasformazione di uno stato iniziale di natura in uno stato obiettivo (il *goal*, coincidente con la soluzione del problema originale) mediante l'applicazione, in cascata, di una sequenza ordinata di azioni ammissibili. Queste ultime sono volte univocamente a mutare lo stato iniziale in altrettanti stati di natura intermedi e questi, a loro volta, in altri stati intermedi successivi, e così via fino ad arrivare al goal. La struttura di dati adottata per questo tipo di approcci è nella forma di un grafo di ricerca, i cui nodi sono i possibili stati di natura, la cui radice è lo stato iniziale e i cui archi mettono in relazione ogni stato con quelli che da esso si possono derivare come conseguenza delle possibili azioni. Sebbene, perlomeno in astratto, sia sempre possibile immaginare di sviluppare l'intero grafo di ricerca (eventualmente di profondità infinita) derivante dall'applicazione di tutte le azioni ammissibili in ogni ordine possibile, garantendosi in linea di principio l'individuazione di almeno uno stato *goal* (qualora una soluzione al problema originale effettivamente esista), argomenti tanto di implementabilità concreta quanto di complessità (teorica) in tempo e in spazio richiedono di sviluppare algoritmi che procedano all'espansione[7] dei nodi del grafo in modo

7 Per espansione di uno stato si intende l'esame degli stati ad esso successivi, derivanti dalle singole azioni applicate

parsimonioso e mirato, orientato cioè all'individuazione della soluzione più economica ed efficiente possibile. In sostanza, ad ogni progressiva espansione viene applicata una regola di decisione che sceglie quale nodo successivo esplorare.

Esempio 3: riconoscimento di forme (pattern recognition). Un'ampia famiglia di macchine intelligenti è costituita dai cosiddetti "classificatori", il cui compito è quello di stabilire (ovvero, decidere) con la massima attendibilità possibile la categoria di appartenenza di specifici eventi **x** del mondo reale. In questo ambito sono molto popolari, per esempio, i riconoscitori ottici di caratteri, i classificatori di immagini, i riconoscitori automatici del parlato.

Esempio 4: apprendimento attivo (active learning). Generalmente le macchine dotate di facoltà di apprendimento, quali le reti neurali artificiali, fanno leva su algoritmi supervisionati. Questi operano a partire da una collezione di esempi etichettati manualmente da un istruttore, rappresentativi del tipo di fenomeno che si desidera che la rete apprenda. Le etichette esprimono, in sintesi, la risposta che dalla rete ci si aspetta in output a fronte del corrispondente stimolo in input. Numerosi problemi applicativi di grande attualità (si pensi al dominio del *world-wide web*) richiedono alla macchina di esaminare una mole di dati di volume tale da renderne impossibile, disponendo di risorse finite, l'etichettatura esaustiva. In questi casi si procede all'etichettatura di un sottoinsieme limitato dei dati, affidando alla macchina il compito di sfruttare al meglio l'informazione contenuta nelle etichette disponibili e nei dati non-supervisionati rimanenti (si parla a questo proposito di algoritmi di apprendimento semi-supervisionato). Un modo di procedere in questa direzione è quello dell'*apprendimento attivo*, in cui la macchina seleziona (parsimoniosamente) specifici esempi non etichettati che ritiene essere particolarmente critici o rilevanti per le proprie finalità di apprendimento, richiedendone poi l'etichettatura da parte del proprio istruttore. Il processo, ancora una volta, si fonda in nuce sulla capacità della macchina di decidere, in maniera opportuna e autonoma, quali esempi fare etichettare e quali no.

Formalmente, assumeremo che le decisioni prese dalla macchina siano la conseguenza dell'applicazione di una funzione *phi*(.) avente due argomenti: (i) la rappresentazione **x** della porzione di mondo su cui si deve prendere la decisione; e (ii) lo *stato interno S* della macchina. Scriveremo *phi*(**x**, *S*) = *omega_i* per denotare la particolare decisione *omega_i* presa univocamente dalla macchina tra le possibili scelte alternative *omega_1*, …, *omega_c*. Nell'informatica classica il problema di implementare un programma per calcolatore capace di realizzare una siffatta funzione *phi*(.) verrebbe affrontato codificando il processo decisionale attraverso un sistema di regole deterministiche di tipo condizionale quali:

*if P(***x***,S) then omega_i;*
*if Q(***x***,S) then omega_j;*
*if R(***x***,S) then omega_k;*
else omega_l;

dove *P(.)*, *Q(.)* e *R(.)* rappresentano predicati logici su **x** e su *S*. Anche ad un'analisi superficiale, questo approccio ingenuo basato su una collezione di regole di natura strettamente deterministica appare chiaramente inadeguato. Esso, infatti, richiederebbe al progettista di attuare una codifica rigida ed esaustiva di tutte le decisioni conseguenti al soddisfacimento di altrettanti possibili predicati (condizioni) sul mondo **x** e sullo stato *S*. Si può immaginare di scrivere (abbastanza facilmente, invero) un programma siffatto solo

quel medesimo stato.

laddove si voglia controllare il comportamento di macchine estremamente semplici, dotate di un ventaglio limitato di scelte alternative tra cui dover decidere, operanti in un mondo che sia possibile descrivere efficacemente mediante una casistica parimenti limitata di possibili valori dei rispettivi attributi **x**.

