

# L'Intelligenza Artificiale: la storia e le idee

di *Roberto Cordeschi e Guglielmo Tamburrini*

Scopo di questo capitolo è di introdurre il lettore ai principali argomenti che, nel corso della sua breve storia, l'Intelligenza Artificiale (IA) ha affrontato sia nella variante applicativa o ingegneristica sia in quella teorica o cognitiva. Alla fine di questo capitolo il lettore dovrebbe

- avere presente l'evoluzione di alcune delle tendenze di ricerca più influenti in IA;
- essere al corrente delle più recenti posizioni nel dibattito attuale all'interno dell'IA;
- essere brevemente introdotto ad alcuni classici problemi filosofici ed epistemologici affrontati dall'IA.

## I.1

### Introduzione

È difficile dare una definizione dell'IA e dei suoi obiettivi che sia unanimemente condivisa dai ricercatori. L'origine della difficoltà sta anche nel fatto che da sempre l'IA si è presentata sotto un duplice profilo: quello di disciplina *ingegneristica*, il cui obiettivo è di costruire macchine in grado di assistere l'uomo, e magari di competere con esso, in compiti soprattutto intellettuali, e quello di disciplina *psicologica*, il cui obiettivo è di costruire macchine le quali, riproducendo da vicino le caratteristiche essenziali dell'attività cognitiva umana, gettino una nuova luce su alcuni tradizionali enigmi della mente, ad esempio sul cosiddetto problema mente-corpo. Forse la base programmatica dell'IA più comunemente accettata è ancora quella utilizzata nella presentazione del seminario organizzato da John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon nel giugno del 1956 negli Stati Uniti, a Dartmouth (New Hampshire), nella quale si legge:

In linea di principio si può descrivere ogni aspetto dell'apprendimento e dell'intelligenza con una precisione tale da permetterne la simulazione con mac-

chine appositamente costruite. Si cercherà di costruire macchine in grado di usare il linguaggio, di formare astrazioni e concetti, di migliorare se stesse e risolvere problemi che sono ancora di esclusiva pertinenza degli esseri umani.

Nel corso di quello storico seminario si gettarono le basi dell'IA, individuando alcune aree di ricerca rimaste classiche e presentando i primi programmi per calcolatore cosiddetti "intelligenti". Le macchine alle quali pensavano i pionieri dell'IA sono i calcolatori digitali. Come vedremo nel PAR. 1.2, è possibile circoscrivere con una certa precisione alcune proprietà fondamentali di queste macchine, proprietà che ci aiutano anche a comprendere perché di esse si è parlato come di "macchine intelligenti". Ricorderemo varie caratteristiche dei calcolatori fisicamente realizzati e di loro modelli matematici, facendo riferimento a risultati cruciali della teoria della calcolabilità, le cui basi vennero gettate da alcuni logici nell'arco del quinquennio che va dal 1931 al 1936 – e cioè *prima* che i moderni calcolatori digitali venissero effettivamente costruiti.

I pionieri della nuova disciplina videro nel calcolatore digitale uno strumento con capacità di elaborazione simbolica ineguagliate da qualsiasi macchina sviluppata nella storia dell'umanità e per questo motivo adatta al confronto con alcuni aspetti tra i più peculiari dell'intelligenza umana. Nei giorni di Dartmouth, pronto per "girare" su calcolatore era un programma ideato da Newell, Shaw e Simon, l'ormai mitico Logic Theorist, che dimostrava alcuni teoremi del calcolo proposizionale. Ma bisogna ricordare almeno un altro programma la cui sperimentazione era cominciata qualche tempo prima, quello di Arthur Samuel per il gioco della dama, in grado di migliorare le proprie prestazioni al punto di arrivare a battere ottimi giocatori. Altri programmi sarebbero presto seguiti: dalla Geometry Machine di Herbert Gelertner e Rochester ai primi programmi per gli scacchi, stimolati da alcune precedenti intuizioni di Shannon (Somenzi, Cordeschi, 1994).

I protagonisti della prima IA si concentrarono su problemi relativi ad ambiti ben delimitati, per risolvere i quali bastano regole esplicite per l'elaborazione simbolica e poca conoscenza specializzata. In questa scelta iniziale sono confluiti sia motivi pratici, come la modesta potenza di calcolo dei calcolatori dell'epoca, sia motivi più teorici, come la diffusa concezione che l'elaborazione simbolica sia la vera marca dell'intelligenza. Ma la descrizione degli intenti dell'IA sopra ricordata si rivela in questo caso opportunamente generica, poiché non assegna esplicitamente all'elaborazione simbolica un ruolo privilegiato. Come vedremo nel PAR. 1.3, programmi di ricerca ispirati ad altre concezioni dell'intelligenza sono stati portati avanti nell'ambito dell'IA, spostando l'accento

dall'elaborazione simbolica alla conoscenza tacita, alle abilità sensorie, alle capacità di adattamento all'ambiente naturale o alle interazioni sociali con altri agenti naturali o artificiali.

Programmi di ricerca così diversi tra loro sono emersi anche in reazione agli ostacoli che questa disciplina ha incontrato lungo un percorso non lineare, per quanto costellato da notevoli successi. La diffusione capillare di tecniche e sistemi dell'IA in molti prodotti della tecnologia avanzata testimonia il successo dell'IA intesa come disciplina ingegneristica. Ma quale bilancio possiamo trarre a proposito della costruzione di macchine che riproducano caratteristiche essenziali dell'attività cognitiva umana? Cosa aggiunge alla nostra conoscenza del mentale un sistema dell'IA che simula con successo un essere umano nello svolgimento di un certo compito cognitivo? E che posto occupa l'IA, così intesa, nell'ambito di ricerche interdisciplinari sul sistema cervello-mente che coinvolgono anche le neuroscienze e la psicologia? È di questi problemi epistemologici che ci occuperemo nel PAR. 1.4, mentre delle discussioni relative alla natura degli stati cognitivi, che hanno accompagnato lo sviluppo dei sistemi dell'IA, ci occuperemo nel PAR. 1.5. In particolare, vedremo come le tesi del funzionalismo, della realizzabilità di sistemi intelligenti con sostrati materiali differenti, dell'indipendenza di vari livelli di analisi e spiegazione del mentale emergono nelle riflessioni dei ricercatori dell'IA prima di ricevere una formulazione canonica da parte dei filosofi della mente. Infine, toccheremo brevemente i problemi delle emozioni e dei cosiddetti *qualia* e il problema "difficile" della coscienza intesa come esperienza soggettiva, che resta un enigma per l'IA così come per ogni altra impostazione scientifica allo studio della mente.

## 1.2

### **Macchine astratte e calcolatori potenzialmente universali**

I calcolatori digitali sono dunque le macchine prescelte per realizzare i sistemi dell'IA. Nella forma di calcolatori "generali" (*general purpose*), essi hanno la fondamentale caratteristica di essere macchine *simboliche*, che non eseguono solo un repertorio limitato di calcoli numerici, come le ordinarie macchine calcolatrici. Con i calcolatori *general purpose* si comincia a disporre per la prima volta di una macchina capace di manipolare strutture di simboli che il programmatore fa corrispondere in modo naturale alle entità più diverse: parole di una lingua naturale, espressioni matematiche, posizioni del gioco degli scacchi, oggetti da riconoscere e classificare. Questa capacità di elaborazione simbolica, insieme alla presenza dell'istruzione di "salto condizionato" nei pro-

grammi, sollecitò da subito a parlare di “macchine intelligenti”. L’istruzione di salto condizionato consente di cambiare l’ordine di esecuzione delle istruzioni: *se* una data condizione è soddisfatta, *allora* vengono effettuate operazioni specificate da una certa parte del programma, *altrimenti* ne vengono eseguite altre, specificate da una diversa parte del programma. Il salto condizionato conferisce a un programma una certa capacità *discriminativa*, che sfocia nella selezione della sequenza di istruzioni da eseguire.

Nell’EDSAC le due caratteristiche che abbiamo ricordato erano pienamente realizzate (cfr. la scheda 1.1 che ricorda alcune tappe dell’evoluzione dei primi calcolatori). L’EDSAC, realizzato nel 1949, era il primo grande calcolatore con *programma memorizzato*: nella sua memoria interna erano depositati non solo i dati, ma anche le istruzioni per manipolarli, ovvero il programma, che poteva essere modificato non meno dei dati mediante operazioni algoritmiche. Era questa l’architettura dei calcolatori che John von Neumann aveva sintetizzato nel celebre *First Draft* del 1945 e che sarebbe rimasta sostanzialmente immutata negli anni a venire (un comune calcolatore da tavolo è una macchina di questo tipo).

I programmi che determinano i processi di elaborazione dei calcolatori digitali sono *procedimenti algoritmici* specificati in un qualche linguaggio di programmazione dato. La nozione generale di programma dipende dunque da quella di procedimento algoritmico, che è del tutto intuitiva e dai contorni sfumati – per quanto sia sufficientemente chiara da consentire a tutti di riconoscere e di eseguire vari procedimenti algoritmici, per calcolare almeno funzioni molto semplici, come la somma o il prodotto di numeri naturali. Alcune proprietà fondamentali dei procedimenti algoritmici sono state tuttavia individuate con precisione, soprattutto grazie alle ricerche di vari logici, come Alonzo Church, Kurt Gödel, Stephen Kleene, Emil Post e Alan Turing, compiute prima degli sviluppi ricordati nella scheda 1.1. In particolare, attraverso la cosiddetta tesi di Church-Turing, enunciata indipendentemente da Church e Turing tra il 1935 e il 1936, è stata fornita una caratterizzazione precisa e generalmente ritenuta soddisfacente delle *funzioni* calcolabili mediante procedimenti algoritmici. Nella versione direttamente collegata al lavoro di Turing (1937), la tesi asserisce che *ogni funzione calcolabile mediante un procedimento algoritmico è calcolabile da una macchina di Turing*. Dunque, per la tesi di Church-Turing la classe delle funzioni calcolabili mediante procedimenti algoritmici è inclusa nella classe delle funzioni calcolabili dalle macchine di Turing. Quest’ultima classe, al contrario della prima, ha una definizione precisa, che è basata a sua volta sulla definizione di *macchina di Turing*.

