INTELLIGENZA ARTIFICIALE (DRAFT)

I. Che cos'è l'intelligenza artificiale - II. Che cosa fa l'intelligenza artificiale - III. Che cosa l'intelligenza artificiale non può fare - IV. L'intelligenza artificiale fra scienza e coscienza.

I. Che cos'è l'intelligenza artificiale

1. Introduzione. Si può definire l'«intelligenza artificiale», che qui abbreviamo in IA (si usa spesso anche "AI", acronimo dell'inglese *Artificial Intelligence*), come l'«insieme di studi e tecniche che tendono alla realizzazione di macchine, specialmente calcolatori elettronici, in grado di risolvere problemi e di riprodurre attività proprie dell'intelligenza umana» (T. De Mauro, *Grande dizionario italiano dell'uso*, Torino 2000). La locuzione "intelligenza artificiale" è un evidente ossimoro, in quanto attribuisce all'"artificiale" qualcosa che è essenzialmente "naturale" in quanto è la prerogativa più gelosa della natura umana: l'intelligenza. E l'ossimoro è piuttosto provocatorio, poiché c'è chi molto seriamente si domanda se la macchina possa essere davvero "intelligente", nel senso in cui questo termine è attribuito alla mente dell'uomo (vedi *infra,* IV). D'altra parte definizioni come questa appaiono troppo generiche, giacché si adattano altrettanto bene a tutta l'informatica e, per esempio, alle "tecniche di automazione": discipline che non fanno parte dell'IA. Nella prossima sezione, in cui si cercherà di delineare una traccia storica, si vedranno meglio questi apparentamenti. Ma è obiettivamente difficile dare definizioni più precise, sia perché si tratta di una materia in forte evoluzione, sicché una definizione che delimitasse il territorio di sua competenza rischierebbe di escludere *a priori* sviluppi futuri che naturalmente le potrebbero appartenere, sia perché essa è contemporaneamente una scienza e una tecnica, ed è una disciplina di frontiera, una specie di affascinante "punto multiplo" in cui s'incontrano diversi domini del sapere: logica, informatica, psicologia, neuroscienze, filosofia. Per cui, piuttosto che delimitare, si preferisce elencare e descrivere le sue caratteristiche fondamentali e le sue principali aree d'applicazione. Tuttavia si è tentato di precisare, ma le precisazioni sono discordanti fra loro sicché ne sono risultate definizioni alquanto diverse. Ne vediamo qualcuna. Russell e Norvig (1998) propongono due distinzioni fondamentali. La prima è fra macchine che "pensano" e macchine che si limitano a "operare" in maniera, in qualche misura, simile a quella degli esseri umani. La seconda riguarda il termine di confronto per valutare le loro prestazioni, termine che può essere l'uomo reale o una sua idealizzata razionalità (cfr. *Intelligenza artificiale*, p. 4). Si deve dire che le applicazioni tecniche (vedi *infra,* II) riguardano principalmente l'"operare razionale", mentre il dibattito filosofico, che sarà oggetto proprio dell'ultima sezione, insiste sulla possibilità di prestazioni "umane" e soprattutto di un "pensiero umano". Un'altra distinzione fondamentale, sulla quale il dibattito filosofico è acceso, è fra le cosiddette «IA debole» e «IA forte»: chi sostiene l'IA debole si accontenta di considerare macchine che agiscono "come se" fossero

intelligenti; l'IA forte, invece, asserisce la possibilità di macchine simili all'uomo fino a possedere un'autocoscienza. S'intende facilmente come queste distinzioni s'intreccino fra loro: in particolare l'IA forte riguarda esclusivamente "macchine che pensano in maniera umana", mentre l'IA debole concerne, di preferenza, "macchine che operano".

Infine, sul piano della realizzazione, si può distinguere fra un'impostazione "funzionale" o "comportamentistica", per la quale non importa quale struttura abbia l'elaboratore nel quale risiede l'"intelligenza", e un'impostazione "strutturale" o "costruttivistica" o "connessionistica" che vuole ottenere le stesse prestazioni del cervello umano riproducendo, in qualche modo, la sua struttura. Con un leggero spostamento di prospettiva, la prima impostazione è stata chiamata "emulazionistica" e la seconda "simulazionistica": i sostenitori di quest'ultima ritengono che solo riproducendo il più fedelmente possibile il cervello si possono ottenere prestazioni ad esso paragonabili; chi sostiene la prima, al contrario, è convinto che l'essenza del funzionamento del cervello non risieda nella sua struttura ma nelle sue prestazioni, e che queste possano essere ottenute anche, e magari meglio, da strutture anche completamente diverse. Entrambe le impostazioni sono fertili di risultati ma la seconda, benché minoritaria, ha una speciale importanza perché ha condotto alla realizzazione delle «reti neurali». Queste sono imitazioni del cervello animale, estremamente rozze, ma di grande interesse conoscitivo oltre che tecnico, perché stabiliscono un rapporto con le neuroscienze di grande utilità sia per l'IA sia per le neuroscienze stesse. In pratica, le due impostazioni convergono perché dopo qualche tentativo di produrre strutture fisiche "dedicate" (cioè a livello di *hardware*), ora le reti neurali sono realizzate piuttosto con programmi di calcolo (cioè a livello di *software*) che vengono eseguiti su calcolatori di tipo generico.

2. Un po' di storia. L'idea di delegare a congegni "meccanici" talune operazioni tipiche della mente è molto antica. Per esempio basti pensare alle operazioni aritmetiche svolte con l'abaco, probabilmente inventato dai Cinesi attorno al 5000 a.C.; oppure al controllo automatico: per esempio, se si vuole procedere con auto ad una certa velocità, si regola la pressione sul pedale dell'acceleratore in funzione dell'indicazione del tachimetro. Anche in questo caso il primo dispositivo realizzato di proposito, che si conosca, è dovuto ai Cinesi, e serviva a regolare il livello dell'acqua nelle risaie: un galleggiante muoveva una paratoia, che riduceva la portata d'acqua quando il livello tendeva ad aumentare. In epoca moderna, le prime macchine da calcolo sono dovute a (↗) Pascal (1623-1662) che, nella metà del Seicento, costruì una addizionatrice meccanica (la *pascaline*), e a (↗) Leibniz (1646-1716) che, alla fine del secolo, la perfezionò per consentire le moltiplicazioni e le divisioni. La prima macchina "programmabile", ossia capace di eseguire automaticamente sequenze di operazioni, fu concepita da Charles Babbage (1792-1871) attorno al 1830, ma non fu mai costruita a causa delle difficoltà meccaniche. Sul versante dell'automatica si deve menzionare il regolatore di velocità di James Watt (1736-1819), che a metà del Settecento aprì la via all'automazione industriale. Queste notizie documentano l'interesse a trasferire alle macchine non solo il lavoro

materiale — quello che comporta dispendio di energie fisiche — ma anche la fatica intellettuale richiesta, vuoi per eseguire tediose sequenze di calcolo, vuoi per sorvegliare e controllare il corretto funzionamento di altri dispositivi (l'automazione è stata definita con la formula «macchine che controllano altre macchine»). Ma gli sviluppi che più specificamente interessano la nascita dell'IA avvengono attorno alla metà del Novecento. Si devono ad Alan Turing (1912-1954) due contributi fondamentali. Nel 1936 egli propose un modello ideale di calcolatore automatico "universale" (conosciuto, appunto, come «macchina di Turing»): esso è il prototipo di tutti gli elaboratori elettronici, che poi furono sviluppati a partire dalla metà degli anni Quaranta. Nel 1950, poi, Turing propose il «gioco dell'imitazione», ossia un paradigma per stabilire se una macchina è "intelligente". Nel un suo noto articolo *Computing Machinery and Intelligence* (1950), egli suggeriva di porre un osservatore di fronte a due telescriventi. Una delle due è comandata da un uomo, l'altra da una donna. L'osservatore, che non sa a quale terminale corrisponda l'uomo e a quale la donna, lo può accertare ponendo loro qualunque tipo di domanda. Uno dei due interlocutori deve rispondere con sincerità, l'altro invece deve fingere d'essere dell'altro sesso. Poi all'interlocutore mendace si sostituisce un calcolatore programmato in modo da "fingere" di essere una persona umana. Quando il numero d'errori nell'identificare il calcolatore sarà uguale a quello nell'identificare l'interlocutore mendace, allora si potrà dire che il calcolatore è "intelligente". Va detto che il gioco dell'imitazione è stato tentato più volte con risultati, finora, piuttosto deludenti. Nella classificazione di Russell e Norvig, l'esperimento concettuale di Turing fornisce un valido esempio di "macchina che si comporta in maniera umana"; esso rappresenta una posizione "comportamentista", per la quale Turing è stato criticato da chi riteneva che l'IA dovesse essere qualcosa di più forte.