Per sfuggire all'implicito (pre-)determinismo o, quantomeno, per compiere un primo, importante passo in questa direzione, si conviene di abbandonare il formalismo del sistema a regole in favore di un quadro di riferimento stocastico, noto come teoria bayesiana delle decisioni. Il richiamo alla memoria del reverendo Thomas Bayes trova giustificazione nel ruolo centrale rivestito nell'ambito della teoria da una delle relazioni fondamentali tra quantità probabilistiche condizionate e congiunte, relazione scoperta dallo stesso Bayes. Trovano collocazione in questo quadro molti paradigmi attuali dell'IA, quali (sotto opportune condizioni, peraltro abbastanza generali) le reti neurali artificiali e i modelli grafici probabilistici. La cosiddetta regola di decisione *phi*(**x**) di Bayes trova espressione nella seguente doppia implicazione (per concisione scriviamo qui la *phi*(.) come una funzione di un solo argomento, giacché la regola dipende solo dalla particolare descrizione **x** del mondo, pensata come una quantità aleatoria):

$$phi(\mathbf{x}) = omega\_i \leftrightarrow P(omega\_i \mid \mathbf{x}) >= P(omega\_j \mid \mathbf{x})$$

dove *P(omega* | **x***)* denota la verosimiglianza *a posteriori* della scelta *omega* condizionatamente al fatto di avere osservato il manifestarsi dell'evento **x** (nell'universo percettivo della macchina). Questa probabilità a posteriori può essere opportunamente fattorizzata, in virtù del teorema di Bayes, nel seguente modo: *P(omega* | **x***) = p(***x** | *omega) P(omega)/p(***x***).* Nell'esprimere questa relazione si e' adottata una notazione standard, per cui "*P(.)*" denota una probabilità (su uno spazio discreto di eventi), mentre "*p(.)*" rappresenta il valore di una funzione densità di probabilità, valutata su un vettore aleatorio a valori continui. Il teorema traduce una conoscenza "a priori" (o pregiudizio) *P(omega)* sulla distribuzione di probabilità delle diverse scelte *omega* a disposizione della macchina, in una conoscenza più solida e obiettiva, "a posteriori", a seguito dell'osservazione di **x**. La regola di decisione prescrive di optare per quella specifica scelta *omega_i* che risulti la più verosimile, a posteriori, dato **x**.

Nelle situazioni concrete, le quantità probabilistiche in gioco non sono note all'artefatto intelligente, né al suo progettista. Esse vengono piuttosto stimate, algoritmicamente, a partire da campioni raccolti sul campo, ossia da osservazioni empiriche e statisticamente rappresentative dei fenomeni di cui la macchina sta cercando di sviluppare un proprio modello. A differenza dei sistemi a regole esse non sono dunque cablate a priori e in modo rigido da un programmatore, bensì *apprese* spontaneamente dalla macchina. Inoltre, esse non sono di tipo deterministico, bensì di natura aleatoria: ogni decisione viene presa sulla base dello specifico valore delle distribuzioni di probabilità in esame (si riconosce, insomma, un'aleatorietà ai fenomeni presi in considerazione, cioè il mondo **x** e l'implicito stato interno *S*).

Le tre domande fondamentali che a questo punto è legittimo porsi, e alle quali cercheremo di dare risposta, sono le seguenti:

D1) un sistema decisionale siffatto può godere di libero arbitrio?

Ad oggi la definizione di "libero arbitrio" è, apertamente, oggetto di discussione; a fortiori, essa non è certo enunciabile nei rigorosi termini della matematica. Di conseguenza, adotterò qui un approccio in qualche modo complementare, che si rivelerà fruttuoso sul

piano pratico. Piuttosto che cercare di postulare in maniera ineludibilmente forzosa un ipotetico predicato *L(A)* che asserisca la libertà d'arbitrio dell'ente intelligente *A*, ci limiteremo a introdurre una condizione sufficiente (ancorché non necessaria) a garantire l'assenza di libero arbitrio in *A*. Diremo che *A* non possiede libero arbitrio se la decisione da esso presa in un certo istante temporale *t* a fronte dell'osservazione ("stato del mondo") x è determinata, oltre che da x, solamente dallo stato interno *S* di *A* e dalle specifiche leggi che di *A* governano il comportamento. È semplice verificare, come vedremo tra poco, che la risposta a D1 non può che essere negativa.

D2) Può goderne, in generale, una macchina calcolatrice di altra natura?

Il quadro decisionale bayesiano, per quanto utile e consolidato, non è certo l'unico possibile paradigma atto alla definizione di regole di decisione implementabili in un calcolatore. Le branche dell'IA note come *pattern recognition* statistica e *machine learning* hanno visto emergere, nel corso degli ultimi decenni, nuovi ed efficaci sistemi decisionali automatici ad esso alternativi. Non solo: le filosofie sottostanti a queste transizioni paradigmatiche hanno, non di rado, portato (più o meno implicitamente) alla formulazione di teorie della calcolabilità e a macchine calcolatrici alternative a quelle classiche. A titolo di esempio si pensi ai calcolatori quantistici, o a quelli analogici, paralleli e distribuiti che chiamiamo reti neurali artificiali. Qualunque sia la risposta che daremo a D1, è quindi necessario chiedersi se essa continui a valere anche per tali macchine, a dispetto della loro diversa natura, o meno.

D3) In caso negativo, che riflessi potrebbe mai questo avere sul libero arbitrio nell'uomo?

Nelle scienze cognitive vige un continuo processo di reciproca contaminazione tra i risultati conseguiti in aree diverse quali le neuroscienze, la psicologia, l'IA e la filosofia della mente. È lecito chiedersi se (e come) le risposte alle precedenti due domande possano suggerire possibili conclusioni sul fenomeno del libero arbitrio nell'uomo. Mentre le teorie matematiche della calcolabilità e della complessità consentono una trattazione rigorosa di D1 e D2, allo stato dell'arte nessuna teoria scientifica della mente mette a nostra disposizione gli strumenti formali necessari per procedere, parallelamente, alla ricerca di una risposta esatta ed univoca a D3. Nondimeno, il conseguente e necessario sconfinare in ambito prettamente filosofico si rivelerà gravido di fruttuose conseguenze. Le sezioni che seguono cercano di dare risposta alle tre domande fondamentali che ci siamo qui posti.