SCHEDA I.1

I primi calcolatori (1941-51)

1941 Z<sub>3</sub> di K. Zuse. Il primo calcolatore programmabile digitale e *general purpose* effettivamente costruito. Era elettromeccanico, con una memoria composta di 2.600 relè telefonici. Poteva convertire i decimali in numerazione binaria e viceversa. Utilizzava un nastro magnetico perforabile per il programma in ingresso.

1943 Primo dei calcolatori COLOSSUS di T. H. Flowers, W. W. Chandler e altri. Era un grande calcolatore elettronico, con una memoria composta di 1.500 valvole. Usava cinque processori paralleli e poteva leggere un nastro perforato alla velocità di 5.000 caratteri al secondo.

1944 Harvard Mark I di H. Aiken. Era elettromeccanico e usava la numerazione binaria. Era specializzato in problemi di tipo matematico, ma aveva già il programma memorizzato.

1946 ENIAC di J. P. Eckert e J. W. Mauchly. Il primo grande e veloce calcolatore elettronico, *general purpose* e programmabile. Era composto di più di 18.000 valvole e utilizzava un sistema di calcolo parallelo.

1949 EDSAC (Electronic Delay Storage Automatic Computer) di M. Wilkes, W. Renwick, D. J. Wheeler e collaboratori. Il primo calcolatore elettronico *general purpose* con programma memorizzato.

MADM (Manchester Automatic Digital Machine) di F. Williams, T. Kilburn e M. H. A. Newman (con la collaborazione di A. M. Turing per gli aspetti relativi alla programmazione). Da questa macchina prototipo fu costruito nel 1951, in collaborazione con la Ferranti Ltd., il Ferranti Mark I.

BINAC (Binary Automatic Computer) di J. P. Eckert e J. W. Mauchly. Era di tipo *general purpose*, il primo calcolatore costruito da privati fuori da centri di ricerca.

1950 ACE (Automatic Computing Engine) di G. G. Alway, D. Davies, J. H. Wilkinson e M. Woodger (progetto di A. M. Turing del 1945). Costruito presso il National Physical Laboratory di Teddington, era *general purpose* e usava la numerazione binaria

1951 EDVAC (Electronic Discrete Variable Automatic Computer) di R. L. Snyder, S. E. Gluck e W. H. Boghosian. Il progetto iniziale, tra gli altri di J. von Neumann, lo descriveva come un calcolatore a programma memorizzato. Era composto di 3.600 valvole e usava schede perforate.

UNIVAC (Universal Automatic Computer) di J. P. Eckert e J. W. Mauchly. Il primo grande calcolatore elettronico prodotto per scopi commerciali.

Una macchina di Turing è un calcolatore *idealizzato*, in quanto si suppone che essa non sia soggetta ad alcune limitazioni proprie dei calcolatori fisicamente realizzati: non si impone un limite alle sue capacità di memoria (la memoria di una macchina di Turing è costituita da un nastro *indefinitamente espandibile*, uniformemente diviso in caselle, ciascuna delle quali può contenere al più un simbolo da un alfabeto finito e prefissato); non si impone nemmeno un limite alla durata dei processi di calcolo e si assume che la macchina funzioni sempre perfettamente. Le operazioni eseguibili da una macchina di Turing sono estremamente elementari: dotata di una testina di lettura per osservare una casella per volta, una macchina di Turing che si trova in una determinata configurazione interna può stampare un simbolo sulla casella osservata o cancellare il simbolo già stampato su quest'ultima, spostare la testina di lettura di una casella a sinistra o a destra della casella osservata e cambiare la propria configurazione interna. Quali di queste operazioni vengono eseguite a ogni dato istante dipende esclusivamente dal contenuto della casella osservata e dalla configurazione interna della macchina.

Gli aspetti fondamentali di questa descrizione informale e intuitivamente visualizzabile di una macchina di Turing possono essere espressi in un linguaggio più astratto e rigoroso. Si consideri la lista di simboli  $s_1, s_2, s_3, \dots; q_1, q_2, q_3, \dots; D, S$ . Un'espressione è una successione finita di simboli presi da questa lista. Un'istruzione è una *quintupla*, cioè un'espressione della forma

$$q_i s_j s_k q_l M$$

dove M sta per la lettera D o per la lettera S.

Intuitivamente, un'istruzione si interpreta in questo modo: se la configurazione interna è  $q_i$  e il simbolo osservato è  $s_j$ , allora la macchina scriverà al suo posto il simbolo  $s_k$  entrando nella configurazione interna  $q_l$  e spostandosi di una casella, a destra o a sinistra, a seconda che M sia D oppure S. Si conviene generalmente che il simbolo  $s_1$  indichi la casella vuota; pertanto rimpiazzare  $s_j$  con  $s_1$  equivale all'azione di cancellare  $s_j$ .

*Definizione:* Una macchina di Turing è un insieme finito (ma non vuoto) di quintuple che non contiene nessuna coppia di quintuple che coincidano nei loro primi due simboli.

Questa definizione impone cruciali condizioni di *finitezza* (relative, per esempio, al numero di istruzioni, all'alfabeto di simboli sul quale opera la macchina e alle sue possibili configurazioni interne) e di *determinatezza* (ad ogni istante dato una macchina di Turing potrà applicare al più

un'istruzione). Essa identifica una macchina di Turing con un programma formato da istruzioni scritte in una forma standard. È anche possibile dare una descrizione rigorosa di quali siano, ad ogni passo di esecuzione di un tale programma, la casella osservata, i contenuti del nastro e la configurazione interna della macchina (per una trattazione esauriente cfr. Davis, 1982, pp. 5-8). Una semplice macchina di Turing che esegue la somma di due numeri naturali è descritta nella scheda 1.2.

Alcune macchine di Turing sono dette *universali*, perché in grado di simulare fedelmente il comportamento di una qualsiasi macchina di Turing dato un qualsiasi ingresso. Poiché una macchina universale di Turing può calcolare tutte le funzioni calcolabili da ogni macchina di Turing, dalla tesi di Church-Turing segue che una tale macchina universale può calcolare ogni funzione calcolabile da un calcolatore digitale. Tra i calcolatori *general purpose* fisicamente realizzati, anche i calcolatori da tavolo e i portatili più comuni hanno una stretta relazione con le macchine universali di Turing. Facendo astrazione da possibili guasti e dalle limitazioni relative alla memoria e alla durata dei calcoli, i nostri calcolatori da tavolo hanno la stessa potenza di calcolo di una macchina universale di Turing. Per questo motivo siamo autorizzati a chiamarli calcolatori *potenzialmente universali*. Pertanto si preserva la massima generalità rispetto alla classe di funzioni calcolabili se si sceglie, per simulare dei comportamenti intelligenti, una macchina potenzialmente universale. Questo punto è stato fortemente sottolineato agli albori dell'IA dallo stesso Turing (1950) e rielaborato con originalità da Newell e Simon attraverso la nozione di sistema fisico di simboli (cfr. PAR. 1.5).

Il lavoro che portò alla tesi di Church-Turing fu principalmente motivato dal problema di generalizzare i teoremi di incompletezza, scoperti da Gödel nel 1931. Nella forma generalizzata ottenuta facendo ricorso alla tesi di Church-Turing, il primo teorema di Gödel implica che non vi è nessun procedimento algoritmico che consenta di generare *tutti* gli enunciati veri dell'aritmetica elementare senza generare al contempo *nessun* enunciato falso. In altre parole, se un procedimento algoritmico consente di operare in modo *corretto* nell'ambito dell'aritmetica elementare (cioè non permette di generare enunciati falsi), allora esso risulta essere *incompleto* (cioè non permette di generare tutti gli enunciati aritmetici veri). Il problema di generare tutte e sole le proposizioni vere dell'aritmetica elementare non è risolvibile, nemmeno in linea di principio, mediante un procedimento algoritmico. Si è ampiamente discusso se gli esseri umani siano soggetti ad analoghe limitazioni e, in caso di risposta negativa, se ciò ponga dei limiti alle ambizioni dell'IA. Su questo punto torneremo brevemente nel PAR. 1.5.

## SCHEDA 1.2

## Una macchina di Turing

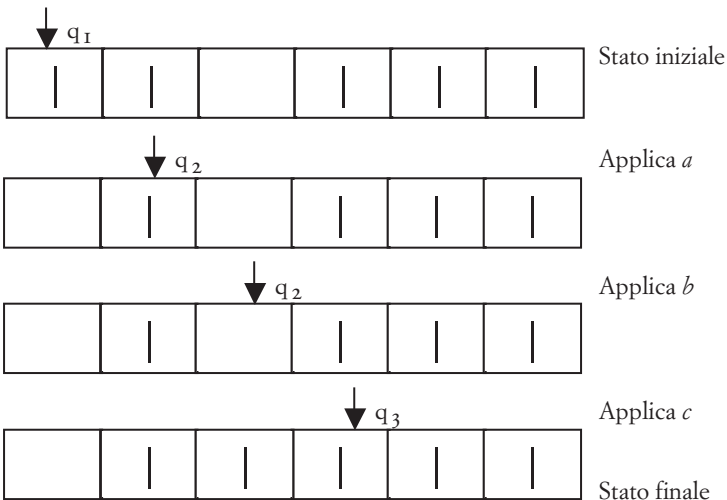
La macchina di Turing

a)  $q_1 \mid s_1 q_2 D$

b)  $q_2 \mid \mid q_2 D$

c)  $q_2 s_1 \mid q_3 D$

esegue la somma di due interi positivi  $n$  ed  $m$  se si rispettano le seguenti convenzioni: a  $n$  ed  $m$  si associano le espressioni nastro formate rispettivamente da  $n$  ed  $m$  barrette consecutive, separate tra loro da una sola casella vuota, denotata nelle istruzioni a) e c) con il simbolo  $s_1$ , e poste su nastro altrimenti vuoto; la macchina viene avviata nello stato  $q_1$  osservando la casella non vuota posta più a sinistra sul nastro.



La macchina si arresterà nello stato  $q_3$  con un nastro altrimenti vuoto che contiene  $n+m$  barrette consecutive, cioè l'espressione nastro corrispondente all'intero positivo  $n+m$ .

I passi di calcolo sull'ingresso formato dalle espressioni nastro corrispondenti ai numeri 2 e 3 sono indicati nella figura (la freccia indica la posizione delle testine di lettura). All'inizio, la macchina si trova nello stato  $q_1$  e osserva il simbolo  $|$ . Deve dunque cancellarlo (scrivere al suo posto il simbolo  $s_1$ ), passare nello stato  $q_2$  e spostarsi a destra. Il risultato di queste operazioni è descritto nella seconda situazione della figura, gli altri nelle successive situazioni.