Per la verità, il primo lavoro che si usa ascrivere all'intelligenza artificiale risale al 1943, quando Warren McCulloch (1899-1969) e Walter Pitt progettarono una rete neurale. Ma gli sviluppi più importanti — sia sul piano teorico sia quanto alla redazione di programmi di calcolo, che hanno valore di prototipo per le esperienze successive — sono da ascrivere al decennio posteriore alla provocazione di Turing. In particolare, nel 1956 un altro dei pionieri, John McCarthy (***-***) riunì a Dartmouth i principali studiosi del tempo (fra i quali Marvin Minsky, Allen Newell, Claude Shannon e Herbert Simon) in un seminario in cui, fra l'altro, egli propose il nome di «intelligenza artificiale». Il 1958 fu particolarmente fertile di risultati: McCarthy produsse il *Lisp*, un linguaggio di programmazione ad alto livello dedicato specialmente all'IA (poi seguito, nel 1973, dal *Prolog*), e incominciò a sviluppare programmi generali per la soluzione di problemi. S'incominciò anche a studiare quelli che oggi sono chiamati «algoritmi genetici», ossia programmi capaci di modificarsi automaticamente in modo da migliorare le proprie prestazioni. Nei decenni successivi la ricerca proseguì con alterne vicende. Gli anni Sessanta furono caratterizzati da risultati forse non eccezionali se valutati col metro odierno, ma allora entusiasmanti, sia a causa della limitatezza degli strumenti di calcolo con i quali erano ottenuti, sia perché sistematicamente smentivano gli scettici che sostenevano che «la tal cosa non si potrà mai fare». In quegli anni si registrarono

anche interessanti sviluppi, soprattutto teorici, della ricerca sulle reti neurali. Ma si incontrarono anche le prime difficoltà, e si fu costretti a prendere coscienza di limiti, che ancor oggi appaiono insuperabili. Una grave difficoltà è l'«esplosione combinatoria», ossia l'aumento esplosivo del tempo di calcolo quando aumenta il numero di variabili del problema; un limite è il fatto, sul quale si dovrà tornare, che l'elaboratore può trattare soltanto i legami "sintattici" e non i contenuti "semantici", ossia il significato delle variabili sulle quali sta operando. Gli anni Settanta videro la nascita dei «sistemi esperti» e le loro prime applicazioni alla diagnostica medica, e i primi tentativi di "comprensione" del linguaggio naturale (nel senso restrittivo di dare risposte preordinate a un limitato numero di domande).

A partire dal 1980 l'IA è uscita dai laboratori scientifici e ha trovato applicazioni pratiche significative, varie delle quali saranno descritte nella prossima sezione. Contemporaneamente, e per conseguenza, aziende industriali specialmente americane e giapponesi hanno incominciato a mettere in commercio programmi dedicati ai sistemi esperti, al riconoscimento di configurazioni e così via, ed hanno costruito microcircuiti ed interi elaboratori specializzati per applicazioni dell'IA. Le reti neurali, dopo poco meno di vent'anni di quasi completo disinteresse, hanno ricevuto nuova attenzione a partire dal 1985, in particolare a causa della definizione di nuovi, più potenti algoritmi di ottimizzazione. Nell'ultimo decennio del secolo, al perfezionamento delle reti neurali si è affiancato lo sviluppo di nuovi procedimenti di calcolo, soprattutto derivati dalla teoria delle probabilità e delle decisioni; e, sul versante delle applicazioni, sono stati sviluppati metodi efficaci per la costruzione dei sistemi esperti e per il riconoscimento del parlato e delle forme, questi ultimi specialmente destinati alla robotica e alla visione artificiale.

II. Che cosa fa l'intelligenza artificiale

In questa sezione vedremo gli aspetti più propriamente tecnici dell'IA, mentre la prossima sarà dedicata a questioni più propriamente filosofiche. Dal punto di vista "ingegneristico" e riduttivamente pragmatico, l'IA è valutata semplicemente per le sue capacità e prestazioni, indipendentemente dai metodi e meccanismi che sono utilizzati per realizzarla. Il punto di vista è dunque "emulazionistico" e non "simulazionistico": l'idea che sta alla base è quella di costruire macchine che non necessariamente "simulino" riproducendo il comportamento del cervello umano, ma siano più semplicemente in grado di "emularlo", selettivamente, nel risultato finale di certe operazioni. È questa la tesi sostenuta da A. Turing nel "gioco dell'imitazione" che abbiamo già descritto: egli propone di "valutare" l'intelligenza di una macchina solamente dalla sua capacità di presentare un comportamento comunicativo indistinguibile da un essere parlante umano. Quest'impostazione è stata sicuramente dominante nella storia dell'IA ed ha portato alla costruzione programmi che raggiungono un alto livello di competenza nella conoscenza e nella risoluzione di problemi ritenuti

complessi. Tali programmi sono costruiti come "manipolatori" di simboli formali non-interpretati, per cui la macchina può essere concepita semplicemente come un trasformatore sintattico senza alcuna conoscenza "semantica" del problema (vedi *infra,* III.4).

1. Architettura di base dei sistemi di intelligenza artificiale. L'applicazione *software* alla base di un sistema di IA non è un insieme di istruzioni immutabili che rappresentano la soluzione di un problema, ma un "ambiente" in cui rappresentare, utilizzare e modificare una base di conoscenza. Il sistema esamina un largo numero di possibilità e costruisce dinamicamente una soluzione. Ogni sistema di tal genere deve riuscire ad esprimere due tipi di conoscenza in modo separato e modulare: una «base di conoscenza» e «un motore inferenziale». Per «base di conoscenza» si intende il "modulo" che raccoglie la conoscenza sul "dominio", cioè sul problema. È possibile dettagliare la base di conoscenza suddividendola in due sezioni: a) Il blocco delle asserzioni o fatti (memoria temporanea o a breve termine), b) il blocco delle relazioni e regole (memoria a lungo termine). La memoria temporanea contiene la "conoscenza dichiarativa" su di un particolare problema da risolvere. Si ha una rappresentazione costituita da fatti veri introdotti all'inizio della consultazione o dimostrati veri dal sistema nel corso della sessione di lavoro. Nella memoria a lungo termine vengono invece mantenute regole che forniscono un insieme di raccomandazioni, consigli, direttive strategiche atti a costruire il patrimonio di conoscenza disponibile per risolvere il problema. Le regole sono costituite attraverso dichiarazioni composte di due unità. La prima è detta «antecedente» ed esprime una situazione od una premessa, mentre la seconda è chiamata «conseguente», e in essa viene avviata l'azione da applicarsi in caso ci sia riscontro di verità nella premessa. La sintassi generale è pertanto: «se "antecedente" allora "conseguente"». Il «motore inferenziale» è il modulo che utilizza la base di conoscenza per giungere alla soluzione del problema proposto e per fornire spiegazioni. Al motore inferenziale è delegata la scelta di quale conoscenza è opportuno utilizzare, istante dopo istante, nel processo risolutivo. Dunque vengono combinate le varie cellule di conoscenza, che considerate singolarmente, apparirebbero di uso limitato, al fine di trarre nuove conclusioni ed esprimere nuovi fatti. Ciascuna regola dell'insieme che rappresenta il dominio della conoscenza, per risultare valida in una particolare istanza, dev'essere confrontata con un insieme di fatti che rappresentano la conoscenza attuale sul caso corrente, e quindi soddisfatta. Ciò viene fatto attraverso un'operazione di *matching* (raccordo), in cui si tenta di accostare l'antecedente della regola con i vari fatti presenti nella memoria temporanea. Se il *matching* ha successo, si procede all'esecuzione delle eventuali azioni elencate nel conseguente. Nel caso quest'ultimo contenga invece una conclusione, il soddisfacimento dell'antecedente consente di omologare tale asserzione come nuovo fatto della memoria a breve termine. L'operazione di *matching* genera delle «catene inferenziali», che indicano il modo in cui il sistema utilizza le regole per effettuare nuove inferenze. Permettono inoltre di dare all'utente una spiegazione su come vengono emesse certe conclusioni. Per generare delle catene inferenziali a partire da un insieme di regole i metodi adoperati sono sostanzialmente due: a) il «concatenamento in avanti» (*forward*

chaining). Tale tecnica cerca di giungere ad una conclusione partendo dai fatti presenti all'inizio nella memoria temporanea e applicando in avanti le regole di produzione. Si dice che l'inferenza è guidata dall'antecedente, in quanto la ricerca delle regole da applicare si basa sul *matching* tra i vari fatti in memoria e quelli combinati logicamente nell'antecedente della regola attiva; b) il «concatenamento all'indietro» (*backward chaining*). In questo caso si procede mediante riduzione dell'obiettivo principale (*goal*) a sottoproblemi. Allora, una volta individuata la tesi da dimostrare, si applicano all'indietro le regole di produzione, cercando di trovare coerenza con i dati iniziali. L'interprete ricerca, se esiste, una regola che abbia come conseguente l'asserzione di cui deve provare la veridicità. Da qui si volge a provare i fatti che costituiscono l'antecedente della regola trovata. Si parla perciò di inferenza guidata dal conseguente.