3. La risposta (negativa) alla prima domanda

Nonostante la regola di decisione bayesiana faccia leva su quantità non deterministiche, esse vengono di fatto trattate come tali solo durante l'apprendimento (ossia durante il processo di stima delle medesime quantità a partire da osservazioni empiriche). L'applicazione della regola in sé è invece deterministica e univoca. È chiaro che questo esclude ogni possibilità di libero arbitrio per le macchine che prendano le proprie decisioni sulla base della massima probabilità a posteriori *P(omega* | **x***)*, laddove questa non tenga conto del cangiante stato interno della macchina e allorquando il suo calcolo abbia luogo sulla base di stime statistiche effettuate a monte del processo decisionale e incapaci, dunque, di adattarsi alle mutevoli esperienze di vita dell'artefatto. Tutto ciò ci porta necessariamente a dare risposta negativa a D1.

È possibile però immaginare delle estensioni realmente non deterministiche della regola di

decisione di Bayes. Una prima estensione potrebbe essere quella di estrarre la decisione *omega* per campionamento dalla distribuzione di probabilità P(*omega* | **x**), anziché definirla deterministicamente come quella *omega_i* per cui la probabilità a posteriori è massima. Così facendo, la macchina prenderà comunque, il più delle volte, la decisione bayesiana *omega_i,* riservandosi però, di tanto in tanto, la "libertà" di decidere diversamente (e casualmente, secondo una distribuzione generalmente non uniforme). A meno di non avere una concezione radicalmente riduzionista del libero arbitrio, è però evidente che questo meccanismo decisionale non ha nulla di libero; né, tutto sommato, di decisionale, dato che si traduce nello scegliere tirando a caso (sebbene si possa ragionevolmente ipotizzare che i processi decisionali nell'uomo non siano del tutto avulsi da questa stessa dinamica aleatoria. Ma all'uomo avremo modo di tornare). Per dirla in termini più formali, e con riferimento alla "definizione complementare" di libero arbitrio formulata in precedenza, questa regola di decisione non può condurre a libero arbitrio in quanto è determinata dalla mera applicazione di una ben precisa legge (di probabilità).

Una seconda estensione del quadro bayesiano, assai più interessante, consiste nel rendere la regola di decisione *phi(.)* adattativa in tempo reale, come avviene nel caso delle reti neurali con apprendimento cosiddetto *online*. In pratica, anziché avere una fase preliminare di stima delle quantità probabilistiche a partire da un campione prefissato di osservazioni, la macchina viene messa nelle condizioni di modificare le leggi aleatorie sottostanti al proprio stesso comportamento, basandosi sull'informazione contenuta nelle nuove esperienze (osservazioni) da essa compiute. Mantenendo la notazione precedente possiamo dunque asserire che, in questa prospettiva, la funzione *phi(.)* e lo stato interno *S* divengono l'una lo specchio dell'altro, e converrà scrivere esplicitamente *phi(***x***, S)* per enfatizzare la dipendenza di ogni decisione tanto dall'immagine sensoriale del mondo **x** all'istante attuale, quanto dallo stato interno (una sorta di memoria storica) della macchina. Nel senso più ampio, la metafora nobile di questa relazione è rappresentata dalla funzione cerebrale nella specie uomo. Alla stregua dell'atto decisionale, in questo quadro anche l'adattamento di *phi(.)* deve avvenire in funzione tanto di **x** che di *S*. In altre parole, le leggi di probabilità che stanno alla base del processo decisionale stesso sono stimate, adattivamente, tanto dall'osservazione del mondo (cioè dalla distribuzione statistica, in esso, degli eventi aleatori sensibili) quanto dal vissuto della macchina (pensato, anch'esso, come fenomeno aleatorio). Una siffatta IA, in sostanza, "sceglierebbe" come raffinare i propri stessi meccanismi di scelta, automodificandosi fino all'istante esatto in cui le necessitasse di prendere una decisione. L'influenza degli stati interni sul come si evolverebbe una siffatta regola decisionale *phi(.)* rende, a questo punto della trattazione, difficile capire se si stiano creando o meno le condizioni per l'insorgere di un libero arbitrio. Le apparenze, sovente ingannevoli, sembrano quantomeno suggerire un cauto possibilismo: è forte la tentazione di lasciarsi sedurre dall'idea, per esempio, di una rete neurale artificiale che, "come accade nel cervello umano", realizzi detta *phi(.)* e che, almeno in potenza, possa portare all'emergere di (una sorta di) libero arbitrio. Per sfuggire all'inganno di questo gioco di specchi è necessario fare pulizia mentale, smettere di pensare in astratto alla macchina nel suo incedere nel mondo, ricondursi agli universali (della matematica). Tutto ciò ci è consentito facendo ricorso alla teoria classica della calcolabilità, ovvero ragionando sulla calcolabilità (o meno) di una siffatta regola automodificante di "libera" decisione, *phi(.)*.

È noto che le macchine calcolatrici digitali fanno una sola cosa: calcolano funzioni. Formalmente, ogni funzione e' *intrinsecamente* calcolabile oppure no. Il merito di questa straordinaria scoperta va prevalentemente a Kurt Goedel, matematico del '900 afferente al Circolo di Vienna, amico di Einstein e che tanta importanza ha avuto (ed ha) per l'informatica. Grazie ad un processo da egli concepito, la *goedelizzazione* (qui riducibile

alla costruzione di una relazione biunivoca tra programmi per calcolatore e numeri interi), è infatti possibile dimostrare un risultato fondamentale della teoria della calcolabilità e parallelo al celebre Teorema di incompletezza, noto come teorema di Cantor (in virtù dello schema cantoriano su cui la sua dimostrazione si fonda). Questo teorema asserisce proprio che, contrariamente all'intuizione comune, esistono (infinite) funzioni che intrinsecamente non sono calcolabili, né lo saranno mai, a prescindere dalla forma e dalla potenza del calcolatore digitale utilizzato. Alla luce di questo risultato appare dunque tutt'altro che ozioso chiedersi se una eventuale regola di decisione *phi(.)* dotata di libero arbitrio si trovi o meno, per sua stessa natura, nella condizione di possedere la proprietà dell'intrinseca calcolabilità. Si tenga presente che un'eventuale risposta positiva a questa domanda non implica affatto che mai si riesca concretamente a scrivere un programma che realizzi *phi(.)*. In definitva, chiedersi se una macchina possa mai godere di libero arbitrio comporta dunque, a rigore, chiedersi se (i) il libero arbitrio sia formalizzabile come una funzione o, più in generale, come un insieme di funzioni (diciamole “regole di libera decisione”), e se (ii) esse siano calcolabili oppure no.