## I.3

**Elaborazione simbolica e intelligenza**

La partizione in problemi risolvibili o non risolvibili mediante un procedimento algoritmico è ancora troppo grossolana per gli scopi dell'IA. Nel caso di molti problemi risolvibili “in linea di principio”, accade che le risorse di calcolo da impiegare per arrivare alla soluzione desiderata eccedano la potenza di un qualsiasi calcolatore già realizzato o perfino fisicamente realizzabile. In questi casi, la locuzione “in linea di principio” indica una mera possibilità matematica, che non implica possibilità fisica e tanto meno la concreta realizzabilità di sistemi in grado di risolvere tali problemi.

La teoria della complessità algoritmica o computazionale studia le risorse di calcolo necessarie per la soluzione dei problemi (per un'introduzione pertinente ai temi ai quali qui accenniamo cfr. Burattini, 1993). Si sono rivelate proibitive, in base alle stime della teoria della complessità, le risorse di calcolo necessarie per risolvere problemi che hanno attratto da subito l'attenzione dei pionieri dell'IA, come quello di determinare una strategia ottimale in una partita a scacchi. È stato sostenuto che questi risultati pongono un ostacolo teorico alle ambizioni dell'IA. Ma i ricercatori di questa disciplina hanno adottato un punto di vista molto diverso da quello della teoria della complessità, che ha permesso loro di aggirare il presunto ostacolo in molti casi significativi.

I primi programmi dell'IA, come abbiamo ricordato, elaborano *strutture di simboli* e, soprattutto attraverso sequenze di applicazioni di regole di salto condizionato, simulano almeno i prodromi di una capacità ritenuta peculiare dell'intelligenza umana: quella di scegliere, davanti a un problema che dà luogo all'esplosione combinatoria delle mosse lecite, solo *alcune* sequenze di mosse che potrebbero portare alla soluzione. Classico è il caso del giocatore di scacchi, che all'inizio della partita ha di fronte a sé un numero ultra-astronomico di possibili mosse alternative, che Shannon calcolò nell'ordine di  $10^{120}$ . L'intelligenza fu primariamente identificata con questa capacità *selettiva* della mente umana, del resto ben documentata dalla psicologia dei processi del pensiero. Fu proprio con quest'ultima che l'IA delle origini venne a confrontarsi, sia nelle vesti della *Information Processing Psychology*, o “psicologia dei processi dell'informazione”, che aspirava a simulare i processi cognitivi umani in modo psicologicamente realistico (cfr. PAR. 1.4), sia nella costruzione di programmi per calcolatore che, senza essere psicologicamente realistici, avessero comunque prestazioni *efficienti*. Il successo dei primi programmi alimentò la convinzione che compito

principale dell'IA fosse lo studio delle strategie di soluzione di problemi efficacemente selettive, o "euristiche".

La *programmazione euristica*, il settore di ricerca nel quale l'IA raccolse i suoi primi successi, ha risposto alle sfide della complessità algoritmica tenendo conto delle strategie effettivamente adottate dagli esseri umani. Questi ultimi, dovendo fornire risposte adeguate ai problemi in tempi ragionevoli, per lo più evitano di condurre una ricerca "cieca" o basata sulla forza bruta di soluzioni *ottimali* a un problema dato, limitandosi a esplorare solo una parte dei percorsi che, in base alle informazioni in loro possesso, potrebbero portare a una soluzione accettabile del problema dato. Ci si comporta in questo modo in tante situazioni della vita di tutti i giorni (quando, poniamo, si sceglie una scuola per i propri figli o una località per le vacanze), accontentandosi di individuare una soluzione che soddisfi alcuni requisiti irrinunciabili. Ma così facendo si corre il rischio che *tutte* le soluzioni soddisfacenti al problema dato non si trovino lungo i percorsi prescelti; per questo motivo, in IA si pone grande cura nello sviluppo di metodi per individuare i percorsi risolutivi più promettenti.

Gli algoritmi euristici dell'IA non sono ispirati solo al comportamento cognitivo degli esseri umani (come la strategia "mezzi-fine" introdotta all'origine dell'IA da Allen Newell e Herbert Simon: cfr. PAR. 1.4 e più in generale sulle euristiche il CAP. 2). L'evoluzionismo darwiniano, con le nozioni di mutazione e selezione, ha ispirato gli algoritmi evolutivi, applicati sia in IA (cfr. CAP. 5) sia nella cosiddetta Vita Artificiale (Langton, 1995). Vi sono inoltre euristiche che prendono lo spunto da comportamenti individuali o collettivi osservati nel mondo animale, come le strategie cooperative di ricerca e segnalazione di fonti di cibo messe in opera da colonie di formiche (Dorigo, Gambardella, 1997).

La programmazione euristica ha continuato comunque a costituire un settore di ricerca importante dell'IA. Tipicamente fondati sulla programmazione euristica sono i *sistemi basati sulla conoscenza*, e tra questi i cosiddetti *sistemi esperti* (descritti nel CAP. 11), che generalmente richiedono un supporto hardware più potente di quello dei primi calcolatori. Un sistema esperto risolve problemi in un ambito specialistico (ad esempio in un qualche settore diagnostico) con prestazioni, almeno nei casi riusciti, comparabili a quelle di un esperto umano del settore. Una metodologia comunemente impiegata nella costruzione di un sistema esperto prevede anzitutto la registrazione di conoscenze e di tecniche di ragionamento euristico dell'esperto umano. Tali conoscenze sono registrate, sia pure in forme drasticamente semplificate, in strutture dati chiamate *basi di conoscenza*, che generalmente assumono vaste dimensioni. Sulla base di conoscenza opera un motore inferenziale, che risol-

ve i problemi posti anche simulando tecniche di ragionamento euristico dell'esperto umano.

Nei grandi sistemi basati sulla conoscenza, al di là di differenze, anche marcate, negli stili di *rappresentazione* ed elaborazione della conoscenza (sui quali cfr. CAP. 3), si riconosce uno dei prodotti principali di quella che viene comunemente chiamata "IA simbolica". Un bilancio equilibrato del lavoro svolto in questo ambito impone di ricordare, ma senza nessuna pretesa di esaustività, alcuni problemi che le metodologie soggiacenti ai sistemi basati sulla conoscenza non permettono di affrontare adeguatamente.

1. *Conoscenza tacita*. Varie forme di comportamento intelligente sono guidate da ciò che chiamiamo comunemente il "saper fare", cioè da un tipo di abilità che appare difficile e spesso innaturale descrivere attraverso conoscenze dichiarative e regole esplicite di manipolazione simbolica.

2. *Azione in tempo reale*. Gli algoritmi euristici dell'IA simbolica, anche se soddisfacenti rispetto a ricerche esaustive o a quelle con un grado controllato di approssimazione all'ottimalità, non sempre riescono a generare risposte adeguate in tempi utili. Ostacoli del genere sono stati incontrati, per esempio, nella progettazione di sistemi esperti per la diagnosi di guasti in situazioni critiche (cfr. CAP. 11) o di sistemi per la pianificazione delle azioni di un robot in ambienti poco prevedibili (cfr. CAP. 5).

3. *Robustezza*. I sistemi che operano in base a rappresentazioni esplicite della conoscenza si rivelano fragili, spesso incapaci di operare in situazioni diverse da quelle previste dai progettisti. Piccole discrepanze tra lo stato percepito del mondo e le condizioni previste per l'esecuzione di un procedimento algoritmico talvolta portano a un rapido degrado delle prestazioni. Ciò non si osserva negli esseri umani e in altre specie animali, le cui prestazioni sono caratterizzate da una buona tolleranza al rumore e alle variazioni ambientali.

Negli ultimi venti anni, attraverso il *connessionismo*, la *nuova robotica* e la *cognizione situata*, la tematizzazione di queste e altre difficoltà ha permesso di delineare programmi di ricerca che si collocano all'interno di quella che viene talvolta chiamata "nuova IA". Ci limitiamo a indicarne qui alcune caratteristiche generali, rimandando a capitoli successivi di questo volume (in particolare ai CAPP. 4 e 5) per la descrizione dei principali strumenti algoritmici ad essi associati.

Il connessionismo attuale trae ispirazione remota da schematici modelli funzionali delle cellule nervose, introdotti nella scia del lavoro pionieristico di Warren McCulloch e Walter Pitts del 1943. Successiva-

mente, attraverso la mediazione delle celebri ricerche sui meccanismi dell'apprendimento di Donald Hebb del 1949, lo studio dei processi di apprendimento e classificazione in reti con neuroni *à la* McCulloch e Pitts si è incontrato, nel Percettrone di Rosenblatt, con la tradizione del connessionismo psicologico e neurologico risalente a Ivan Pavlov, Edward Thorndike e Clark Hull (alcuni testi di McCulloch e Pitts, di Rosenblatt e di Hebb sono raccolti in Anderson, Rosenfeld, 1988). Dopo un periodo di relativa stasi – attribuita da molti sia all'impossibilità di sintetizzare alcune semplici funzioni booleane con Percettroni formati da un solo "strato" di neuroni sia alla mancanza di validi algoritmi di apprendimento per reti a più strati – lo studio delle reti neurali è ripreso con grande vigore intorno al 1985, grazie all'introduzione di nuovi modelli di neuroni e di nuovi algoritmi di apprendimento per reti multistrato (come quello di retropropagazione), ma anche grazie alla disponibilità di calcolatori più potenti per simulare il comportamento di reti formate da un elevato numero di unità.

Il connessionismo ha affrontato, non di rado con maggior successo della prima IA, quelli che abbiamo chiamato problemi di conoscenza tacita e di robustezza, ad esempio in relazione all'elaborazione di segnali sensoriali. La nuova robotica e la cognizione situata hanno invece affrontato i problemi dell'azione in tempo reale. Contrastando una tendenza, manifestatasi all'interno della prima IA, a scindere nettamente tra sistemi percettivi, sistemi motori e sistemi cognitivi in senso stretto, la cognizione situata ha variamente tematizzato l'interazione tra mente, corpo e ambiente circostante (cfr. Clark, 1997). Il modello ideale di un sistema di successo dell'IA è diventato così un robot immerso in un ambiente non predeterminato dal progettista e in grado di mostrare comportamenti adattivi in tempo reale, piuttosto che un calcolatore in grado di sostenere, poniamo, un'educata conversazione salottiera, come nel gioco dell'imitazione di Turing (1950), o di battere alla maniera di Deep Blue un grande maestro di scacchi.