2. Sistemi esperti. I sistemi esperti sono l'esempio applicativo più noto derivante da questo approccio. Un sistema esperto, ossia un «sistema basato sulla conoscenza», è uno strumento in grado di risolvere problemi in un dominio limitato, ma con prestazioni simili a quelle di un esperto umano del dominio stesso. Questo significa che il compito fondamentale di un sistema esperto è quello di coadiuvare l'attività di utenze professionali, laddove è usualmente richiesta la consulenza di uno specialista umano dotato di competenza (*expertise*) e capacità di giudizio. Le ricerche di IA hanno posto luce sui problemi realizzativi di tali strumenti, affermando la necessità di restringere, per quanto possibile, il campo di applicazione. Dunque, rispetto ad un esperto umano, questi applicativi si rivelano certamente più limitati e superficiali, non disponendo di quella completezza che costituisce la conoscenza culturale della persona competente. Inoltre non è possibile sperare che un sistema esperto possa giungere a conclusioni in maniera intuitiva o saltando alcuni passaggi logici, affidandosi al "buon senso" o al meccanismo della (↗) analogia, com'è invece prerogativa dell'uomo. In definitiva, viene simulato un esperto umano con tratti più o meno abbozzati, e lo si fornisce della capacità di risolvere compiti ristretti, temporanei o secondari. Il primo e più noto di tali sistemi è *Mycin,* sviluppato da E.M. Shortleffe a partire dal 1972 ed applicato in campo medico. Per quanto riguarda più specificatamente i tipi di problemi che un sistema esperto può essere chiamato a risolvere, si può stendere una lista di argomenti, ovviamente non esaustiva: a) "diagnosi": si tratta di individuare, in base al riconoscimento di determinati sintomi, le possibili cause di "malfunzionamento" e suggerire un cammino di cura; b) "monitoraggio": viene seguito lo sviluppo temporale di un processo; si procede al controllo dell'acquisizione e dell'elaborazione di dati di vario tipo, fornendo in uscita informazioni sintetiche sullo stato e stime sulla sua evoluzione; c) "pianificazione": note le risorse a disposizione, se ne individua l'impiego ottimo allo scopo di conseguire un certo obiettivo entro un dato tempo; parallelamente si indirizza l'acquisizione di nuove risorse; d) "interpretazione di informazioni e segnali": avendo in ingresso una serie di dati relativi ad un certo àmbito, si vuole effettuare una valutazione complessiva al fine di riconoscere il presentarsi di alcune situazioni predeterminate.

3. Giochi. Un altro campo applicativo in cui questo tipo di approccio, simbolico e ingegnerisico, ha avuto notevoli successi è quello dei giochi. L'intelligenza artificiale considera generalmente giochi a due giocatori in cui le mosse sono alternate e interpreta lo svolgersi del gioco come un "albero" in cui la "radice" è la posizione di partenza e le "foglie" sono le posizioni finali (vincenti o perdenti). Ovviamente, a causa della complessità dei giochi trattati, sarebbe impensabile, anche per un potentissimo *computer,* di sviluppare completamente tutto l'albero per decidere la mossa "migliore". Ecco quindi la necessità di applicare opportune euristiche per "potare" alcuni rami dell'albero e rendere il problema trattabile. Si pensi al gioco degli scacchi in cui la dimensione del problema è enorme. Solo all'inizio partita le mosse possibili sono 400, diventano più di 144.000 alla seconda mossa. Sviluppando l'albero di gioco avremmo circa 35^{100} nodi. Applicando tecniche di manipolazione simbolica e utilizzando metodi potenti per ridurre la dimensione dello spazio di ricerca, altrimenti intrattabile, si sono prodotti comunque sistemi in grado di giocare a scacchi meglio dell'uomo, anche se, ovviamente, utilizzando tecniche ben differenti da quelle umane. È infatti noto che nel maggio 1997, a New York, una macchina (*Deep Blue*) ha battuto in un *match* di sei partite il campione del mondo Kasparov. Interessante è sottolineare che tale macchina, appositamente progettata a livello *hardware* per riuscire a sviluppare ed esaminare spazi di ricerca in parallelo in tempi rapidissimi (si pensi che *Deep Blue* arriva ad esplorare 10^{11} posizioni in circa 3 minuti) utilizza la "forza bruta" piuttosto che tecniche euristiche raffinate per giungere rapidamente alla soluzione migliore.

4. Dimostrazioni matematiche e linguaggi di programmazione logica. L'utilizzo della (↗) logica e l'automazione delle dimostrazioni matematiche è un altro campo applicativo in cui l'IA ha raggiunto notevoli risultati. La logica è sicuramente uno degli strumenti più antichi, assestati e rigorosi utilizzati dall'uomo per formalizzare e spiegare il proprio ragionamento. È semanticamente ben definita, altamente dichiarativa, ed ha un apparato deduttivo assolutamente generale. Questo spiega perché la logica classica (in particolare quella del primo ordine) sia tanto utilizzata in IA per rappresentare la conoscenza su un problema, anche se questa scelta ha delle limitazioni evidenti (vedi *infra,* III.3) e non trova un consenso unanime. Minsky sostiene al riguardo che le formule ed i metodi di deduzione logici non sono il modo più naturale con cui ragionare e non sono i metodi con cui l'uomo organizza la sua conoscenza e mostra un comportamento intelligente. La base di conoscenza diventa in questo caso una collezione di asserzioni della logica dei predicati del primo ordine. Le regole di inferenza permettono di dedurre nuove asserzioni ("teoremi") non esplicitamente contenute nella base di conoscenza iniziale. La sequenza di regole di inferenza utilizzate nella derivazione del teorema si chiama «prova del teorema». Ovviamente, volendo automatizzare il procedimento, l'efficienza della prova diventa un requisito fondamentale. Gran parte dei programmi che utilizzano la logica in IA sono basati sugli studi sulla dimostrazione automatica di teoremi nella logica, ed in particolare sul metodo di risoluzione messo a punto da J.A. Robinson negli anni '60 ed allo sviluppo di strategie per rendere più efficiente la dimostrazione. Figli di questi studi sono anche la programmazione logica ed il linguaggio *Prolog*

(da *PROgramming in LOGic*) in particolare, che si sta affermando come uno dei più interessanti ed innovativi paradigmi di programmazione per lo sviluppo di applicazioni "intelligenti".

La nozione di «programmazione logica» nasce agli inizi degli anni '70, per merito soprattutto di alcuni ricercatori delle università di Edimburgo e di Marsiglia. A Robert Kowalski, allora all'università di Edimburgo, si deve il merito di aver definito precisamente i fondamenti teorici della programmazione logica, ed in particolare di aver proposto un'interpretazione procedurale delle clausole della logica che permette di ridurre il processo di dimostrazione di un teorema al più tradizionale processo di computazione dei linguaggi di programmazione. Al gruppo di Alain Colmerauer a Marsiglia va invece il merito di avere realizzato, per primo, nel 1972, un "interprete" per il linguaggio *Prolog*, dimostrando così la fattibilità pratica della nozione di programmazione logica. La programmazione logica si differenzia radicalmente dalle tecniche di programmazione che vengono normalmente utilizzate per scrivere programmi nei linguaggi tradizionali. I linguaggi di programmazione più diffusi, dal *Fortran* al *Pascal* al *C,* sono infatti basati sul paradigma imperativo, secondo il quale un programma consiste di una sequenza di comandi che specificano in modo estremamente dettagliato le operazioni che dovranno essere eseguite dall'elaboratore per risolvere il problema dato. Viceversa, in programmazione logica un problema viene descritto in termini molto più astratti con un insieme di formule della logica. Questo modo di rappresentare i problemi consente una comprensione dichiarativa della conoscenza, con cui si descrive un problema senza specificare in modo dettagliato come si potrà ottenere la soluzione. In altre parole, la programmazione logica condivide con la dimostrazione automatica dei teoremi l'uso della logica per rappresentare la conoscenza e l'uso della deduzione per risolvere problemi. Tuttavia, essa pone l'accento sul fatto che la logica può essere usata per esprimere programmi e che particolari tecniche di dimostrazione possono essere usate per eseguire i programmi.