Si può dimostrare che una funzione è calcolabile se esiste una macchina di Turing (MdT) che la calcoli, ovvero se essa è ricorsiva parziale (cioè definibile attraverso la composizione e la ricorsione di determinate funzioni primitive[8]). Per praticità, il paradigma cui nel seguito farò riferimento è dunque proprio quello della MdT[9]. Supponendo che il libero arbitrio *phi(.)* sia calcolabile, esiste quindi una MdT di cui *phi*(.) è la funzione calcolata. Per conferire rigore al nostro argomento, è necessario specificare che una MdT *M* è una quintupla *M=(A, Q, q_0, F, R)* dove *A = {a_1, …, a_n} è* un alfabeto, *Q = {q_1, …, q_m} è* un insieme finito e discreto di possibili stati interni, uno dei quali (*q_0)* è lo stato iniziale; il sottoinsieme *F = {f_1, …, f_j}* di *Q* contiene gli stati finali (o di accettazione), mentre *R = {r_1,..., r_k} è* la collezione delle regole di transizione (il “programma”) che governano il comportamento della MdT (sebbene nell'informatica teorica sia comune definire piuttosto *R* come una funzione parziale di transizione). Una generica regola di transizione *r_i* si presenta nella forma *(x,s) → (x',s',z)* dove *x* e *x'* sono simboli dell'alfabeto *A, s* ed *s'* sono stati interni in *Q*, mentre *z* denota un elemento in *{left, right}*. Il significato di questo formalismo è sostanzialmente il seguente. Si immagina che la MdT sia costituita da una testina di lettura/scrittura che scorre lungo un nastro monodimensionale infinito suddiviso in celle atomiche indivisibili e discrete. Ad ogni istante temporale *t* di un immaginario orologio tempo-discreto di sistema, ogni cella può essere vuota o contenere un (solo) simbolo dell'alfabeto *A.* All'istante *t = 0* la macchina viene posta nel suo stato iniziale *q_0* e la testina viene collocata in corrispondenza della prima cella non vuota del nastro, coincidente con il primo carattere della specifica stringa di input *X = x_1, …, x_u.* Nel corso degli istanti temporali successivi *t = 1, 2, …* si sviluppa una sorta di danza della testina sul nastro, cui corrispondono una manipolazione dei contenuti simbolici delle celle ed un'evoluzione interiore e progressiva dello stato interno della macchina, secondo i dettami delle regole di transizione. Ad ogni rintocco dell'orologio di sistema, infatti, la MdT andrà a cercare una regola del tipo *(x,s) → (x',s',z)* in cui le componenti della coppia *(x,s)* coincidano, rispettivamente, con il simbolo *x* presente nella cella del nastro corrispondente all'attuale posizione della testina, e con lo stato *s* in cui la macchina attualmente si trova. Se una tale regola viene trovata, la conseguenza della sua applicazione è quella di sovrascrivere il nuovo simbolo *x'* al posto di *x* sul nastro, di abbandonare lo stato *s* per entrare nel nuovo stato *s',* e di far compiere alla testina un passo discreto sul nastro verso sinistra (se *z = left*) o verso destra (se *z = right*). Il modello è deterministico se esiste al più una regola per ogni possibile coppia *(x,s)*, non-

8 G. Ausiello, *Complessita' di calcolo delle funzioni*, Bollati-Boringhieri, Torino, 1975.

9 A. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem,* Proceedings of the London Mathematical Society, 42: 230-265, 1936.

deterministico qualora si possano avere più regole applicabili al medesimo simbolo *x* e al medesimo stato interno *s (*in questo secondo caso la MdT sceglie casualmente quale applicare, di volta in volta e a parità di condizioni). La danza può proseguire indefinitamente, se si trovano sempre regole di transizione appropriate da applicare. Se ad un certo istante *t = T* non vi sono invece regole definite per il simbolo e lo stato attuali, la danza ha termine. In questo ultimo caso, qualora la macchina si fermi essendo in uno stato interno di accettazione (cioè appartenente a *F*) la stringa *Y = y_1, …, y_v* di simboli di *A* presente sul nastro definisce l'immagine (output) associata al particolare input *X* dalla funzione parziale *f_M(.)* calcolata dalla MdT *M*, *Y = f_M(X), f_M: A* → A** (essendo *A** il cosiddetto linguaggio universale sull'alfabeto *A*). Il libero decisore *phi(.) è* dunque calcolabile se e solo se esiste (almeno) una MdT *M* per cui si abbia *f_M(.) = phi(.).*