Inizialmente, la proposizione di questo nuovo modello ideale è sfociata in posizioni estreme, che hanno proposto l'eliminazione di ogni forma di rappresentazione ed elaborazione simbolica dal repertorio degli strumenti dell'IA. In alcuni scritti che si collocano alle origini della nuova robotica, Rodney Brooks ha assunto una posizione non lontana da un atteggiamento eliminativista "forte" nei confronti delle rappresentazioni. Le concezioni di Brooks hanno guidato lo sviluppo di architetture innovative per il controllo dell'azione in robot che si sono mostrati capaci di agire in tempo reale e in ambienti non completamente strutturati, senza fare ricorso a sistemi di controllo centralizzato (cfr. ancora CAP. 5). Brooks ha anche riproposto importanti questioni epistemologiche e

ontologiche relative all'analisi e alla simulazione dei sistemi intelligenti. Alle carenze nel controllo dell'azione in tempo reale si aggiunge, tra i motivi di insoddisfazione verso le rappresentazioni simboliche della prima IA, il fatto di non riuscire a definire in modo soddisfacente che cosa veramente siano le rappresentazioni (ad esempio quali siano le loro condizioni di identità) o quale sia il loro ruolo nei sistemi intelligenti, naturali e artificiali. Sebbene queste difficoltà siano innegabili, è veramente arduo rinunciare alle rappresentazioni nella costruzione e nella spiegazione del comportamento di sistemi intelligenti, come indicano gli stessi ostacoli incontrati dal programma di ricerca di Brooks.

Per questa ragione, il richiamo a non trascurare l'interazione tra corpo, cognizione e mondo in altri casi è sfociato in proposte anche molto diverse di integrazione tra gli strumenti della prima IA e quelli dell'IA "nuova". Come risultato si è avuta la progettazione di sistemi con connessioni retroattive più profonde fra i processi percettivo-motori e i processi di rappresentazione ed elaborazione simbolica (cfr. ad esempio Carlucci Aiello, Nardi, Pirri, 2001). L'assenza di modelli interni del mondo e di elaborazioni simboliche rende infatti il comportamento dei sistemi *à la* Brooks troppo dipendente da stimoli sensoriali locali. Il punto di vista originario di Brooks si espone in effetti a obiezioni simili a quelle già sollevate sia contro impostazioni comportamentiste allo studio dell'intelligenza, basate esclusivamente sulle correlazioni fra stimoli e risposte, sia contro i primi modelli ciberneticici dell'azione rivolta a uno scopo, nei quali lo scopo veniva identificato con un oggetto fisico presente nel raggio di azione dei sensori del sistema modellato (Somenzi, Cordeschi, 1994, pp. 23-7). È inoltre difficile immaginare come gli austeri strumenti consentiti da Brooks possano fornire una base adeguata per simulare capacità di ragionamento che troviamo anche nei sistemi della prima IA, oppure per affrontare le problematiche dell'IA *distribuita* e dei *sistemi multiagente* (sui quali cfr. CAP. 9). Un sistema intelligente immerso in un ambiente nel quale operano altri agenti deve essere infatti capace di comunicare con altri agenti, formarsi aspettative sui risultati di possibili interazioni, ragionare a proposito degli obiettivi e delle capacità degli altri agenti. Proprio la dimensione sociale dell'intelligenza è diventata in anni recenti uno dei settori principali di indagine dell'IA, anche perché, oltre a Internet, le interfacce uomo-macchina e sistemi diversi di controllo negli elettrodomestici pullulano di piccoli agenti intelligenti specializzati (cfr. CAP. 11 per ulteriori esempi), i quali devono interagire con gli utenti umani o almeno, se "invisibili" a questi ultimi, con altri agenti artificiali.

I vantaggi largamente complementari mostrati dai sistemi basati sulla conoscenza e da quelli connessionisti o da quelli ispirati alle tesi

della cognizione situata vengono spesso sfruttati nella progettazione di agenti “ibridi” in IA. Questa spregiudicatezza metodologica è legittima e auspicabile in una visione ingegneristica dell’IA. Ma per l’IA più strettamente collegata alla scienza cognitiva si pone il vincolo ulteriore e più severo di una maggiore aderenza ai processi cognitivi degli esseri umani o di altre specie animali. È opportuno notare a questo riguardo come la nozione di agente, attorno alla quale si è formato un vocabolario condiviso da ricercatori operanti nei vari programmi di ricerca che abbiamo menzionato, abbia un forte impatto metaforico, mentre richiederebbe un confronto più puntuale con quelli che consideriamo esempi *genuini* di agenti cognitivi. Nel diffuso impiego metaforico di un vocabolario cognitivista bisogna indicare uno dei vizi ricorrenti dell’IA, che ha suscitato fin dalle origini giuste critiche di vuoto ideologismo e una serie di polemiche verbali, a detrimento di una più equilibrata valutazione dei risultati sia ingegneristici sia teorici ottenuti da questa disciplina nel corso della sua breve storia. Quali sono, dunque, le metodologie che potrebbero essere adottate per condurre un confronto sistematico soddisfacente tra i sistemi algoritmici dell’IA e altri agenti cognitivi?

#### 1.4

### Modelli algoritmici nello studio della mente

Turing (1950) ha discusso la possibilità di controllare, attraverso ciò che egli chiamava «gioco dell’imitazione», se un calcolatore opportunamente programmato potesse sostenere una conversazione nei modi tipici di un essere umano. La simulazione di comportamenti intelligenti, seppure in forme diverse da quelle contemplate nel gioco dell’imitazione (oggi noto anche come *test di Turing*), è rimasta un obiettivo centrale dell’IA. Ma per quanto riguarda l’IA come “nuova scienza della mente”, la tesi che la simulazione, purché funzioni, possa essere conseguita sulla base di *qualunque* mezzo algoritmico ha dato luogo a diversi e gravi equivoci. La storia dell’IA ha mostrato che si possono simulare con successo, ad esempio, varie forme di comportamento verbale senza gettare alcuna luce sui meccanismi che presiedono alla generazione o alla comprensione del linguaggio naturale (i difficili problemi in questo settore di ricerca sono discussi nei CAPP. 6, 7). Esempi noti e spettacolari sono i programmi vincitori del premio Loebner, o ELIZA di Joseph Weizenbaum, un programma che dialogando con un utente umano «scimmiettava», per dirla con Schank (1984), la comprensione della lingua inglese. Viceversa, la costruzione di programmi simulativi che siano rilevanti sul piano della *spiegazione* di certe capacità cognitive è per certi aspetti assimilabile al normale processo di costruzione e revisione dei

modelli scientifici. Per comprendere anche sviluppi più recenti in questa direzione, è bene partire da un breve esame della metodologia messa a punto da Newell, Shaw e Simon (cfr. Cordeschi, 1984, per maggiori dettagli). Essi furono infatti i primi a porsi esplicitamente il problema della simulazione come strategia di costruzione di modelli del mentale.

La psicologia dell'elaborazione dell'informazione (PPI d'ora in avanti) aveva appunto questo obiettivo: costruire modelli algoritmici, cioè programmi per calcolatore che simulassero i *processi cognitivi* umani in modo psicologicamente realistico. Non meno dell'IA delle origini, la PPI era interessata in primo luogo alla questione del *controllo*. Una routine di controllo che selezionava le regole in funzione dell'obiettivo venne sperimentata da Newell, Shaw e Simon nella soluzione di problemi di logica proposizionale da parte del Logic Theorist. Si trattava di un procedimento euristico "guidato" dall'obiettivo, diventato presto noto come analisi mezzi-fine: invece di applicare sistematicamente tutte le regole della logica proposizionale, la routine, attraverso il confronto tra la formula che rappresentava lo stato iniziale del problema e quella che rappresentava lo stato finale, selezionava solo quelle regole che permettevano di *eliminare le differenze* tra le due formule. La soluzione del problema consisteva nel generare una successione di formule progressivamente più simili a quella che rappresentava lo stato finale, coincidente con l'ultima formula della successione.

Questa forma di controllo è stata ampiamente sperimentata in numerosi programmi di IA dedicati a compiti diversi (come viene illustrato nel CAP. 2). Nella PPI essa divenne inizialmente il fulcro di una metodologia simulativa che, sviluppata in particolare da Newell e Simon, è entrata nella storia delle scienze della mente del Novecento. Essi pensarono di registrare il protocollo verbale di un soggetto che risolveva un problema riferendo "ad alta voce" i procedimenti selettivi o euristici che usava e di implementare questi ultimi in un programma. Il confronto tra il protocollo verbale e la traccia del programma avrebbe dovuto mostrare se e fino a che punto la simulazione aveva avuto successo, cioè se e fino a che punto *processi* umani e *processi* algoritmici di soluzione di problemi erano gli stessi, almeno sotto il profilo dell'elaborazione dell'informazione.

Un aspetto importante di questa metodologia simulativa era che il programma doveva tener conto dei *limiti* effettivi del solutore di problemi umano, limiti di memoria, di velocità di elaborazione dei dati e così via. Il programma doveva cioè soddisfare alcune *restrizioni* se voleva essere una simulazione psicologicamente realistica del comportamento del solutore di problemi umano. Non bastava che la prestazione della macchina fosse la stessa dell'essere umano, che la macchina, per

così dire, si *sostituìsse* a quest'ultimo nel dare la risposta giusta (la soluzione del problema). Il confronto protocollo-traccia doveva garantire che la simulazione si estendesse ai *processi* di soluzione del problema. Era questa condizione che distinse subito l'impostazione di Newell, Shaw e Simon, e anche l'uso stesso del termine "euristica" fatto da questi ultimi, rispetto a varianti più ingegneristiche dell'IA.

Dunque non ogni programma dell'IA poteva essere considerato una simulazione interessante sotto il profilo della spiegazione psicologica, in breve un "modello". Il modello prendeva la forma di un programma di un tipo ben definito, che fu anche chiamato «microteoria» da Newell e Simon. Una microteoria riguardava il comportamento di un *singolo* solutore umano di problemi, mentre la teoria *generale* dell'elaborazione umana dell'informazione era un corpus di generalizzazioni qualitative (relative alle caratteristiche della memoria a breve e a lungo termine di tutti i solutori umani di problemi, ai loro tempi di reazione ecc.), non più esprimibili sotto forma di programmi, ma ricavabili dallo studio dei singoli programmi simulativi o microteorie (Newell, Simon, 1972).