5. L'apprendimento. Al di là della doverosa elencazione di questi sistemi ritenuti di successo dal punto di vista applicativo, anche se con indubbi limiti se valutati in un'ottica meno riduttiva, è quasi universalmente riconosciuto che le macchine non potranno dirsi «intelligenti» fino a quando non saranno in grado di accrescere le proprie conoscenze e di migliorare le proprie abilità. Scrive Simon (1983***): «L'apprendimento consiste in cambiamenti del sistema che siano adattativi, nel senso che mettono in grado il sistema di svolgere la prossima volta lo stesso compito in modo più efficiente ed effettivo». Un metodo per risolvere, anche se molto parzialmente, questo problema è dotare le macchine simboliche di capacità ragionamento induttivo oltre che deduttivo. Il ragionamento induttivo procede da asserzioni singolari riguardanti particolari fatti o fenomeni ("esempi") ad asserzioni universali esprimibili mediante ipotesi o teorie che spieghino i fatti dati e siano in grado di predirne di nuovi. Mentre però l'inferenza deduttiva preserva la "verità" (nel senso di correttezza logica), l'inferenza induttiva non garantisce ciò, e quindi tali sistemi possono tendere ad un'eccessiva generalizzazione e produrre errori. Si tratta sempre di un approccio simbolico in

quanto i risultati di tale procedimento sono una nuova teoria, nuove regole e, in generale, una base di conoscenza nuova o aggiornata. Uno dei più noti programmi di apprendimento dagli esempi è *ID3*, sviluppato da J. Ross Quinlan (1979-1983), da cui sono nati prodotti commerciali per la classificazione automatica. *ID3* e i suoi "discendenti" hanno esplorato migliaia di basi di dati producendo regole di identificazione in differenti aree (ad esempio diagnosi di malattie). Attualmente i programmi di apprendimento sono ampiamente utilizzati dal punto di vista pratico per far fronte all'esigenza di sfruttare il patrimonio informativo contenuto nelle grandi raccolte di dati accessibili su rete, o nelle basi di dati aziendali, per estrarre regolarità fra i dati, informazioni e conoscenze nascoste (*data mining*).

6. Le reti neurali. Le reti neurali rappresentano un approccio significativamente diverso da quello simbolico analizzato precedentemente, e rientrano nel filone dell'IA che abbiamo citato come "strutturale" o "connessionistico" (vedi *supra,* I.1). L'idea di base è di riprodurre l'intelligenza e, in particolare, l'apprendimento simulando all'elaboratore la struttura neurale del cervello animale. I calcolatori possono memorizzare con facilità grandi quantità di informazioni, operano in nanosecondi e possono svolgere enormi moli di calcoli aritmetici senza errore, mentre gli uomini non sono in grado di avvicinarsi a tali prestazioni. È indubbio, però che gli uomini, normalmente, svolgono "semplici" compiti come camminare, parlare, interpretare una scena visiva o comprendere una frase, ragionare su eventi di senso comune, trattare situazioni incerte, in modo molto più brillante ed efficiente dei più raffinati e costosi programmi di IA risultanti dall'approccio simbolico e funzionale.

L'idea di costruire una macchina intelligente a partire da neuroni artificiali si può fare risalire alla nascita dell'IA, e già alcuni risultati furono ottenuti da McCulloch e Pitts nel 1943 quando nacque il primo modello neurale; essi furono poi approfonditi da altri ricercatori. Nel 1962 Rosenblatt propose un nuovo modello di neurone, il «percettrone», capace di apprendere mediante esempi. Un percettrone descrive il funzionamento di un neurone eseguendo una somma pesata dei suoi ingressi ed emettendo un'uscita "1" se la somma è maggiore di un certo valore di soglia modificabile, o "0" altrimenti. L'apprendimento, così inteso, è un processo di modifica dei valori dei pesi. Il grande entusiasmo verso questo approccio subì una brusca riduzione pochi anni dopo, quando Minsky e Papert evidenziano i grandi limiti di apprendimento del percettrone. Più recentemente sono state proposte nuove architetture di reti neurali non più soggette alle limitazioni teoriche dei percettroni, dette "connessioniste", che utilizzano potenti algoritmi di apprendimento (propagazione all'indietro). Questo ha risvegliato un forte interesse per le reti neurali ed ha consentito lo sviluppo di applicazioni di successo. L'architettura "connessionista", alternativa a quella di Von Neumann è caratterizzata da: a) un grande numero di elementi di elaborazione molto semplici, simili a neuroni; b) un grande numero di connessioni (sinapsi) pesate tra gli elementi; c) un controllo distribuito altamente parallelo. I pesi codificano, di fatto, la conoscenza di una rete. Le variazioni che si hanno durante l'apprendimento si possono considerare come variazioni

dinamiche dei pesi delle connessioni. Si possono distinguere varie modalità di apprendimento in dipendenza da come la rete viene "addestrata". In particolare, i paradigmi di apprendimento possono essere suddivisi in tre classi fondamentali: a) apprendimento con supervisione mediante esempi (*Supervised Learning*): un insegnante fornisce alla rete le risposte desiderate che i neuroni dovrebbero produrre dopo la fase di addestramento; b) apprendimento senza supervisione (*Unsupervised learning*): i neuroni si specializzano mediante una competizione interna al fine di discriminare gli stimoli presentati in ingresso; c) apprendimento mediante rinforzo (*Reinforcement Learning*): alla rete viene fornita solo una informazione qualitativa sulla bontà della sua risposta; un critico valuta la risposta della rete ed invia ai neuroni un segnale di rinforzo positivo se la valutazione è buona, negativo altrimenti.

Nei sistemi connessionisti sembra essere più facile realizzare sistemi di apprendimento, ma tale apprendimento, nascosto in variazioni di valori numerici reali, rimane cablato all'interno della rete e non può essere esplicitato in forma simbolica. Le reti neurali sono quindi più adatte a compiti di classificazione e di percezione concettualmente "di basso livello" anche se tecnicamente ardui, quali il riconoscimento del parlato, il controllo di processi e il riconoscimento di immagini, mentre problemi concettualmente complessi quali progettazione, diagnosi, pianificazione, rimangono dominio della IA simbolica. Mentre i modelli a reti neurali sono basati sulla simulazione del cervello umano, molte altre tecniche di IA si ispirano all'evoluzione del mondo animale e delle sue società.. Gli algoritmi genetici, ad esempio, sono algoritmi basati sull'evoluzione, in cui l'apprendimento avviene attraverso un processo selettivo a partire da una vasta popolazione di programmi casuali.

7. Limiti e nuovi traguardi. Molte critiche sono state portate agli attuali sistemi di IA: sicuramente essi sono poveri e deludenti se confrontati con le prime aspettative dell'intelligenza artificiale. Non si sono visti, in effetti, passi da gigante ed i problemi più impegnativi quali, ad esempio, l'apprendimento e la rappresentazione del comune buon senso, anche se affrontati e risolti parzialmente, sono ben lontani da una completa soluzione. Per quello che riguarda l'approccio funzionale all'IA, nonostante i molti punti a favore del modello architetturale dei sistemi basati sulla conoscenza come la modularità dell'architettura e la possibilità di una crescita incrementale della conoscenza, solo pochi sistemi esperti commerciali sono veramente operativi, costituendo peraltro una percentuale molto bassa rispetto ai programmi convenzionali. Un pesante "collo di bottiglia" per la loro diffusione è senz'altro quello dell'acquisizione della conoscenza. È particolarmente complesso, infatti, estrarre in modo completo la conoscenza dall'esperto e riuscire a formalizzarla nella base di conoscenza. Inoltre tali sistemi hanno un alto costo di mantenimento ed aggiornamento. D'altro canto, l'alternativa all'approccio funzionale rappresentata da connessionismo e reti neurali trova applicazioni anche di successo, ma spesso limitate alla soluzione di problemi ritenuti di più basso livello quali la percezione e il riconoscimento.

Per quanto riguarda le prospettive future, attualmente la rivoluzione tecnologica che porta alla società dell'informazione dà la possibilità di accedere ad un'enorme mole di materiale informativo, che deve però essere gestito ed interpretato in maniera corretta. Si va dai grandi archivi aziendali all'informazione *on line,* aggiornata "in tempo reale"; dalla capacità di cogliere lo scibile nelle sue manifestazioni più pratiche —come l'esperienza maturata "sul campo" dallo specialista — all'indagine del dettaglio, volta a traguardi sempre più accurati. Ogni proposito di sviluppo si trova ad affrontare una mole non strutturata di dati eterogenei e ridondanti. Pare dunque giustificato cercare non solo di rafforzare, ma soprattutto di rivoluzionare gli strumenti di estrazione e di analisi dell'informazione, al fine di utilizzare questo grande patrimonio conoscitivo al meglio delle sue potenzialità. Diventa quindi fondamentale l'uso delle metodologie per l'estrazione della conoscenza precedentemente citate, che utilizzano tecniche di apprendimento simbolico e reti neurali.