In ogni istante *t* lo stato interno della macchina e la sua memoria (il contenuto del nastro) sono una conseguenza di ciò che essa era all'atto del concepimento (ovvero la sua conoscenza iniziale, come codificata dal progettista in forma di regole di transizione) congiuntamente alla summa di tutte le esperienze sensibili (rappresentate dalla stringa di input) e interiori (rappresentate dall'evoluzione degli stati interni e della memoria) da essa attraversate nel corso del proprio vissuto. La macchina può essere “fotografata” al tempo *t* tramite un'istantanea *xi(t) = (h(t), X(t), s(t))* che la descriva completamente, dove *h(t)* esprime la posizione della testina sul nastro, *X(t)* in *A* è* il contenuto corrente del nastro, e *s(t)* lo specifico stato interno. La MdT prende vita all'istante *t = 0* essendo in *xi(0),* dove *h(0)* coincide con il primo simbolo *(x_1)* della stringa di input *X(0)*, *X(0) = x_1, … x_u*, e *s(0) = q_0.* Viene poi applicata una regola di transizione, passando a *xi(1)*, e così via. Chiamiamo *successore* la trasformazione *succ(xi(t)) = xi(t+1)*, per *t = 0, 1, …,* che associa ad una generica istantanea *xi(t)* l'istantanea successiva *xi(t+1)*. L'estensione allo scenario non-deterministico è immediata ponendo come dominio e codominio della funzione *succ(.)* l'insieme della parti dell'universo delle possibili istantanee, e definendo *succ({xi(t)})* come l'insieme di tutte le istantanee successive a *xi(t)* conseguenti dall'applicazione di tutte le regole di transizione appropriate a *xi(t)*. Il paradigma non-deterministico verrà esplicitamente ripreso nella prossima sezione, mentre per semplicità notazionale mi atterrò da qui in avanti al solo quadro deterministico. Il lettore non faticherà a verificare che il seguito del mio argomento rimane implicitamente valido anche nel caso non-deterministico, una volta che si sia ridefinita adeguatamente la funzione successore.

La *computazione* realizzata da una MdT su input *X* consiste nella corrispondente successione di istantanee *xi(0), …, xi(t-1), xi(t), xi(t+1), …* ottenute per applicazione reiterata della funzione successore, ovvero per composizione del successore con se stesso: *xi(t) = succ(succ(... succ(xi(0)) …)).* In pratica, posta all'istante *t* di fronte ad una qualsiasi scelta, o nella necessità di dover decidere un'azione, la MdT non potrà fare altro che cercare la regola di transizione adatta al proprio attuale stato interno *s(t)* e al contenuto della propria memoria sul nastro *X(t)*, come specificati dall'istantanea attuale *xi(t)*, realizzando un'applicazione della funzione successore. Posta infinite volte in *xi(t)*, in cui è incapsulato tutto il suo vissuto e la rappresentazione codificata del suo universo percettivo, la macchina intraprenderà, reiteratamente e sempre uguale a se stessa, la medesima “azione”, sia questa una risposta motoria, una frase pronunciata, un'alterazione temporanea o permanente della propria memoria o del proprio stato interno. Se si obietta che, se lo fa, lo fa solo perché in passato ha fatto evolvere la propria interiorità *S(.)* “in modo da essere portata a prendere decisioni di un certo tipo” allora bisogna chiedersi *quando*, esattamente, quest'evoluzione autodeterminata possa essersi verificata. Necessariamente in un certo istante precedente *t'*, con *t' < t;* ma in *t'* essa si trovava nell'istantanea *xi(t'),* come conseguenza ineludibile di *succ(succ(… succ(xi(0) …)),* e così via, regredendo fino all'istante iniziale *t = 0.* Un osservatore attento che fosse a

conoscenza del “codice genetico” della specifica macchina, ovvero *(A, Q, q_0, F, R),* delle leggi (regole di transizione) che la governano, nonché della sequenza di tutte le sue esperienze sensibili (la stringa di simboli di input sul nastro), sarebbe in grado di anticipare la specifica decisione presa dalla MdT in un qualsiasi istante temporale futuro *t''*. Gli risulterebbe financo prevedibile, nel dettaglio, tutto quello che la macchina farà nel corso della propria esistenza, incluse tutte le mutazioni interiori che la spingeranno ad esibire comportamenti che solo all'occhio di un osservatore occasionale e meno attento apparirebbero originali e imprevedibili.

In qualsivoglia forma si presenti, insomma, un ipotetico libero decisore *phi(.)* non è dunque calcolabile (né lo sarà mai, a prescindere dalla potenza del calcolatore digitale impiegato e dal livello di sofisticazione del software che lo implementa: ricordiamo, infatti, che la calcolabilità è una proprietà intrinseca delle funzioni). Una macchina intelligente dotata di coscienza di sé potrebbe forse sentirsi libera di decidere, ma questo sentimento non sarebbe nulla più che un genuino autoinganno.

4. La risposta (negativa) alla seconda domanda

Una volta realizzato che nessuna MdT deterministica potrà mai godere di libero arbitrio, c'è da chiedersi (domanda D2) se l'impossibilità non sia dovuta semplicemente alla natura semplicistica di questa famiglia di macchine. Paradigmi di calcolo più sofisticati, ispirati magari alla struttura e alle funzionalità di quella macchina biologica che molti ritengono essere in grado di ingenerare libero arbitrio, il cervello umano, potrebbero rivelarsi capaci di andare oltre i confini esigui entro cui è consentito di spaziare alle MdT tradizionali. Iniziamo con il ribadire che, in quest'ottica, a nulla vale il mero rompere i vincoli imposti dal quadro deterministico. L'abbracciare il non-determinismo, infatti, comporta solamente che l'azione della macchina al tempo *t* può cambiare per effetto dell'aleatorietà, cioè in base al valore casuale e contingente assunto, di volta in volta, da una variabile aleatoria secondo la corrispondente legge (distribuzione) di probabilità. Qualora il lettore non dovesse essere rimasto convinto dall'estensione a MdT non-deterministiche dell'argomento proposto nella sezione precedente - tramite l'opportuna ridefinizione di dominio e codominio della funzione *succ(.)* cui ho ivi accennato - valga qui una prova indiretta inoppugnabile: la teoria della calcolabilità dimostra (in forma di teorema) che la funzione calcolata da una qualsiasi MdT non-deterministica è calcolabile anche tramite appropriate MdT deterministiche. Se *phi(.)*, il libero decisore, fosse calcolabile nel quadro aleatorio, lo sarebbe ipso facto anche nel caso deterministico, il che risulterebbe impietosamente contraddetto da quanto sopra esposto. Mi si permetta anche di rilevare, pure solo in termini qualitativi, che un'entità che prendesse le proprie decisioni a caso (quantunque secondo una ben precisa distribuzione di probabilità, non necessariamente uniforme), ancorché non prevedibile univocamente nel proprio agire da un eventuale osservatore esterno, certo non per questo godrebbe di “libertà di scelta” (a meno di non fare implicito riferimento ad accezioni assai blande e sterili tanto di “libertà” che di “scelta”). In definitiva, dovesse anche la “mente” realizzata da una tale macchina sentirsi libera di scegliere e convinta, a fortiori, di avere preso una libera decisione, essa non avrebbe, di fatto, deciso alcunché.