Quando il programma "girava" sul calcolatore, diventava possibile individuarne le lacune, suggerendo le modifiche per migliorarlo e trovando nella sua prestazione anche corroborazioni o smentite della teoria generale. Per esempio, del Logic Theorist, che sotto questo profilo risultò essere una simulazione molto grossolana, venne data ben presto una nuova versione, basata su un confronto più puntuale delle euristiche implementate nel programma con i protocolli verbali di singoli soggetti alle prese con problemi di logica. Parliamo del ben noto General Problem Solver che, a giudizio dei suoi ideatori, dimostrava in molti casi di essere una buona simulazione del protocollo verbale di un singolo soggetto umano alle prese con la dimostrazione di un semplice teorema di logica proposizionale.

Sulla base di queste esperienze, Newell e Simon hanno ritenuto che il test di Turing fosse un test «debole», come essi dicevano, proprio perché non riguardava i *processi* del pensiero, ma solo la prestazione finale. E per la stessa ragione la sconfitta di un essere umano da parte di un calcolatore che giochi a scacchi, come nel caso di Deep Blue, benché interessante, non è un risultato particolarmente significativo dal punto di vista di quanti sono rimasti interessati al realismo psicologico della simulazione. «Processi diversi danno luogo a risultati simili»: questo è del resto il commento dei costruttori di Deep Blue, i quali, con parole che Newell e Simon avrebbero probabilmente sottoscritto, sottolineano il «contrasto di stili» tra uomo e macchina durante il gioco (come si legge nella pagina web dell'IBM dedicata al confronto tra Kasparov e Deep Blue).

In definitiva, i programmi sviluppati da Newell e Simon ai tempi della prima IA possono essere visti come altrettanti tentativi di costruire *modelli* cognitivi, impiegati per attribuire una struttura funzionale interna a parti del sistema cognitivo e per dare conto su questa base di varie proprietà esibite dal sistema stesso. Questi modelli vennero considerati significativi anche sotto il profilo predittivo. Attraverso il confronto con il comportamento dei soggetti umani, una simulazione algoritmica diventava uno strumento per controllare e raffinare ipotesi empiriche sui processi di pensiero.

Può un approccio del genere portare alla costruzione di *buoni* modelli del mentale? Tra le varie critiche sollevate in merito a questo punto, ci sono state quelle che hanno sottolineato l'incompletezza dei protocolli verbali: poiché non tutti i processi mentali coinvolti nella soluzione di un problema sono accessibili all'introspezione, la costruzione di un programma simulativo deve attingere a ipotesi arbitrarie su processi di pensiero non documentati nel protocollo. Per quali vie, allora, si possono introdurre ulteriori vincoli empirici sul modello, riducendone gli aspetti arbitrari?

Mentre Simon ha continuato a percorrere la strada della costruzione di microteorie o di teorie di «media ampiezza», come egli disse, raffinando le tecniche dell'analisi dei protocolli verbali, Newell finì ben presto per ritenere che la costruzione di singole microteorie simulative portasse in un vicolo cieco, perché non permetteva di generalizzare plausibilmente i risultati raggiunti in direzione di una teoria «generale» o «unificata» della cognizione (l'antico obiettivo che la PPI aveva affrontato in termini di generalizzazioni *qualitative*). Egli propose di costruire i programmi simulativi sulla base di nuove restrizioni, indipendenti dai singoli compiti cognitivi e dai singoli soggetti, e che includessero proprietà invarianti del sistema cervello-mente (Newell, 1990). L'individuazione di proprietà invarianti dell'architettura funzionale del sistema cognitivo è stata al centro anche di un approccio molto diverso alla scienza cognitiva, quello di Pylyshyn (1984). Altri, come Fodor (1983), hanno sostenuto la tesi di un'architettura *modulare* della mente, dunque opposta a quella dell'architettura *unificata* che Newell e diversi collaboratori hanno sperimentato nel sistema SOAR.

Si tratta di una questione che è tuttora oggetto di diverse valutazioni all'interno della scienza cognitiva, ma va sottolineato che, rispetto alla PPI, fondamentale legata alla psicologia cognitivista, il contesto interdisciplinare nel quale si collocano le ricerche su SOAR si apre alle neuroscienze, ponendo nuove e difficili sfide epistemologiche per i modelli cognitivi algoritmici. La convergenza in un singolo modello di restrizioni provenienti da ambiti disciplinari diversi richiede che ci si

interroghi sulla compatibilità di assunzioni teoriche e metodologiche soggiacenti le varie discipline coinvolte.

Un'impostazione diversa, molto influente nell'IA e nella scienza cognitiva, al problema della costruzione di modelli algoritmici è stata suggerita da David Marr. Secondo Marr (1982), per costruire un buon modello algoritmico di una qualche capacità cognitiva bisogna anzitutto specificare *che cosa* dovrebbe essere in grado di fare un soggetto umano che possiede quella capacità, e cioè quali corrispondenze algoritmiche tra ingressi e uscite (ovvero quali funzioni Turing-calcolabili) caratterizzano l'esercizio di quella particolare capacità cognitiva. In assenza di tale analisi non si ha un'idea sufficientemente chiara di che cosa si vuole simulare o modellare. Newell e Simon, a giudizio di Marr, avrebbero mostrato poca attenzione verso questo problema, concentrandosi prematuramente sui *modi* in cui i soggetti umani svolgono un compito cognitivo – un problema, questo, subordinato epistemologicamente e temporalmente all'analisi *astratta* del compito.

In un contesto ancora preteorico, Marr ha proposto di identificare il compito del sistema visivo umano con la trasformazione dei valori di intensità luminosa registrati dai fotorecettori della retina (gli ingressi del sistema visivo) in informazioni relative alla forma e alla posizione degli oggetti nel campo visivo (le uscite finali del sistema visivo). All'interno di questo sistema complessivo Marr ha distinto vari sottosistemi deputati a compiti più specifici e ha proposto di analizzarli con il medesimo impianto metodologico: bisogna anzitutto mostrare che le corrispondenze fra gli ingressi e le uscite del sottosistema oggetto di indagine possono essere ottenute, in linea di principio, mediante un procedimento algoritmico. Una volta fatto ciò, si può tentare di sviluppare il modello algoritmico analizzando i processi mediante i quali i soggetti umani determinano queste corrispondenze. Marr ha proposto di astrarre, in una prima fase di costruzione del modello, dai concreti processi di esecuzione del compito e dalle limitazioni del sistema cognitivo degli esseri umani che influenzano l'esecuzione del compito (su questa concezione del sistema visivo umano, e su altre avanzate dopo Marr nel campo della Visione artificiale, cfr. CAP. 6). È evidente a questo punto il conflitto con Newell e Simon, secondo i quali, per arrivare a modelli algoritmici con un qualche potere predittivo ed esplicativo, bisogna da subito prendere in considerazione i limiti del sistema cognitivo e le sue strategie euristiche.

Nei modelli connessionisti, l'analisi astratta richiesta da Marr si riduce generalmente alla selezione di esempi significativi di ingressi o di uscite associati all'esecuzione del compito preso in considerazione. Elaborando tali esempi mediante un appropriato algoritmo di appren-

dimento, la rete neurale dovrebbe imparare a svolgere correttamente il compito anche in casi non considerati nella fase di apprendimento. Il confronto tra processi cognitivi e processi di elaborazione neurale, al di là di mere relazioni ingresso/uscita, è reso problematico dalla stessa natura dell'elaborazione neurale, che avviene attraverso la propagazione di una miriade di segnali spesso privi di una plausibile interpretazione cognitiva. Nei modelli connessionisti più rappresentativi l'unità significativa sul piano cognitivo o simbolico è un gruppo più o meno ampio di neuroni in uno stato complessivo determinato. Ma va riconosciuto che i limiti dei modelli cognitivi connessionisti, nonostante la ricordata rinascita del connessionismo, restano per ora quelli di sempre: non è ancora chiaro se e come essi possano riuscire a simulare le capacità cognitive "superiori" (ad esempio inferenziali) e non solo quelle "inferiori" (associazioni, semplici forme di apprendimento). Come hanno concluso gli stessi Anderson e Rosenfeld:

al momento le nostre reti, dopo trent'anni di progressi, funzionano ancora come "cervelli lesionati" [incapaci di comportamento simbolico, secondo l'espressione di Rosenblatt]. Resta aperta la questione di quali severe modifiche apportare alla teoria delle reti neurali affinché esse riescano a raggiungere le funzioni cognitive superiori (Anderson, Rosenfeld, 1988, p. 91).

Nei modelli situati, è la dinamica del sistema formato da un agente algoritmico dotato di un corpo e dall'ambiente circostante a diventare oggetto principale di indagine. Anche in questo caso, i vantaggi dei modelli situati (robustezza, azione in tempo reale e, particolarmente nel caso dei robot evolutivi, anche miglioramento delle prestazioni) si pagano con alcuni limiti (incapacità di ragionamento e di pianificazione complessa). E anche in questo caso si sperimentano approcci variamente "ibridi", che cercano di combinare al meglio le caratteristiche positive dei diversi modelli. Ma nonostante tutto, i modelli situati non sembrano ancora accontentare i teorici radicali della cognizione situata. Clancey (1997) ritiene ad esempio che i robot *à la* Brooks o *behavior based* sono ancora troppo "pre-programmati" dal progettista, e non sono in grado di sviluppare in modo sufficientemente autonomo l'interazione con l'ambiente.

## I.5

### IA e filosofia della mente

Una delle prime discussioni filosofiche intorno all'IA ha preso le mosse dai teoremi di incompletezza e di indecidibilità. Già Turing (1950) aveva discusso il problema da una prospettiva particolarmente adatta agli scopi

di una modellistica cognitiva, utilizzando il teorema della fermata da lui dimostrato quindici anni prima. In base alla tesi di Church-Turing, questo risultato ci permette di concludere che nessun algoritmo consente di rispondere correttamente a *tutte* le infinite domande della seguente forma: «Sia  $X$  un programma e  $Y$  un particolare ingresso per  $X$ . Si dà il caso che  $X$  si fermerà in un tempo finito fornendo un'uscita determinata in corrispondenza dell'ingresso  $Y$ ?». E inoltre, per ogni dato programma  $a$ , è possibile esibire una domanda di questo tipo tale che, *se*  $a$  risponde in un modo qualsiasi, *allora* la sua risposta sarà dimostrabilmente errata. Turing commentava:

Ogni volta che viene posta a una di queste macchine un'opportuna domanda critica, ed essa dà una risposta definita, noi sappiamo che questa risposta deve essere errata, e questo ci dà un senso di superiorità. Questo senso di superiorità è illusorio? Certamente esso è genuino, ma non credo vi si possa attribuire troppa importanza. Diamo troppo spesso risposte errate anche noi, per sentirci giustificati nel provar piacere davanti a tali prove della possibilità di errore da parte della macchina (Turing, 1950, p. 179).