Attualmente vi è anche una forte spinta all'integrazione dei sistemi di IA, ed in particolare dei sistemi esperti con il resto del mondo dell'ingegneria dell'informazione, ove ritroviamo l'uso corrente di tecnologie quali la programmazione o la costruzione di basi di dati "orientate agli oggetti" (*Object Oriented Programming* e *Object Oriented Data Base*) e le "interfacce grafiche" (*Graphic User Interfaces*), talune delle quali sono originariamente nate nell'ambito dell'IA. Un fenomeno importante è la tendenza all'estinzione del sistema esperto, inteso come applicazione a sé stante, a vantaggio di una visione integrata: si tende infatti a realizzare moduli che producano *task* "intelligenti", strettamente integrati nelle applicazioni *software* e nei sistemi informativi generali. L'idea è quindi quella di costruire "agenti intelligenti" con capacità di ragionamento deduttive ed induttive, preposti a particolari compiti e in grado di coordinarsi con altri agenti in ambito distribuito al fine di raggiungere insieme un unico obiettivo. Le funzioni realizzate da tali agenti intelligenti devono integrarsi con le funzioni realizzate da altri moduli, magari preesistenti al sistema stesso, quali l'interfaccia con l'operatore, i sistemi di gestione degli archivi (*DBMS*), i moduli di acquisizione dati e i sistemi grafici. In questi campi l'IA sta già cogliendo e potrà cogliere nuovi prestigiosi successi.

III. Che cosa l'intelligenza artificiale non può fare

Il dibattito sull'IA è stato negli ultimi due decenni del Novecento, ed è ancora, fra i più appassionati e appassionanti della ricerca filosofica. Ed è naturale perché l'IA riapre, con grande forza di provocazione, il problema di che cosa siano la mente, l'intelligenza o l'intelligenza cosciente (↗ MENTE-CORPO, RAPPORTO). La discussione riguarda due grandi temi: che cosa può fare l'IA, e che cosa è lecito fare con l'IA.

1. IA debole e IA forte. Conviene, fin dal principio, distinguere fra "IA debole" e "IA forte": al significato di questi termini s'è già accennato

inizialmente (vedi *supra,* I.1), ma ora conviene richiamarlo e precisarlo. L'IA debole vuol costruire macchine che si comportino "come se" fossero intelligenti: ossia macchine capaci di risolvere "tutti" i problemi che l'intelligenza umana sa risolvere. L'IA forte vuole di più: afferma che la macchina che agisce in modo intelligente deve avere una "intelligenza cosciente", una mente cosciente indistinguibile dalla mente umana. Si noti che l'IA debole si occupa della concreta costruzione o costruibilità di macchine "pensanti", mentre l'IA forte vuol dare risposta al problema astratto di che cosa sia il loro "pensare". Perciò, come osservano Russell e Norvig, si può credere all'IA forte ed essere scettici sull'IA debole (cfr. *Intelligenza artificiale*, 1998, p. 884): ossia pensare che le macchine intelligenti, se fossero costruite, avrebbero un'intelligenza cosciente; ma ritenere che esse non possano essere costruite.

Alcune delle obiezioni riguardano l'IA debole, ma le più radicali sono quelle portate all'IA forte. Il dibattito parte da asserzioni di questo tipo: «il cervello è una macchina e perciò, in linea di principio, si può costruire una macchina che faccia tutto ciò che fa il cervello». Applicato alle proprietà della mente, questo enunciato assume un chiaro sapore riduzionistico poiché assume implicitamente che "mente" coincida con "cervello". Questa impostazione è chiaramente materialista (↗ MATERIALISMO, IV; MATERIA, VIII) ed estremamente discutibile ma, poiché tutto il dibattito si svolge al suo interno, di essa si potrà discutere più avanti.

2. Che cosa l'IA debole non può fare. Incominciamo dalle critiche all'IA debole, pur facendo presente che molte di esse si applicano, per naturale estensione, anche all'IA forte. Un primo gruppo consiste in affermazioni apodittiche del tipo: «la macchina non potrà mai fare la tal cosa». Obiezioni di questo genere hanno sempre accompagnato le innovazioni della tecnica, e sono soprattutto manifestazione del rifiuto psicologico, del timor panico di fronte ad un "nuovo" incomprensibile: perciò devono essere guardate con estrema diffidenza. In generale esse non hanno suscitato discussione: sono state smentite quasi sempre dai fatti. Più esattamente, i tecnici hanno raccolto le sfide e si sono adoperati per costruire precisamente macchine che facessero ciò che si dichiarava impossibile. Clamoroso è l'esempio del gioco degli scacchi di cui si è parlato (vedi *supra,* II.3), anche perché la convinzione che la macchina non avrebbe mai potuto battere il grande maestro è durata forse più a lungo di ogni altra, e più duramente ha sfidato i tecnici. Ma è istruttivo vedere "come" l'elaboratore ha battuto il maestro: quest'ultimo "vede" sulla tastiera la mossa giusta, la "intuisce" in un modo apparentemente simile all'intuizione dell'artista, con un processo mentale al quale diamo il nome di "genio", ma del quale non sappiamo nulla. A tutto ciò la macchina oppone la "forza bruta" di un numero enorme di circuiti velocissimi, che le consentono di fare un gran numero di tentativi in cerca della mossa che assicura la maggior probabilità di vittoria. S'incomincia a intravedere, in questo, una differenza fondamentale fra uomo e macchina che sfugge alle analisi di tipo riduzionistico.

Però Turing, nello stesso articolo in cui propone il "gioco dell'imitazione", presenta anche un fantasioso elenco di operazioni che "la macchina non potrà mai fare", e prudenza vuole che alle provocazioni di Turing si presti qualche attenzione. Alcuni punti dell'elenco, come «apprendere dall'esperienza», sono stati, almeno ad un certo livello, realizzati: lo scetticismo di Turing non è stato da meno di quello degli sprovveduti avversari dell'IA. Altri, semplicemente, non riguardano la macchina in quanto soggetto dell'IA. Per esempio, «essere bello» oppure «far innamorare qualcuno di sé»: la narrativa di fantascienza favoleggia di bellissimi *robot* umanoidi e di umani che se ne innamorano, ma questo potrebbe riguardare, se mai, la robotica (se in qualche tempo futuro volesse dilettarsi nel costruire *robot* umanoidi), ma certamente non l'IA. Oppure «gustare delle fragole con panna»: si può immaginare (e qualcosa di simile è stato costruito) un *robot* dotato di sensori del gusto e dell'olfatto, e di un programma che discrimini i gusti gradevoli. Anche in questo caso la questione riguarderebbe soprattutto la robotica, ma con un problema in più: "gustare" implica già una facoltà tipica della mente umana e quindi chiama in causa il problema di fondo dell'IA forte. Altrettanto può dirsi della più inquietante: «essere l'oggetto del proprio pensiero», ossia essere autocosciente.

Una variante di queste obiezioni è: «la macchina può fare soltanto ciò che noi le sappiamo ordinare di fare», cioè è priva di volontà libera e le scelte che compie sono condizionate. A questo si può rispondere in due modi opposti. Da un lato gli algoritmi genetici e molte delle applicazioni descritte (vedi *supra,* II) mostrano che la macchina può allargare di molto e modificare la gamma delle sue possibilità. Dall'altro, però, si potrebbe osservare che anche queste modificazioni sono, in qualche modo, previste e perciò potenzialmente incluse nella programmazione originaria: e questo metterebbe in luce l'essenziale, indissolubile dipendenza della macchina dall'uomo. Ma questo mette in evidenza anche un fatto di sostanziale novità: la trasgressione del paradigma metodologico fondamentale dell'ingegneria, la progettualità. Essa vuole che ogni oggetto tecnico sia disegnato in ogni sua parte prima che s'intraprenda la sua costruzione. Ma i sistemi di apprendimento, e *massime* le reti neurali, sono inizialmente oggetti in una certa misura informi: il tecnico definisce la loro struttura formale ma i "pesi" delle connessioni, che caratterizzano quella particolare rete neurale in ordine al compito che deve svolgere, si vanno precisando durante l'apprendimento, in una maniera dipendente non dall'intenzionalità del progettista ma dall'informazione fornita; e, alla fine (ammesso che vi sia una fine, ossia che il processo d'apprendimento non duri per tutta la vita del sistema), essi assumono valori numerici del tutto imprevedibili e, d'altronde, privi d'interesse per il tecnico che ha progettato la rete e la usa. In questo la tecnica dell'IA sembra anticipare una tendenza che poi si è diffusa in molti campi dell'ingegneria, e soprattutto nell'ingegneria dell'informazione: la generazione e l'impiego di sistemi "non progettati". Si veda, per esempio, il caso di *Internet* (➶ INFORMAZIONE, VI).