Più in generale, è una verità dimostrata nell'ambito della teoria assiomatica della calcolabilità che ogni funzione calcolabile mediante una qualsiasi macchina calcolatrice digitale (centralizzata o distribuita, sequenziale o parallela, con uno o più “nastri”, dotata di un linguaggio più o meno ricco e sofisticato, rapida come un lampo o lenta come una lumaca) debba necessariamente esserlo anche tramite una MdT come quella descritta

nella sezione precedente. Anche l'introduzione recente di paradigmi alternativi, quali i calcolatori neurali (come ha osservato Hava Siegelmann[10]) e i calcolatori quantistici, non hanno apportato alterazioni significative al quadro: queste estensioni non inficiano, infatti, i capisaldi della calcolabilità classica, bensì i mutui rapporti di forza tra le diverse classi di complessità di calcolo delle funzioni[11]. In particolare, ogni simulazione software delle reti neurali artificiali implementata su di un calcolatore digitale non potrebbe dunque mai arrivare a calcolare un libero decisore *phi(.)*, perché altrimenti questo ultimo sarebbe, altresì, calcolabile mediante una MdT deterministica; fatto, questo, ancora una volta contraddetto dall'ormai familiare argomento (con buona pace di quanti dovessero qui provare ad avvalersi delle tesi churchlandiane, che vorrebbero la mente emergere come conseguenza del *processo* derivante dall'esecuzione di un programma da parte di un'architettura distribuita e parallela). Per quanto riguarda le macchine analogiche (operanti, cioè, su valori continui), premesso doverosamente che i numeri reali possono essere approssimati con qualunque precisione si desideri attraverso opportune rappresentazioni discrete sul nastro di una MdT, non è comunque verosimile che un eventuale libero arbitrio possa scaturire dal mero passaggio dal dominio discreto a quello continuo. Le più recenti teorie neuroscientifiche assumono, peraltro, fenomeni discreti nel funzionamento del cervello, persino nei meccanismi che si celano dietro la percezione continua dello scorrere del tempo. L'eventuale implementazione di *phi(.)* su di un calcolatore analogico (eventualmente anche tempo-continuo) sarebbe in ogni caso passibile dello stesso ragionamento esposto nella sezione precedente, una volta che si fossero definite le istantanee a valori continui per descrivere lo stato della macchina (ivi incluso quello iniziale), e le regole analogiche di transizione (le funzioni di attivazione realizzate dai neuroni sui rispettivi input). Ancora una volta la funzione calcolata dalla macchina al generico tempo *t* sarebbe l'ineluttabile conseguenza della sua specifica natura e della sequenza di input ricevuti dal mondo esterno a partire dall'istante iniziale *0*.

Uscendo dalla cornice rigida della teoria della calcolabilità e muovendoci verso i lidi ridenti delle neuroscienze computazionali (Churchland docet), che accadrebbe se simulassimo invece l'intero cervello umano, in ogni suo più minuto dettaglio morfologico e funzionale (incluso, incidentalmente e per estensione, l'intero soma; così da tenere conto delle interazioni a carico del sistema nervoso periferico)? Digitale o analogica che fosse, una tale simulazione al calcolatore sarebbe comunque, per definizione, calcolabile: di nuovo, in forza di quanto concluso finora, incapace quindi di generare libero arbitrio. Se mai si arriverà a sviluppare una mente artificiale (o, quantomeno, a dimostrarne in potenza la calcolabilità), questa mente (ancorché autocosciente) non potrà mai possedere libero arbitrio. È mia opinione che questo risultato ponga un punto fermo significativo nell'ambito delle diatribe sull'IA. Spingendosi ancora oltre, la simulabilità del soma e la calcolabilità della mente umana proverebbero, a fortiori, il sussistere di una (e una sola) delle due condizioni seguenti: (1) l'assenza di libero arbitrio anche nell'essere umano (quantomeno nel soggetto il cui cervello è oggetto della simulazione perfetta al calcolatore); (2) la presenza di libero arbitrio nell'uomo, spiegabile però solo con il necessario concorso di fenomeni trascendenti. Questa speculazione ci porta, nella prossima sezione, a cercare risposta alla domanda D3.

5. La terza domanda: quali implicazioni ha tutto questo sull'essere umano, ammesso che ne abbia?

Molti studiosi di scienze cognitive e IA (forse la maggior parte, anche se pare non esistano

10 H.T. Siegelmann, *Computation Beyond the Turing Limit*, Science, 238(28): 632-637, 1995.
11 E. Bernstein e U. Vazirani, *Quantum Complexity Theory*, SIAM Journal on Computing, 26(5): 1411-1424, 1997.

dati statistici precisi sulla loro distribuzione) ritengono che l'essere umano goda di libero arbitrio e che, al tempo stesso, sia possibile realizzare una mente artificiale funzionalmente analoga a quella umana (dotata, cioè, delle medesime prerogative, coscienza di sé e libero arbitrio inclusi). La chiusa della sezione precedente mostra che questa posizione è di fatto autocontraddittoria, a meno di non ipotizzare futuribili paradigmi di calcolo che sfuggano i limiti della teoria classica della calcolabilità (la qual cosa, a questo punto della storia, appare quantomeno poco verosimile se si considera l'universalità della nozione di calcolabilità come proprietà intrinseca delle funzioni). Gli argomenti portati fino a questo punto all'attenzione del lettore dovrebbero essere sufficienti a convincerlo che, se si rigettano spiegazioni metafisiche, sono ammissibili solo due scenari disgiunti e complementari: (i) l'essere umano gode, come l'esperienza soggettiva comune vorrebbe, di libero arbitrio, e nessuna macchina esibirà dunque mai una mente “umana” a tutto tondo (IA debole); (ii) la mente “umana” e' un fenomeno calcolabile (in altre parole, hanno ragione i cultori dell'IA forte) e proprio per questo non vi è nell'animale uomo alcun libero arbitrio. *Tertium non datur*.