Turing sosteneva che bisogna concedere la possibilità di sbagliare alle macchine che intendiamo usare per imitare o modellare i comportamenti cognitivi degli esseri umani. Dopotutto, ammettiamo senza battere ciglio questa possibilità nel caso dei matematici più dotati. Ma se ammettiamo che le macchine sbagliano, allora non sarà più possibile concludere, in base al teorema della fermata, che le macchine alla quali siamo interessati nella modellistica cognitiva non sono in grado di fornire risposte a certe domande critiche. In altre parole, ammettere macchine fallibili come modelli ragionevoli delle capacità cognitive umane blocca alla radice la possibilità di sfruttare il teorema della fermata per argomentare che gli esseri umani sanno dare risposte corrette a certi problemi matematici mentre i loro presunti modelli algoritmici non sono in grado di farlo. Analoghe considerazioni valgono per tentativi simili che prendono lo spunto dai teoremi di incompletezza di Gödel (cfr. Tamburrini, 1997).

Turing sottolineava come *forti* idealizzazioni a proposito delle capacità matematiche degli esseri umani sono necessarie (ma non sufficienti) per ipotizzare una qualche rilevanza dei teoremi limitativi in questo contesto, ma suggeriva che l'utilità di tali idealizzazioni è perlomeno dubbia. Per quanto la storia della scienza mostri come in molti casi le giuste idealizzazioni siano state cruciali per arrivare a leggi generali nell'ambito della fisica, non disponiamo di una base teorica e sperimentale analoga per decidere se e quali idealizzazioni siano le più opportune

nello studio algoritmico di determinate capacità cognitive, e comunque non c'è alcun accordo tra i ricercatori su questo punto. Abbiamo ricordato, tra i tanti possibili esempi, come Marr sostenesse l'opportunità di introdurre forti idealizzazioni nei modelli algoritmici, astruendo da vari fattori che limitano di fatto le prestazioni cognitive, mentre, a partire dai primi lavori di PPI, Newell e Simon abbiano percorso altre strade, a loro volta molto diverse tra loro.

Fin dalle origini l'IA ha esercitato una notevole influenza sui filosofi interessati al problema mente-corpo e, più in generale, allo statuto epistemologico delle scienze tradizionalmente coinvolte nella spiegazione della vita mentale, come la psicologia e le neuroscienze. L'esempio più chiaro di tale influsso, evidente già a partire dal celebrato articolo di Putnam (1960), è la tesi del funzionalismo, collegata strettamente alla tesi della realizzabilità multipla dell'intelligenza e, conseguentemente, all'ipotizzata autonomia dei *livelli* di spiegazione del mentale. Queste varie tesi sono state indipendentemente dibattute dagli stessi ricercatori della prima IA.

Tanto la PPI quanto l'IA successivamente coinvolta nella scienza cognitiva hanno considerato la mente umana come un sistema di elaborazione dell'informazione, proponendo di indagarne i processi a uno specifico livello di analisi attraverso la costruzione di modelli algoritmici. Inizialmente in questo progetto non erano *direttamente* coinvolte le neuroscienze. Per la PPI, ad esempio, il futuro si sarebbe incaricato di mostrare se e fino a che punto i modelli algoritmici possono essere messi in una qualche relazione (eventualmente di *riduzione*) con teorie biologiche o fisiche del cervello. In assenza di conoscenze approfondite sulla biologia e la fisica dei processi cognitivi, la psicologia, come PPI, avrebbe potuto svilupparsi *autonomamente* al livello dell'elaborazione dell'informazione (proprio come la chimica si è consolidata inizialmente al livello molecolare senza aspettare che fosse nota la struttura atomica della materia).

La relativa autonomia dei livelli di spiegazione trovava ispirazione e insieme conferma nell'architettura del calcolatore. È proprio questa che suggerì ai padri dell'IA di respingere la posizione del riduzionismo estremo, o "laplaciano", che svela la propria paradossalità allorché si immagini di applicarlo ai possibili livelli di descrizione di un calcolatore. È come se noi, volendo spiegarne il funzionamento, rifiutassimo di descriverlo al livello del linguaggio di programmazione (il LISP o il C, poniamo), scegliendo quello, sottostante, del linguaggio macchina, o addirittura quello dello hardware, cioè del funzionamento del calcolatore come macchina fisica – un livello solo illusoriamente ("laplacianamente") assumibile come "più fondamentale", ma in realtà «acme della

riduzione e dell'incomprensibilità», per usare le parole di Simon (1973, p. 26).

Questa concezione dei livelli del calcolatore ha fornito il nucleo teorico a differenti programmi di ricerca prima in PPI e poi in scienza cognitiva, e anche all'impostazione dei rapporti tra le varie discipline coinvolte, in particolare la psicologia e le neuroscienze. Una tesi "forte" sul ruolo del calcolatore in scienza cognitiva è quella relativa ai Sistemi fisici di simboli (SFS) proposta da Newell e Simon (1976). Un SFS è un calcolatore potenzialmente universale, relativamente al quale certe espressioni simboliche hanno una funzione *denotante* e possono essere *interpretate*. In particolare, si dice che un SFS interpreta un'espressione simbolica se questa denota un processo algoritmico e l'SFS può eseguire tale processo quando riceve in ingresso quell'espressione simbolica. La tesi "forte" di Newell e Simon afferma che gli SFS, con le nozioni collegate di denotazione e interpretazione, forniscono una base esplicativa per tutto il comportamento intelligente degli esseri umani. Al giusto livello di analisi, le menti umane sarebbero identificabili con degli SFS, i quali sono dotati delle capacità di elaborazione simbolica *necessarie* e *sufficienti* per esibire intelligenza.

Newell (1982) ha successivamente interpretato l'impresa della scienza cognitiva attraverso l'individuazione di tre livelli fondamentali di descrizione di un SFS: il livello della conoscenza, quello sintattico o dei simboli e quello fisico. Questa idea è stata ripresa da Pylyshyn, che l'ha coniugata con l'ipotesi del "linguaggio del pensiero" di Fodor (sul quale cfr. Di Francesco, 2001). Per quest'ultimo, gli stati mentali sono *rappresentazioni*. Le rappresentazioni condividono con gli enunciati di un linguaggio naturale o formale una struttura combinatoria o, come più spesso si dice, compositiva: a partire dagli elementi base del linguaggio, iterando l'applicazione di regole per formare enunciati sintatticamente corretti, si può generare un numero illimitato di enunciati. Le regole generative degli enunciati dei più comuni sistemi logico-formali esemplificano in modo elementare questa proprietà. Un'altra analogia con la logica è fondamentale per il linguaggio del pensiero di Fodor. Le regole sintattiche della logica hanno una giustificazione semantica. In alcuni sistemi di logica proposizionale, ad esempio, da una congiunzione della forma  $A \& B$  possiamo immediatamente ottenere l'enunciato  $A$ . Questa transizione è semanticamente corretta (necessariamente *se*  $A \& B$  è un enunciato vero *allora* anche  $A$  sarà vero). Anche le regole di manipolazione delle rappresentazioni devono permettere transizioni tra rappresentazioni che abbiano una qualche giustificazione semantica. In altre parole, il linguaggio del pensiero, come ogni altro linguaggio, avrebbe una semantica delle rappresentazioni in grado di giustificare

regole di manipolazione delle rappresentazioni, le quali costituiscono una sorta di “motore inferenziale” cognitivo.

In forme diverse, questa concezione ha permeato molti settori della scienza cognitiva. Le rappresentazioni e le loro regole di manipolazione possono avere *realizzazioni multiple*, in sistemi o agenti *fisici* diversi: nel calcolatore sotto specie di hardware a loro volta diversi, negli esseri umani sotto specie di strutture biologiche. Ma i tre livelli sopra ricordati (della conoscenza, dei simboli, fisico) sono stati anche considerati (ad esempio da Pylyshyn) autonomi e irriducibili l'uno all'altro *in quanto livelli di spiegazione*. Secondo questa concezione, non è possibile descrivere, poniamo, una scelta compiuta da un agente razionale senza rifarci alle sue intenzioni, convinzioni e conoscenze, in breve al vocabolario della psicologia del senso comune. Se rinunciassimo ad analizzare il comportamento al livello della conoscenza, dal momento che quella stessa scelta può essere effettuata in una molteplicità di modi, che coinvolgono diversissimi processi neurologici o fisici dell'agente, è difficile concepire la possibilità di formare leggi e spiegazioni generali del comportamento intelligente.

Le ipotesi del funzionalismo e della realizzabilità multipla, unitamente al ruolo centrale attribuito alle rappresentazioni, hanno caratterizzato molti programmi di ricerca in scienza cognitiva e possono essere sintetizzate nel modo seguente: il calcolatore fornisce un adeguato strumento per un'indagine materialistica sui processi cognitivi, dal momento che esso dimostra come certe strutture fisiche, le quali codificano strutture di simboli, possono svolgere il ruolo di *causa* del comportamento intenzionale di un sistema. Più specificamente, questo ruolo causale è assolto dalle rappresentazioni in virtù delle loro proprietà fisiche. Dunque, «i codici sono “psicologicamente reali”, [...] il cervello è un tipo di sistema che elabora tali codici, e [...] i codici hanno un effettivo contenuto semantico» (Pylyshyn, 1984, p. 40).