Simili alle precedenti sono le obiezioni di tipo "quantitativo" del tipo: «non si riuscirà mai a costruire una macchina abbastanza potente per risolvere questo

problema». Anche per la maggior parte di queste obiezioni i fatti si sono incaricati della smentita, ma qualcosa ne è rimasto. Con particolare forza esse sono state sollevate quando, negli anni Sessanta, la soluzione di problemi matematici si è scontrata con l'"esplosione combinatoria" della quale si è già parlato. Nella matematica applicata si parla di problemi «intrattabili» quando il tempo di calcolo cresce almeno esponenzialmente con il numero di variabili. Si tratta dunque di un'impossibilità "pratica": uno stesso problema può essere risolto se le incognite sono poche, ma diventa insolubile (in un tempo ragionevole) quando il loro numero aumenta. A questo proposito Turing ha obiettato che spesso la "quantità" diventa "qualità", nel senso che al di sopra di certe dimensioni il comportamento dei sistemi può modificarsi sostanzialmente, e rendere improvvisamente possibile ciò che prima non lo era. L'esempio portato da Turing, dei reattori nucleari che, oltre certe dimensioni, passano da "sottocritici" a "critici" e producono energia, non è che un caso particolare della ben nota proprietà della maggior parte dei sistemi dinamici, di passare dalla stabilità all'instabilità al variare del valore di qualche loro parametro. Dunque si tratterebbe di trovare una "struttura" che, quando raggiunga certe dimensioni — per esempio una rete con abbastanza neuroni — diventi capace di dominare l'esplosione combinatoria.

Un altro problema di "dimensioni", ampiamente discusso da Hubert Dreyfus (1988), riguarda l'enorme quantità d'informazione (la «base di conoscenza») necessaria, per esempio, per "contestualizzare" il discorso parlato e così eliminare le sue inevitabili ambiguità. Questa base di conoscenza altro non è che quella che l'uomo accumula con l'apprendimento. Perciò il problema è duplice: realizzare una "memoria" di dimensioni adeguate, e immettervi l'informazione. E il problema dell'immissione si articola in diversi sottoproblemi: a) come costruire una "conoscenza di fondo" a partire dalla quale impostare l'apprendimento; b) come organizzare il processo di apprendimento (che sarà, in generale, un processo «per rinforzo») in modo da ottimizzare il suo rendimento; c) come realizzare i procedimenti «induttivi» che dall'esperienza generano conoscenza; d) come controllare l'acquisizione dei dati sensoriali. Dreyfus aveva proposto questi problemi in forma negativa, con una forte coloritura di pessimismo circa la possibilità di risolverli, ma le sue obiezioni si sono risolte in un potente stimolo alla loro risoluzione.

3. I limiti della matematica e della logica. La realizzazione dell'IA, anche nella forma debole, incontra alcune gravi difficoltà di carattere teorico, sulle quali s'è concentrata una parte ragguardevole delle obiezioni. Per esempio c'è il «problema della terminazione»: l'esecuzione di un certo programma avrà termine, o potrebbe proseguire, in linea toerica, all'infinito? Questo problema non ha soluzione: Turing ha dimostrato che per ogni algoritmo che, applicato a qualsivoglia programma, dovrebbe dire se la sua esecuzione avrà termine, si può trovare invece un programma per il quale quell'algoritmo non potrà dare risposta. La difficoltà più grave deriva dal «teorema d'incompletezza» di (↗) Gödel, e su di essa si è incentrato un dibattito, talvolta aspro. Il teorema di incompletezza afferma che in qualunque sistema logico formale (purché sufficientemente

potente) è possibile formulare proposizioni vere, delle quali tuttavia gli strumenti propri del sistema non permettono di dimostrare la verità. Ne scrive John Lucas in un famoso articolo, *Minds, Machines and Gödel* (1961) che incomincia così: «Mi pare che il teorema di Gödel dimostri che il meccanicismo è falso, cioè che le menti non possono essere equiparate a macchine».

C'è dunque una cosa che le macchine non possono fare: decidere la verità di proposizioni indecidibili. L'uomo invece sì, perché egli sa «porsi fuori dal sistema»: per esempio può applicare il teorema di Gödel al sistema medesimo. Douglas Hofstadter, che pure cita questo articolo a lungo e con ammirazione, replica alquanto sarcasticamente dimostrando che l'uomo "non può" porsi fuori dal sistema, perché ciò condurrebbe a una regressione infinita (cfr. Hofstadter, 1984, pp. 508-510). Il che, da un punto di vista riduzionistico, è corretto, e tuttavia contraddice all'esperienza comune che mostra come l'uomo sappia effettivamente superare il limite della pura logica e guardare i problemi logici "dall'esterno". La questione è stata ripresa da Roger Penrose (1992), che propone una via d'uscita. Penrose, che è un fisico di fama, a sostegno della tesi di Lucas osserva, in primo luogo, che se la mente è in grado di comprendere la matematica non computazionale non può essere soltanto un sistema logico formale; ma a questo aggiunge un argomento della sua propria materia. Egli prende l'avvio dalla constatazione che esiste una radicale dicotomia fra la descrizione matematica della meccanica quantistica e quella della fisica classica che è ancora valida a livello macroscopico; e dal fatto che le leggi della fisica sono reversibili rispetto al tempo, ossia non tengono conto della irreversibilità del tempo mostrata dal secondo principio della termodinamica, e riconoscibile anche nella nostra coscienza. Perciò Penrose ipotizza che si possa scoprire una "nuova" fisica più completa e profonda, «che renda possibile la fusione fra mondo classico e mondo quantistico, che sia "asimmetrica" rispetto al tempo e che fisicamente renda possibile comprendere la natura della mente» (G. Piccinini, 1994, p. 141). Ma questa fisica implicherebbe anche un nuovo tipo di matematica, che dovrebbe «contenere elementi essenzialmente non computabili» (*ibidem*). Questa matematica non computazionale supererebbe i limiti dell'IA, che si basa invece su una matematica computazionale, e potrebbe includere operazioni che il teorema di Gödel vieta all'IA ma non alla mente umana.

4. L'IA forte: elaborazione sintattica e contenuto semantico. John Searle, nel 1980, ha portato contro l'IA forte un'obiezione del tutto diversa, cui ha dato la forma di un divertente apologo: l'«esperimento concettuale della stanza cinese». Non conoscendo il cinese, egli osserva, supponiamo che mi trovi in una stanza piena di ideogrammi cinesi e che mi venga fornito un manuale di regole in base al quale associare ideogrammi cinesi ad altri ideogrammi cinesi. Le regole specificano senza ambiguità gli ideogrammi in base alla loro forma e non richiedono che io li capisca. Supponiamo adesso che fuori dalla stanza vi siano delle persone che capiscono il cinese e che introducano gruppetti di ideogrammi e che, in risposta, io manipoli questi ideogrammi secondo le regole del manuale e restituisca loro altri gruppetti di ideogrammi. Se le regole del manuale specificano abbastanza accuratamente quali gruppi di ideogrammi possono essere

associati agli ideogrammi introdotti, in modo che le "risposte" abbiano senso compiuto e siano coerenti con le domande, chi sta fuori dalla stanza può concludere erroneamente che chi sta dentro conosca il cinese. Ossia che chi sta dentro la stanza abbia eseguito l'elaborazione "sintattica" del messaggio in base alla comprensione della sua "semantica", mentre invece la "semantica" è rimasta fuori dalla stanza (cfr. Searle 1990 e 1992). Ciò è quanto accade in tutti i calcolatori (e non solo nell'IA): essi eseguono operazioni sintattiche sui messaggi introdotti, del tutto indipendenti dal loro contenuto semantico. La semantica si arresta, per così dire, all'ingresso del messaggio nel calcolatore, e gli viene restituita da chi riceve il messaggio all'uscita. Si noti che con questo Searle affronta uno dei problemi che, come s'è già accennato (vedi *supra,* I.2), vent'anni prima avevano alquanto attenuato l'iniziale entusiasmo per la nascente IA.