Non vi è, nell'esposizione fatta finora, motivo di deporre in favore dell'uno o dell'altro dei due scenari. È tuttavia possibile, ed è a mio avviso molto interessante il farlo, cercare di ripercorrere il ragionamento già sviluppato per le MdT contestualizzandolo alla macchina (biologica) che chiamiamo cervello e che, da un certo punto di vista, sembra stare lì a darci conferma della calcolabilità della mente in virtù della Tesi di Church (argomento, quest'ultimo, usato dai Churchland a sostegno dell'IA forte).

Ciò che un determinato cervello fa all'istante t è, come nella MdT, una conseguenza necessaria e ineluttabile di tre fattori: (1) le leggi (semplici) di natura chimica e fisica che ne governano la funzione (le sue “regole di transizione”); (2) la sua anatomia al tempo t, data dall'esatta topologia, forza e chimica delle sue connessioni sinaptiche e dei suoi alberi dendritici; (3) lo specifico e completo pattern di attivazione di tutti i suoi neuroni all'istante t (ovvero il suo “stato interno”). Quel cervello è il frutto di come esso era a livello embrionale (espressione del codice genetico), cioè all'istante iniziale 0; e di come esso, nel tempo, si è adattato progressivamente e plasticamente al proprio vissuto (sensibile e interiore). Se il soggetto dotato di quel particolare cervello si trova, al tempo t, nell'atto di prendere una decisione (e pur percependosi, costui, libero di prenderla), nel suo sistema nervoso centrale si produrranno di fatto solo scariche elettriche e fenomeni chimici all'altezza delle membrane pre- e post-sinaptiche che, secondo le specifiche leggi chimiche e fisiche sottostanti, determineranno ineluttabilmente quali altri neuroni debbano attivarsi all'istante successivo, e così via allo scorrere del tempo. Seguendo il medesimo ragionamento esposto a proposito dei calcolatori automodificantisi, laddove si controbatteva all'obiezione secondo cui le leggi applicate al tempo t non fossero predeterminate ma (anche) conseguenza delle elaborazioni fatte e delle decisioni prese dalla macchina in qualche tempo passato t', con $t' < t$, anche a proposito del cervello è evidente che qualsivoglia elaborazione interiore o decisione pregressa risalenti all'istante t' non siano state “libere”, a loro volta, bensì dettate dai medesimi fattori (1), (2) e (3) testè enunciati; e così via, regredendo fino all'istante $t = 0$ del concepimento. Se, ancora una volta, si esclude dai processi mentali propri dell'essere umano ogni traccia di trascendenza e si assume, dunque, che l'attività del sistema nervoso centrale sia la responsabile unica dei fenomeni mentali (ipotesi condivisa quasi unanimemente dai neuroscienziati e dagli scienziati cognitivi), nemmeno per il cervello (ovvero, per estensione, per la macchina biologica che chiamiamo *homo sapiens*) né per la mente che ne scaturisce vi è tempo né luogo per l'insorgere e l'esprimersi di alcun libero arbitrio (nell'accezione del termine specificata nella sezione 2). L'eventuale aleatorietà di alcune delle leggi che sottostanno a questo automodificarsi dello stato interno del cervello, come

già nel caso delle MdT non-deterministiche, non può certo essere intesa, in sé, come luogo di espressione di un libero arbitrio; così come quest'ultimo non può essere imputato al fatto che nella realtà fisica il tempo diventi (meglio: possa diventare) una variabile a valori continui. L'argomento non verrebbe inficiato neppure se si volesse considerare la presenza dell'intero soma (il sistema nervoso periferico) o, addirittura, dell'intero universo con le proprie leggi (eventualmente aleatorie). “ È certo”, non resta insomma che concludere con Arthur Schopenhauer, “che un uomo può fare ciò che vuole, ma non può volere ciò che vuole”.

6. Conclusione: alcune implicazioni ulteriori

Non di rado l'annosa diatriba tra IA forte e IA debole ha spinto i rispettivi sostenitori a proporre argomenti di tipo filosofico in favore o contro la possibilità di sviluppare macchine dotate di una mente artificiale comprensiva delle medesime funzionalità di quella umana. Per essere posta nella giusta prospettiva la questione necessita, alla pari di altre e similari questioni aperte riguardanti l'IA, di un'adeguata riformulazione in seno alla teoria della calcolabilità, chiedendosi: la mente è calcolabile? Lo sono, quantomeno, alcune delle funzioni fondamentali che la compongono, tra cui la coscienza o il libero arbitrio? Il presente saggio ha cercato di dare un contributo metodologico in tale senso, procedendo nell'affrontare con sistematicità (ancorché in maniera estremamente semplice) una specifica questione, quella della calcolabilità del libero arbitrio, solo apparentemente troppo astratta e lontana dallo stato dell'arte della scienza dei calcolatori per consentirci di trarre conclusioni di sorta. La conclusione centrale cui siamo giunti è l'intrinseca impossibilità di sviluppare calcolatori digitali dotati di libero arbitrio. È legittimo chiedersi se questa conclusione abbia, a sua volta, implicazioni negative sulla calcolabilità dell'intelligenza “a tutto tondo”, ovverossia della mente. Iniziamo con il ribadire che se mai si dovesse arrivare a implementare una mente artificiale, autocosciente e funzionalmente simile a quella umana, essa sarebbe forse indotta dalla propria stessa introspezione a ritenersi libera di scegliere, mentre, nei fatti, essa non potrebbe comunque esserlo, mettendo peraltro così a nudo uno dei tratti più autoingannevoli del fenomeno della (cosiddetta) coscienza di sé. Ne segue che chi non ritenga lecito il parlare di “mente” senza che in essa vi sia l'imprescindibile presenza del libero arbitrio, non possa che trarre da questo saggio conclusioni drasticamente negative sugli obiettivi ultimi dell'IA forte. Rimangono invece immutate le speranze in tale senso per quanti, piuttosto, siano aperti a contemplare, se necessario e in via per lo meno ipotetica, l'assenza di libero arbitrio anche nella nostra specie. Come abbiamo avuto modo di rilevare, sarebbe invece autocontraddittorio il credere al libero arbitrio nell'*homo sapiens* e, al contempo, il propugnare la possibilità di una mente artificiale “come quella umana”.