Critici diversi, come Paul Churchland, Patricia Churchland, Gerald Edelman e i teorici della cognizione situata, nel respingere l'ipotesi degli SFS hanno respinto anche il funzionalismo e quella che considerano un suo corollario, la tesi della realizzabilità multipla. Quest'ultima, mentre stabilisce l'autonomia della spiegazione psicologica, avrebbe contemporaneamente indotto a bandire lo studio del cervello, del corpo e spesso dell'ambiente reale dall'impresa della scienza cognitiva. Ma il funzionalismo e la tesi della realizzabilità multipla non sono un tratto esclusivo dell'IA cosiddetta simbolica: in realtà, è ben difficile che essi possano essere rifiutati da *qualunque* teorico dei modelli, sia esso sostenitore di modelli a reti neurali o di robot situati o di quello che Edelman chiama il «metodo neurale sintetico». In tutti i casi, infatti, i modelli incorpora-

no, in quanto *artefatti* (siano essi simulati o “situati”), le ipotesi di una teoria del comportamento. Questi modelli sono costruiti per condividere l’organizzazione funzionale (non, ovviamente, la struttura specifica, biologica) degli organismi dei quali si vuole spiegare il comportamento, introducendo le restrizioni al livello (o ai livelli) scelti. È la scelta di tali restrizioni, la scelta del livello di idealizzazione o semplificazione, che caratterizza i diversi approcci della scienza cognitiva, piuttosto che l’opzione, che resta inevitabilmente comune, per il funzionalismo e per la tesi della realizzabilità multipla. Per dirla in breve, il metodo dei modelli è dato dal funzionalismo *più* le restrizioni, e la tesi della realizzabilità multipla, al di là delle speculazioni di cui essa è spesso oggetto in filosofia della mente, resta il cuore dell’impresa di qualunque teorico dei modelli.

Una prospettiva che appare effettivamente superata è l’idea che sia *sempre* possibile trascurare, nella costruzione dei modelli di IA, le restrizioni suggerite dalle neuroscienze, limitandosi a quelle che riguardano, per esempio, diversi aspetti dell’architettura cognitiva. Questo punto è stato enfaticamente sottolineato dallo stesso Newell:

L’indipendenza dalla struttura del cervello è stato un elemento retorico importante nella storia della PPI. Ritengo che sia stato all’epoca un atteggiamento importante e giusto; ma non ci può essere più questa voragine tra elaborazione della conoscenza e struttura cerebrale [...]. Il grande vantaggio è che molte altre restrizioni possono ora entrare in gioco [...]. Ma questo è un modo troppo astratto di porre la questione. Il grande vantaggio è che finalmente possiamo raccordare la biologia e la psicologia tramite una superautostrada [...] (Newell, 1990, p. 483).

Altre carenze dei programmi di ricerca legati all’ipotesi degli SFS sono state evidenziate criticamente al suo stesso interno per quanto riguarda, per esempio, la mancata considerazione da parte della prima IA dell’interazione degli agenti con il mondo reale (Vera, Simon, 1993) o della cooperazione tra agenti (Okada, Simon, 1997). Ma è anche opportuno rilevare come i modelli della cognizione che si distaccano maggiormente da questa tradizione, introducendo restrizioni che riguardano il corpo, l’ambiente, l’evoluzione, lo sviluppo, realizzino efficacemente certe prestazioni (risposte in tempo reale, robustezza, abilità percettivo-motorie) ma non riescono a realizzare con altrettanta efficacia funzioni cognitive “superiori” (processi decisionali, dimostrazione automatica, elaborazione del linguaggio naturale), ambiti nei quali riescono meglio modelli algoritmici caratterizzati da altre restrizioni. Sembra perciò poco fondata l’idea che vi sia una classe di restrizioni privilegiate per i

modelli, visto che quelle relative al corpo, o in generale quelle *biologiche*, non consentono di ottenere sempre i modelli algoritmici più soddisfacenti.

Il metodo dei modelli, in tutte le versioni che abbiamo discusso, adotta una strategia inevitabilmente anti-olistica nei confronti di un oggetto così complesso come il sistema cervello-mente, cercando di costruire modelli approssimati di “parti” del sistema. Troppo spesso, tuttavia, non ci si è posto il problema delle relazioni di queste “parti” con *tutto* il resto del sistema. E troppo spesso l'IA ha interpretato questa strategia come l'autorizzazione a trascurare i fattori emotivi che influenzano le capacità e le prestazioni cognitive, nonostante sia da tempo noto l'impatto degli stati emotivi sui processi di apprendimento, oppure il ruolo di selezione svolto dalle emozioni nei processi decisionali. Vi sono, a partire dagli albori dell'IA, importanti eccezioni a questa tendenza a separare nettamente emozione e cognizione (Turing, 1950, p. 150; Simon, 1967; Dyer, 1987). Ma questi spunti sono stati spesso approfonditi entro una prospettiva puramente cognitivista sulle emozioni: i processi emotivi producono *informazioni*, sotto forma di un bilancio dei vantaggi o dei danni potenziali risultanti da una determinata situazione (cfr. LeDoux, 1998, cap. 2; Castelfranchi, 1991).

I limiti di questa prospettiva sono emersi con prepotenza attraverso i lavori di vari neuroscienziati, tra i quali Damasio (1995), che ha studiato il rapporto tra emozioni e processo deliberativo razionale, partendo da osservazioni cliniche su pazienti affetti da lesioni alle corteccie prefrontali. Posti di fronte a una situazione che normalmente evoca forti reazioni emotive, questi pazienti sono incapaci di *provare* alcunché, pur conservando la capacità di descrivere vantaggi o danni potenziali di quella situazione. All'insensibilità emotiva, in questi pazienti si associa una marcata incapacità deliberativa: essi si perdono in lunghi ragionamenti, senza operare una scelta o compiendo scelte disastrose. Damasio ipotizza che solo le emozioni *provate* nei confronti di situazioni reali o immaginate consentono agli esseri umani di scartare *immediatamente* un gran numero di scenari possibili, *potando* drasticamente in questo modo lo spazio di un problema deliberativo e passando all'esame dei meccanismi meno veloci della valutazione razionale solo le poche alternative sopravvissute. Provare emozioni costituirebbe un potente metodo euristico per esplorare lo spazio di un problema decisionale.

La conclusione di Damasio, se corretta, pone l'IA davanti a importanti sfide. Anzitutto, è possibile progettare macchine che abbiano prestazioni comparabili agli esseri umani nel ragionamento pratico, senza che queste provino emozioni? E segnatamente per l'IA che ambisce a fornire modelli algoritmici dei processi deliberativi degli esseri umani:

in che senso un calcolatore può provare piacere o dolore e come bisogna modellare il ruolo causale attribuito da Damasio all'esperire la qualità emotiva di una situazione?

Le emozioni che si provano di fronte a situazioni reali o immaginate possono essere viste come qualità *soggettivamente* esperite di particolari *stati mentali*, cioè come particolari tipi di *qualia*. In filosofia della mente (Di Francesco, 2001), un *quale* è, più genericamente di un'emozione, la qualità soggettivamente esperita di uno stato mentale (per esempio, ciò che si prova nel vedere un oggetto di un determinato colore). Le considerazioni di Damasio indicano un possibile ruolo *causale*, poco evidenziato nella discussione filosofica, di alcuni tipi di *qualia* nel processo cognitivo. Ma le discussioni filosofiche sui *qualia* riguardano soprattutto il tema dell'esperienza cosciente, di quella che viene anche detta *coscienza fenomenica*, spesso distinta dall'*autocoscienza*, che si fonda sul possesso di una idea del sé, dalla *coscienza introspettiva*, per la quale è determinante la capacità di auto-osservazione o da altre accezioni ancora del termine "coscienza". Se sia possibile costruire una macchina *cosciente*, in una o più accezioni di questo termine, è un problema al quale già accennava Turing (1950, p. 179), ed è ora ampiamente discusso nella filosofia della mente contemporanea (cfr. ad esempio Chalmers, 1996); ma nell'ambito dell'IA sono stati considerati solo aspetti della coscienza legati all'introspezione e alla riflessione, attraverso sistemi (come lo stesso SOAR) capaci di fornire resoconti articolati sui propri stati e processi interni e di modificare i propri programmi sulla base di tali resoconti (Cordeschi, Tamburrini, Trautteur, 1999).

## 1.6

### Conclusioni

La breve storia dell'IA è caratterizzata da uno sviluppo tortuoso e anche tumultuoso, segnato da proposte e ripensamenti, entusiasmi e delusioni. Tutto questo è stato spesso descritto, all'interno e all'esterno della comunità dell'IA, come una sorta di lotta tra "paradigmi" contrapposti. A insistere sulla contrapposizione del loro paradigma "subsimbolico" con quello "simbolico" dell'IA sono stati soprattutto i connessionisti degli anni ottanta. In un primo momento, è sembrato di assistere al riaffiorare di una sorta di trauma: Rosenblatt aveva ragione, il libro di Minsky e Papert (1969) sui Percettroni, dando un peso eccessivo ad alcune limitazioni funzionali di vari tipi di Percettroni, aveva ingiustamente cancellato le reti neurali dal mondo della ricerca. È vero che dopo il libro di Minsky e Papert ci fu una drastica riduzione nei finanziamenti della ricerca sulle reti neurali; ma questa fu proseguita da

diversi ricercatori, come James Anderson, Eduardo Caianiello, Stephen Grossberg, Teuvo Kohonen. Un autorevole connessionista, James McClelland, ha dichiarato di non credere che l'evento decisivo per l'arresto della ricerca sulle reti neurali sia stato il libro di Minsky e Papert. Tenendo conto del fatto che la ricerca sulle reti si fa simulandole su calcolatore, semplicemente «non si era pronti per la ricerca sulle reti neurali. [...] La potenza di calcolo dei [calcolatori dei] primi anni sessanta era del tutto insufficiente» (citato da Crevier, 1993, p. 309).

È bene ricordare che questi limiti delle prestazioni dei calcolatori sono gli stessi che hanno condizionato lo sviluppo dell'IA simbolica: a puro titolo d'esempio, si pensi alla scelta del *paradigma* dei sistemi basati sulla conoscenza rispetto al precedente *paradigma* della ricerca euristica sui *toy-problems* o problemi giocattolo – una scelta che semplicemente non si sarebbe posta senza disporre di calcolatori con grandi memorie e una grande potenza di calcolo. Le limitazioni dei primi calcolatori incoraggiavano la sperimentazione di euristiche “deboli” su problemi giocattolo, in quel momento considerati la vera *Drosophila* dell'IA. Visti retrospettivamente, i successi della prima IA possono apparire scontati, ma per l'epoca erano tali da incoraggiare certe scelte piuttosto che altre: la scelta per la “manipolazione euristica di simboli”, *invece* che per l’“imitazione del cervello”, *invece* che per la “rappresentazione della conoscenza”. Nessuna linea di ricerca viene spazzata via da un libro se non è già debole per conto suo.