Forse anche in ragione della sua forza, questa argomentazione ha ricevuto un gran numero di contestazioni, talvolta pittoresche. Paul e Patricia Churchland (marito e moglie, colleghi di Searle all'Università di California, Searle a Berkeley e i Churchland a San Diego), volendo mettere in luce il fatto che, per loro, l'esperimento della stanza cinese non è che un capzioso sillogismo, gli hanno contrapposto l'«esperimento della stanza luminosa» (cfr. Churchland e Churchland, 1990). Ma il punto d'arrivo di queste contestazioni consiste nel negare che vi sia una distinzione essenziale e qualitativa fra sintassi e semantica: poiché ogni processo mentale ha sede nel cervello entrambe sarebbero aspetti, fra loro strettamente correlati, dell'attività cerebrale; e dal momento che la semantica risiede nel cervello la sua apparente differenza dall'elaborazione sintattica sarebbe correlata con l'estrema complessità della struttura cerebrale. Dunque anche la semantica potrebbe essere trasferita alle macchine purché avessero una sufficiente complessità circuitale e algoritmica. Anche questa posizione, in un'ottica riduzionistica, appare ineccepibile. Ma ad essa si oppone, con ulteriori argomentazioni, Hubert Dreyfus (1988) che nega che i calcolatori posseggano non solo competenza semantica, ma anche le capacità sintattiche di livello più alto, quelle che servono a «tematizzare, heideggerianamente, la loro presenza nel mondo, di mettersi cioè in discussione fino al punto di poter arrivare a superare il proprio contesto di partenza, per collocarsi […] in altri contesti di realtà che contengano eventualmente il primo, e sempre avendone coscienza. […] Il limite dell'intelligenza artificiale […] consiste nel fatto […] che l'intelligenza e la coscienza reali, non artificiali, hanno la capacità di connettere livelli logici, sintattici e semantici diversi, e di metterli continuamente in discussione, come nessun computer pensabile su basi fisiche […] appare concepibilmente in grado di fare» (Rossi, 1998, p. ***).

IV. L'intelligenza artificiale fra scienza e coscienza

Con le precedenti considerazioni si giunge ad un punto ritenuto centrale, vale a dire al "problema della coscienza". Lo stesso Searle (***) lo propone mediante l'esperimento concettuale della "protesi cerebrale", ipotizzando che con un

intervento di raffinatissima microchirurgia, si riesca a sostituire uno per uno tutti i neuroni di un cervello con altrettanti microcircuiti elettronici che funzionino esattamente allo stesso modo dei neuroni, e che siano riprodotte tutte le connessioni sinaptiche. Che succederebbe, ci si chiede allora, della coscienza di quell'uomo? Per Searle essa svanirebbe; invece per Hans Moravec, che otto anni dopo ha ripreso la questione e l'ha esaminata da un punto di vista "funzionalistico", essa resterebbe inalterata. Ma si tratta, soprattutto, di definire che cos'è la coscienza. E questo, di fatto, non è un problema risolubile con il metodo scientifico.

1. Critica del riduzionismo. Si è parlato di "prospettiva riduzionistica" (vedi *supra,* IV, 1), intendendo con queste parole il fatto che si identifica la mente con il suo "supporto" materiale, il cervello, e si vede questo come una "macchina" pienamente riproducibile con dispositivi artificiali (↗ MENTE-CORPO, RAPPORTO). Per cui se v'è differenza fra mente e macchina, essa è da attribuire o a temporanea insufficienza della macchina, da rimediare in futuro, o a un limite che la mente non sa ancora d'avere. Più in generale, s'intende qui per (↗) riduzionismo l'idea che la mente umana possa essere simulata, almeno in via di principio, da sistemi artificiali capaci di riprodurne le prestazioni in maniera così perfetta da rendere indistinguibile l'una dagli altri (cfr. Rossi, 1998, pp. 43-44). E, per alcuni autori, la simulazione arriva al punto che il sistema artificiale possieda attributi schiettamente umani quali la coscienza e l'intenzionalità (↗ ANIMA, VII). Quasi nessuno degli autori qui citati precisa se questa sia la sua posizione di principio, quali siano le eventuali diversificazioni al suo interno e per quali motivi: essa appare "naturale", come se fosse l'unica possibile. Jerry A. Fodor (1981) preferisce chiamare questa posizione «materialismo», e la contrappone al «dualismo» cartesiano (↗ DESCARTES, VI) che è rifiutato per «la sua incapacità di rendere conto adeguatamente della causazione mentale. Se la mente è qualcosa di non fisico, non ha una posizione nello spazio fisico. Come è possibile, allora, che una causa mentale dia luogo a un effetto comportamentale che invece ha una posizione nello spazio? Per dirla in modo diverso, come può il non fisico dar luogo al fisico senza violare le leggi di conservazione della massa, dell'energia e della quantità di moto?» (p. ***). All'interno del materialismo si distinguono, poi, posizioni "comportamentistiche" derivate dalla psicologia e posizioni che, invece, privilegiano gli aspetti neurofisiologici. E poi, per Fodor distinte sia dal materialismo sia dal dualismo, ci sono le posizioni "funzionalistiche" che prescindono dalla struttura del cervello o dei sistemi che lo simulano per concentrare l'attenzione sulla «possibilità che sistemi così diversi fra loro come esseri umani, macchine calcolatrici e spiriti disincarnati possano tutti avere stati mentali» (*ibidem*). Ma Fodor non prende in considerazione l'eventualità che i rispettivi "stati mentali", e le facoltà intellettive che li producono, siano essenzialmente diversi fra loro.

Il lettore potrà tentar di riconoscere queste diverse posizioni nell'esposizione che precede. Si tratta, come si vede, di una manifestazione del tutto coerente dell'atteggiamento "antimetafisico" che domina l'intera ricerca scientifica post-galileiana e post-cartesiana. Ma esso, qui, investendo questa zona di frontiera fra

corpo e mente, fra fisico e meta-fisico, raggiunge effetti contraddittori e, in qualche misura, paradossali. Hofstadter nega, con una serie di ragionamenti ardimentosi e affascinanti, il fatto che viceversa ciascuno può sperimentare, che la mente umana non si arresta davanti all'indecidibilità gödeliana. Le repliche al paradosso della stanza cinese ipotizzano che non vi sia alcuna distinzione fra sintassi e semantica. E, sul piano pratico, l'abilità "estetica" del maestro di scacchi è declassata alla laboriosa ricerca del percorso giusto in un albero di decisioni incredibilmente intricato. In genere, queste posizioni privilegiano l'attività razionale-deduttiva della mente rispetto ad altre facoltà. In particolare ignorano l'intelligenza "intuitiva", della quale pur dovrebbe far nascere qualche sospetto il caso del giocatore a scacchi. C'è, è vero, qualche generoso tentativo di superare la barriera della *ratio* (nel senso etimologico di "calcolo") e la conseguente aporia del teorema d'incompletezza. Simili tentativi si possono riscontrare specialmente in Searle, e in Penrose che ipotizza una "razionalità non algoritmica", ma si esauriscono subito: forse per un'insufficiente apertura alla prospettiva metafisica, ma certamente anche a causa dell'irosa reazione degli oppositori, che costringe a una faticosa difesa su posizioni di retroguardia.

Vogliamo qui citare, nell'articolo di Searle prima menzionato, un passo illuminante perché mette in luce il nodo vero della questione, e cioè la diversa "natura" del cervello e della macchina, del "naturale" e dell'"artificiale": «Le simulazioni al calcolatore dei processi cerebrali forniscono modelli degli aspetti formali di questi processi, ma la simulazione non va confusa con la riproduzione. Il modello computazionale dei processi mentali non è più reale di quello di qualsiasi altro fenomeno naturale. Si può immaginare una simulazione al calcolatore precisa fino all'ultima sinapsi dell'azione dei peptidi sull'ipotalamo. Ma si può del pari immaginare una simulazione al calcolatore dell'ossidazione degli idrocarburi in un motore d'automobile o dei processi di digestione in uno stomaco alle prese con una pizza. Nel caso del cervello la simulazione non è più reale che nel caso dell'automobile e dello stomaco. A meno che non avvenga un miracolo, non potremmo far marciare la nostra macchina grazie a una simulazione al calcolatore dell'ossidazione della benzina né potremmo digerire la pizza eseguendo il programma che simula tale digestione. Sembra altrettanto ovvio che la simulazione di un processo cognitivo non produca gli stessi effetti della neurobiologia di quel processo cognitivo» (Searle, 1990, p. ***). Fuori da ogni *vis polemica*, l'IA è soltanto un «modello di simulazione» dell'intelligenza naturale (e, per ora, soltanto di alcuni suoi aspetti), utilissimo ai fini pratici come tutti i modelli di simulazione, ma nulla di più; e, se pur dovesse dar segno di qualcosa di simile alla coscienza o all'intenzionalità, dovremmo dire che si tratta soltanto di una simulazione della coscienza e dell'intenzionalità. Così l'intelligenza artificiale assume il ruolo di un «segno di contraddizione» per la ricerca scientifica. Essa rivela, da un lato, il carattere "meta-scientifico" della scelta che Fodor icasticamente chiama "materialistica" e la vanità del tentativo di giustificarla mantenendosi entro i termini delle scienze positive, che è stato riportato poc'anzi. D'altra parte essa mostra i limiti intrinseci di tale scelta, potenzialmente fonte di insolubili aporie e di risultati in contrasto con

l'esperienza: limiti che altre posizioni, non programmaticamente chiuse ad una prospettiva metafisica, potrebbero forse superare.