L'impossibilità per una macchina calcolatrice di possedere libero arbitrio ha evidenti e profonde implicazioni nel campo di una disciplina di recente fondazione, la *machine ethics*[12], che con la roboetica[13] concorre a formare quella branca dell'etica nota come “etica dell'intelligenza artificiale”. Nella *machine ethics* ci si (pre)occupa dei comportamenti etici esibiti da artefatti dotati di IA, tanto nei riguardi dell'uomo quanto in quelli di altri artefatti pensanti. Gli argomenti che ho esposto finiscono inevitabilmente con lo spogliare le macchine intelligenti di ogni responsabilità etica e morale (essendo esse permanentemente impossibilitate a prendere decisioni libere sul proprio agire) e con il riversarla preponderantemente sull'uomo, loro creatore, tanto nella veste di progettista che

12 W. Wallach e C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, Oxford, 2008.

13 G. Veruggio, *The EURON Roboethics Roadmap*, Proceedings of Humanoids 06, 612-617, 2006.

di istruttore (assumendo, quantomeno, una fase preliminare di apprendimento supervisionato). Sebbene il trarre conclusioni definitive lungo questa direzione sia prematuro e al di là degli obiettivi di questo saggio, appare tuttavia immediato il concludere come, conseguentemente, l'intera *machine ethics* finisca con il rientrare implicitamente sotto la giurisdizione della cugina roboetica, la quale si occupa per l'appunto di condotta etica da parte degli esseri umani impegnati nello sviluppo di robot e di artefatti intelligenti in genere.

Proseguendo lungo questa linea di pensiero e contestualizzandola alla simulazione digitale del sistema nervoso umano e, ancora oltre, alla macchina biologica “cervello”, si intuisce immediatamente quali implicazioni si prospettino sulla nostra stessa capacità decisionale, sulla nostra libertà di pensiero, di sentimento e di azione. In assenza di entità o fenomeni trascendenti, infatti, la mente risulta essere conseguenza unica dell'attività del cervello: le leggi fisiche dell'universo e quelle chimiche che caratterizzano la propagazione dei neurotrasmettitori determinano ineludibilmente ogni nostra scelta, prevenendo ogni forma di libero arbitrio. In questo quadro, etica e morale vanno ripensate, non potendosi più ritenere il soggetto il titolare di meriti o colpe, né il vero responsabile delle proprie azioni. Oltre che a fattori genetici, queste ultime sono imputabili piuttosto all'ambiente esperito e introiettato dal soggetto nel corso del proprio vissuto. Tra i fattori ambientali un ruolo cruciale è quello svolto dalle interazioni con altri individui (quali gli “educatori”), essi stessi a loro volta incapaci di scegliere liberamente e dunque, in ultima analisi, anch'essi non tacciabili di colpe alcune. Questo, in prospettiva, finisce con l'avere anche implicazioni sulla giurisprudenza. È infatti necessario rivedere il concetto stesso di colpa, e dunque di pena, e ricondurre l'idea di legge (nell'accezione giuridica del termine) a ciò che nella realtà dei fatti è: non già “giustizia”, ma un sistema solamente pragmatico di regole (auto-)determinate dalla società per motivi esclusivamente funzionali alla salute della medesima. Infine, non si possono trascurare le implicazioni di carattere religioso. A questo proposito basti ricordare che nel *De libero arbitrio* Agostino di Ippona afferma il libero arbitrio, che risulta anzi necessario per spiegare le origini del Male ed essenziale alla concezione teologica agostiniana (che il cattolicesimo ha fatto propria). In assenza di libero arbitrio, infatti, il concetto stesso di colpa vacilla e, con esso, anche quello corrispondente di peccato.

C'è un aspetto fondamentale che non possiamo permetterci di perdere di vista. Tanto gli argomenti portati nella sezione precedente, quanto le implicazioni testé sommariamente passate in rassegna, sono naturalmente - e fino a prova contraria - destituiti di ogni ragionevole fondamento nella misura in cui non si voglia escludere dall'equazione ogni eventuale variabile di natura trascendente. L'argomento qui esposto è stato infatti concepito per essere il più vicino possibile ad una piccola teoria assiomatica, in cui l'assenza di fattori metafisici è annoverata esplicitamente tra gli assiomi. Ogni teoria assiomatica in seno alla quale si arrivi a dimostrare, per via deduttivamente giustificata, un enunciato contraddetto dall'evidenza empirica, costituisce in sé prova irrefutabile dell'infondatezza di almeno uno dei propri assiomi. Ne consegue che chi dovesse, per via introspettiva o per altra via, ritenere il libero arbitrio essere in sé presente e autoevidente, potrà (e dovrà) fare leva sui nostri argomenti e considerare comprovata, alla luce di quanto appena sottolineato, l'esistenza del trascendente oltre il sensibile.