Prima e dopo la contrapposizione simbolico-subsimbolico, sono state proposte altre contrapposizioni tra paradigmi, anche all'interno dell'IA simbolica. Di volta in volta è stato detto che erano contrapposti il paradigma della ricerca euristica e quello della conoscenza; quello logicista e quello antilogicista; quello dichiarativista e quello proceduralista, quello individuale e quello sociale o multi-agente. Ma è evidente che nessuno di questi può essere riconosciuto come un paradigma nel senso tecnico introdotto da Thomas Kuhn nel suo celebre volume *La struttura delle rivoluzioni scientifiche*. Piuttosto, ciascuno di essi descrive la parola d'ordine di indirizzi di ricerca diversi e anche rivali, in una storia come quella dell'IA in cui si intraprendevano e poi si abbandonavano le strade più diverse, salvo poi riprenderne qualcuna a distanza di tempo. In questo tumultuoso panorama, l'evento dell'emarginazione e poi della ripresa delle reti neurali non appare in fondo così sconcertante come è stato qualche volta descritto. E il fatto che sia possibile accostare e mescolare IA simbolica, connessionismo, cognizione situata in tante realizzazioni “ibride” sembra confermarlo, rendendo un esercizio retorico, se non propagandistico, la loro separazione in paradigmi contrapposti.

## Riferimenti bibliografici

- ANDERSON J. A., ROSENFELD E. (eds.) (1988), *Neurocomputing*, MIT Press, Cambridge (MA).
- BURATTINI E. (1993), *Reti neuroniche e sistemi esperti*, in “Lettera Matematica PRISTEM”, 9, pp. 31-7.
- CARLUCCI AIELLO L., NARDI D., PIRRI F. (2001), *Case Studies in Cognitive Robotics*, in V. Cantoni, V. Di Gesù, A. Setti, D. Tegolo (eds.), *Human and Machine Perception 3: Thinking, Deciding, and Acting*, Kluwer, Dordrecht.
- CASTELFRANCHI C. (1991), *Emozione e regolazione del comportamento*, in T. Magri, F. Mancini (a cura di), *Emozione e conoscenza*, Editori Riuniti, Roma.
- CHALMERS D. J. (1996), *The Conscious Mind*, Oxford University Press, Oxford (trad. it. *La mente cosciente*, McGraw Hill-Italia, Milano 1999).
- CLANCEY W. J. (1993), *Situated Action*, in “Cognitive Science”, 17, pp. 87-116.
- CLARK A. (1997), *Being there*, MIT Press, Cambridge (MA) (trad. it. *Dare corpo alla mente*, McGraw Hill-Italia, Milano 1999).
- CORDESCHI R. (1984), *La teoria dell'elaborazione umana dell'informazione. Aspetti critici e problemi metodologici*, in V. Somenzi (a cura di), *Evoluzione e modelli*, Editori Riuniti, Roma, pp. 321-422.
- CORDESCHI R., TAMBURRINI G., TRAUTTEUR G. (1999), *The Notion of Loop in the Study of Consciousness*, in C. Musio, C. Taddei Ferretti (eds.), *Neuronal Bases and Psychological Aspects of Consciousness*, World Scientific, Singapore.
- CREVIER D. (1993), *AI. The Tumultuous History of the Search for Artificial Intelligence*, Basic Books, New York.
- DAMASIO A. (1995), *L'errore di Cartesio*, Adelphi, Milano.
- DAVIS M. (1982), *Computability and Unsolvability*, Dover, New York.
- DI FRANCESCO M. (2001), *Introduzione alla filosofia della mente*, Carocci, Roma.
- DORIGO M., GAMBARDELLA L. M. (1997), *Ant Colonies for the Traveling Salesman Problem*, in “Biosystems”, 43, pp. 73-81.
- DYER M. G. (1987), *Emotions and their Computations. Three Computer Models*, in “Cognition and Emotion”, 3, pp. 323-47.
- FODOR J. (1983), *The Modularity of Mind*, MIT Press, Cambridge (MA) (trad. it. *La mente modulare*, Il Mulino, Bologna 1988).
- LANGTON C. G. (ed.) (1995), *Artificial Life. An Overview*, MIT Press, Cambridge (MA).
- LEDOUX J. (1998), *Il cervello emotivo*, Baldini e Castoldi, Milano.
- MARR D. (1982), *Vision*, Freeman, New York.
- MINSKY M. L., PAPER T. S. (1969), *Perceptrons*, MIT Press, Cambridge (MA). Ristampato nel 1988 con una prefazione e una postfazione degli autori.
- NEWELL A. (1982), *The Knowledge Level*, in “Artificial Intelligence”, 18, pp. 87-127.
- ID. (1990), *Unified Theories of Cognition*, Harvard University Press, Cambridge (MA).

- NEWELL A., SIMON H. A. (1972), *Human Problem Solving*, Prentice-Hall, Englewood Cliffs (NJ).
- IDD. (1976), *Computer Science as Empirical Inquiry: Symbols and Search*, in "Communications of the ACM", 19, pp. 113-26 (trad. it. in B. G. Bara, a cura di, *Intelligenza artificiale*, Angeli, Milano 1978).
- OKADA T., SIMON H. A. (1997), *Collaborative Discovery in a Scientific Domain*, in "Cognitive Science", 21, pp. 109-46.
- PUTNAM H. (1960), *Minds and Machines*, in S. Hook (ed.), *Dimensions of Mind*, New York University Press, New York. Ristampato in H. Putnam, *Minds, Language and Reality*, Cambridge University Press, Cambridge (MA) 1975 (trad. it. *Menti e macchine*, in *Mente, linguaggio e realtà*, Adelphi, Milano 1987).
- PYLYSHYN Z. W. (1984), *Computation and Cognition. Toward a Foundation for Cognitive Science*, MIT Press, Cambridge (MA).
- SCHANK R. C. (1984), *The Cognitive Computer*, Addison-Wesley, Reading (MA) (trad. it. *Il computer cognitivo*, Giunti, Firenze 1989).
- SIMON H. A. (1967), *Motivational and Emotional Control of Cognition*, in "Psychological Review", 74, pp. 29-39.
- ID. (1973), *The Organization of Complex Systems*, in H. H. Pattee (ed.), *Hierarchy Theory*, Braziller, New York, pp. 1-28.
- SOMENZI V., CORDESCHI R. (1994), *La filosofia degli automi. Origini dell'Intelligenza Artificiale*, Bollati Boringhieri, Torino.
- TAMBURRINI G. (1997), *Mechanistic Theories in Cognitive Science: The Import of Turing's Thesis*, in M. L. Dalla Chiara et al. (eds.), *Logic and Scientific Methods*, Kluwer, Dordrecht.
- TURING A. M. (1937), *On Computable Numbers, with an Application to the Entscheidungsproblem*, in "Proceedings of the London Mathematical Society", 42, pp. 230-65.
- ID. (1950), *Computing Machinery and Intelligence*, in "Mind", 59, pp. 433-60 (trad. it. *Macchine calcolatrici e intelligenza*, in Somenzi, Cordeschi, 1994).
- VERA A. H., SIMON H. A. (1993), *Situated Action: a Symbolic Interpretation*, in "Cognitive Science", 17, pp. 7-48.

### Per saperne di più

I seguenti due volumi sono ottime guide nello studio degli argomenti della scienza cognitiva:

- BECHTEL W., GRAHAM G. (eds.) (1998), *A Companion to Cognitive Science*, Blackwell, Oxford.
- WILSON R. A., KEIL F. C. (eds.) (1999), *The MIT Encyclopedia of Cognitive Sciences*, MIT Press, Cambridge (MA).

Anche da consultare:

- FLORIDI L. (ed.) (2002), *The Blackwell Guide to the Philosophy of Computing and Information*, Blackwell, Oxford.

Inoltre:

- BECHTEL W., ABRAHAMSEN A. (1991), *Connectionism and the Mind*, Blackwell, Oxford. Un'eccellente introduzione ai problemi della filosofia della mente in relazione al connessionismo e alle reti neurali.
- CORDESCHI R. (2002), *The Discovery of the Artificial*, Kluwer, Dordrecht. Una ricerca sui modelli della vita mentale e sulle loro implicazioni filosofiche a partire dal connessionismo di Thorndike e Hull fino alla cibernetica e agli attuali sviluppi della "nuova" IA.
- RANDELL B. (ed.) (1975), *The Origins of Digital Computers*, Springer, Berlin. Una raccolta di testi fondamentali nella storia dei calcolatori. Comprende anche parte del celebre *First Draft* di von Neumann.
- RUSSELL S. J., NORVIG P. (1994), *Artificial Intelligence. A Modern Approach*, Simon and Schuster, Englewood Cliffs (NJ) (trad. it. a cura di L. Carlucci Aiello, *Intelligenza artificiale. Un approccio moderno*, UTET, Torino 1998). Uno dei manuali più completi e aggiornati di IA, con numerose schede sulla sua storia. Il suo studio costituisce l'approfondimento ideale degli argomenti trattati in tutti i capitoli del presente volume. È attualmente in preparazione la seconda edizione inglese.
- TRAUTTEUR G. (a cura di) (1995), *Consciousness: Distinction and Reflection*, Bibliopolis, Napoli. Un raccolta di testi sulla coscienza rilevanti dal punto di vista dell'IA.

### Siti Internet

- <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>. Lo storico documento di Dartmouth nella sua versione completa.
- <http://www.chess.ibm.com/>. Il sito della IBM dedicato ai confronti scacchistici tra uomo e calcolatore.
- <http://vlmp.museophile.com/computing.html>. Un sito completo sulla storia dei calcolatori.
- <http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Turing.html>. Sito dedicato ad Alan Turing.
- <http://heinzi.library.cmu.edu/Newell/>. Archivio degli scritti di Allen Newell presso la biblioteca della Carnegie Mellon University di Pittsburgh.
- <http://www.psy.cmu.edu/psy/faculty/hsimon/hsimon.html>. La home page di Herbert Simon.
- <http://www-formal.stanford.edu/>. La home page del Formal Reasoning Group della Stanford University, dalla quale si può accedere alla home page di John McCarthy.
- <http://www.ai.mit.edu/people/minsky/minsky.html>. La home page di Marvin Minsky.