2. Che cosa l'IA non deve fare. Il problema morale, dei limiti entro i quali e dei modi in cui è lecito far uso delle tecniche di intelligenza artificiale, altro non è che una particolarizzazione del problema generale del corretto uso degli strumenti tecnologici. Ma qui esso assume una connotazione speciale in quanto sembra trattarsi di attribuire alla macchina scelte che toccherebbero all'uomo. Si pensi, per esempio, alle implicazioni bioetiche di terapie automatiche "decise" da un sistema esperto. In altre parole, mentre in generale l'impiego di strumenti tecnici incontra limitazioni quantitative (per esempio: «non si deve usare troppa energia perché ciò esaurisce le riserve e devasta l'ambiente») qui la limitazione apparirebbe piuttosto di carattere qualitativo, giacché si tratterebbe di sapere "quali" tipi d'intervento è lecito affidare all'intelligenza artificiale e quali no. Probabilmente la questione non deve essere drammatizzata. Se valesse il paradigma riduzionistico per cui la macchina sarebbe dotata anche di intenzionalità, allora la delega sarebbe profondamente inquietante. Ma se invece, più ragionevolmente, si ricorda che la macchina è programmata dall'uomo e dipende da lui anche quando è stata dotata di algoritmi "genetici" che sviluppano la programmazione in maniera non preordinata, allora il problema si riduce a stabilire in qual misura affidarsi alla "protesi intellettuale" per progettare una terapia, o un qualche intervento di rilevante peso economico o sociale.

Il problema ritorna così ad essere quello di una questione ancora in qualche modo quantitativa, ossia di prudente uso dello strumento tecnico. Ma non si può negare che la disponibilità di programmi che "decidono per lui" potrebbe indurre l'operatore ad assumere un atteggiamento meno responsabile ed a rinunciare alla sua responsabilità per "delegarla" alla macchina: nel qual caso non si potrebbe più dire che la macchina dipende dall'uomo. In questo senso si può osservare, in termini generali, che questi strumenti apparentemente dotati di "capacità autonome" pongono in forma più netta la necessità, già avvertita per esempio da Romano Guardini (cfr. *La fine dell'epoca moderna. Il potere*, Brescia 1989[2], p. 88) che l'uomo abbia in ogni momento il pieno dominio "morale" sui sistemi tecnologici.

Paola Mello

Bibliografia:

Aspetti prevalentemente tecnici dell'IA: W.S. MCCULLOCH, W. PITTS, *A logical Calculus of the ideas immanent inneural nets,* "Bullettin of Mathematical Biophysics" 5 (1943), pp. 115-137; F. ROSENBLATT, *Principles of Neurodynamics: Perceptrons and the theory of Brain Mechanisms,* Spartan Books, Washington D.C. 1962; J.A. ROBINSON, *A Machine-Oriented Logic based on the resolution principle*, "Journal of ACM" 12 (1965), n. 1, pp. 23-41; M. MINSKY, S. PAPERT, *Perceptrons,* MIT Press, Cambridge (MA) 1969;

E.H. SHORTLIFFE, *Computer Based Medical Consultation*: *Mycin*, Elsevier, North-Holland, New York 1976; R. DAVIS, B. BUCHANAN, E. SHORTLIFFE, *Production rules as a representation for knowledge-based consultation program*, "Artificial Intelligence" 8 (1977), pp. 15-45; H.A. SIMON, *The Sciences of Artificial*, MIT Press, Cambridge (MA) 1981; R. DAVIS, D.B. LENAT, *Knowledge-based Systems in Artificial Intelligence*, McGrow-Hill, New York 1982; S. GROSSBERG, *Learning by neural networks in studies of main and brain*, Reidel, Boston 1982; R.S. MICHALSKI, J.G. CARBONELL, T.M. MITCHELL (a cura di), *Machine Learning: An artificial intelligence approach*, Springer, Berlin 1984; J.R. QUINLAN, *Induction of decision trees*, "Machine Learning" 1 (1986), n. 1, pp. 81-106; Ivan BRATKO, *Programmare in Prolog per l'Intelligenza Artificiale*, Masson e Addison-Wesley, Readings, Massachusetts 1988; E. CHARNIAC, D. MCDERMOTT, *Introduzione all'Intelligenza Artificiale*. Addison-Wesley, Readings, Massachusetts 1988; M. MINSKY, *The Society of Mind*, Touchstone Editions, New York 1988; D. GOLDBERG, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Readings, Massachusetts 1989; J.A. FREEMAN, D.M. SKAPURA, *Neural networks algorithms, applications and programming techniques*. Addison-Wesley, Readings, Massachusetts 1991; J.A. HERTZ, A. KROGH, R.G. PALMER*, Introduction to the theory of neural computation*. Addison-Wesley, Readings, Massachusetts 1991; U.M. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, R. UTHURUSAMY, *Advances in knowledge discovery and data mining*, MIT Press, Cambridge Massachuetts 1992; K. KNIGHT, E. RICH, *Intelligenza Artificiale*, McGraw Hill, New York 1992[2]; R. PENROSE, *La mente nuova dell'imperatore*, Rizzoli, Milano 1992; P.H. WINSTON, *Artificial Intelligence,* Addison-Wesley, Readings, Massachusetts 1992; M. GINSBERG, *Essentials of Artificial Intelligence*, Morgan Kaufman, San Mateo, California 1993; J. DOYLE, T. DEAN, *Strategic Directions in Artificial Intelligence*, in "ACM Computing Surveys"*,* 28 (1996), n. 4, p. 653-670;

D.R. HOFSTADTER e il "Gruppo di ricerca sulle analogie fluide", *Concetti fluidi e analogie creative*. *Modelli per calcolatore dei meccanismi fondamentali del pensiero,* Adelphi, Milano 1996 (spostare su aspetti interdisciplinari??);

L. CONSOLE, E. LAMMA, P. MELLO, M. MILANO, *Programmazione Logica e Prolog*, UTET, Torino 1997[2]; T.M. MITCHELL, *Machine Learning*, McGraw Hill, New York 1997; N.R. JENNINGS, M.J. WOOLDRIGE (a cura di), *Agent Technology*, Springer Verlag, Berlin 1998; S.J. RUSSEL, P. NORVIG, *Intelligenza Artificiale: Un approccio moderno*, Prentice Hall International - UTET, Torino 1998; AA.VV., *Sistemi intelligenti,* n. 1, anno XII (aprile), Il Mulino, Bologna 2000.

Aspetti interdisciplinari: A. TURING, *Computing Machinery and Intelligence*, "Mind" 49 (1950), pp. 433-460; J. LUCAS, *Minds, Machines and Gödel*, in "Philosophy" 37 (1961), pp. 37-39; M. BUNGE, *The Mind-Body*

Problem. A Psychobiological Approach, Oxford Univ. Press, Oxford 1980; J.A. FODOR, *Il problema mente-corpo*, "Le Scienze", n. 151, *** 1981, pp. ***-***; E. NAGEL, J. R. NEWMAN, *La prova di Gödel,* Boringhieri, Torino 1982; K. POPPER, J. ECCLES, *L'Io e il suo cervello,* 3 voll., Armando, Roma 1981; J.-P. CHANGEAUX, *L'uomo neuronale*, Feltrinelli, Milano 1983; D.R. HOFSTADTER, *Gödel, Escher e Bach: un'eterna ghirlanda brillante,* Adelphi, Milano 1984; H. DREYFUS, *Che cosa non possono fare i computer? I limiti dell'intelligenza artificiale*, Armando, Roma 1988; J. SEARLE, *La mente è un programma?,* "Le Scienze", n. 259, marzo 1990, pp. ***-***; P. e P. CHURCHLAND, *Può una macchina pensare?,* in *ibidem*, pp. ***-***; G. BASTI, *Il rapporto mente-corpo nella filosofia e nella scienza,* ESD, Bologna 1991; J. SEARLE, *La riscoperta della mente*, Bollati-Boringhieri, Torino 1992; G. PICCININI, *Su una critica dell'intelligenza artificiale «forte»*, "Rivista di filosofia" 85 (1994), n. 1, pp. ***-*** [p. 141]; R. PENROSE, *Ombre della mente. Alla ricerca della coscienza,* Rizzoli, Milano 1996; A. ROSSI, *Il fantasma dell'intelligenza,* CLUEN, Napoli 1998; F. BERTELÈ, A. OLMI, A. SALUCCI e A. STRUMIA, *Scienza, analogia, astrazione. Tommaso d'Aquino e le scienze della complessità,* Il Poligrafo, Padova 1999.