8. Il connessionismo: reti neurali e fenomeni emergenti

8.1 Menti, cervelli e programmi

Per apprezzare i successi del programma connessionista e individuarne i limiti, è importante stabilire *che cosa si propongono di realizzare*, esattamente, i ricercatori che studiano le reti neurali. Per chiarire l'obiettivo di ricerca dei connessionisti, lo confronteremo con quello dei ricercatori dell'IA convenzionale. Ma, per capire su cosa lavorava l'IA convenzionale, dobbiamo prima di tutto esaminare le differenze tra menti, cervelli e programmi.

Una differenza tra la mente e i programmi convenzionali è stata già messa in luce. Se chiamiamo *cognitive* le attività di alto livello (cioè quelle descrivibili mediante un linguaggio di alto livello, ad esempio mediante un diagramma di flusso), come la pianificazione di un'azione, e *subcognitive* quelle di basso livello, come la percezione, possiamo riproporre tale differenza nello schema seguente (per "computer" si intenda il calcolatore *programmato*):

	Attività "facili"	Attività "difficili"
Mente	*subcognitive*	*cognitive*
Computer	*cognitive*	*subcognitive*

Vi è una differenza importante, inoltre, tra i cervelli e l'hardware dei computer tradizionali. Mentre l'architettura del cervello è *reticolare*, con miliardi di neuroni interconnessi che si eccitano e scaricano *in parallelo* (cioè contemporaneamente), quella dei computer tradizionali - che sono realizzazioni fisiche della macchina di Turing - è seriale, con un unico processore che esegue un'unica operazione alla volta.

Chiamando "di von Neumann" l'architettura dei computer tradizionali, possiamo includere la differenza di "hardware" (in senso lato) nello schema precedente:

	Attività "facili"	Attività "difficili"	"Hardware"
Mente	*subcognitive*	*cognitive*	*reti neurali*
Computer	*cognitive*	*subcognitive*	*von Neumann*

Lo schema non è però completo se non si considera l'*ipotesi della macchina virtuale*. Di che cosa si tratta? Intanto, con «macchina virtuale» ci si riferisce alla simulazione, da parte di una macchina fisica, del comportamento di un'altra macchina: a parità di hardware, programmi diversi definiscono macchine (o spazi) virtuali diverse/i[1]. Le reti neurali simulate su calcolatore sono una classe di macchine virtuali: la "macchina" simulata non è altro che il cervello (o meglio, piccole parti di un suo monostrato cellulare bidimensionale).

L'ipotesi della macchina virtuale viene introdotta così da Dennett:

> Da vari anni circola l'idea che la coscienza umana potrebbe essere l'attività di una specie di macchina virtuale e seriale implementata sull'hardware parallelo del cervello [Dennett 1991, 289].

Ora, a parte il fatto che Dennett (seguendo l'ultima moda) chiama "coscienza" ciò che qui possiamo continuare a chiamare "mente", la sua idea è semplice:

> Proprio come puoi simulare un cervello parallelo su una macchina seriale di von Neumann, così puoi anche, in linea di principio, simulare (qualcosa che assomiglia a) una macchina di von Neumann su un hardware parallelo, e questo è proprio quello che sto suggerendo: le menti umane coscienti sono delle macchine virtuali più o meno seriali implementate [...] sull'hardware parallelo che l'evoluzione ci ha fornito [Dennett 1991, 245].

Si noti, per inciso, che Dennett afferma che la simulazione di una macchina seriale può essere realizzata «*in linea di principio*» dalle reti neurali del cervello. Perché? Perché, come avevamo già accennato, una rete neurale è *equivalente* (a meno della

[1]L'esempio canonico è quello della macchina da scrivere elettronica, cioè la macchina virtuale che i personal computer simulano quando sono programmati per l'elaborazione di testi (*word processing*). Uno spazio virtuale, analogamente, è qualsiasi ambiente simulato (si pensi al "mondo dei blocchi"). «Virtuale», in pratica, significa «simulato».

memoria) a una macchina di Turing. Il "principio" a cui si riferisce Dennett è noto come *tesi di McCulloch-Pitts-von Neumann*:

> qualsiasi funzionamento - nel senso di descrizione dei risultati che certi effetti sugli ingressi del sistema devono produrre all'uscita - che sia definibile logicamente, rigorosamente e senza ambiguità con un numero finito di parole può essere realizzato da una rete neurale formale [Lolli 1995, 255].

Sull'ipotesi della macchina virtuale (e sul concetto di *meme*), Dennett basa addirittura la propria definizione di *coscienza* (o mente cosciente): essa è

> un enorme complesso di memi (o più esattamente, di effetti provocati dai memi nel cervello) che si può comprendere egregiamente pensando al funzionamento di una macchina virtuale «*neumanniana*» implementata sull'architettura parallela di un cervello [Dennett 1991, 236].

L'ipotesi della macchina virtuale, sostenuta da numerosi autori, suggerisce di completare il nostro schema nel modo seguente:

	Attività "facili"	Attività "difficili"	"Hardware"	Macchine virtuali
Mente	*subcognitive*	*cognitive*	*reti neurali*	*von Neumann*
Computer	*cognitive*	*subcognitive*	*von Neumann*	*reti neurali*

Osservando lo schema, si vede come vi sia, tra le caratteristiche funzionali (attività facili/difficili) e architettoniche (hardware e macchine virtuali) della mente e dei calcolatori, una precisa corrispondenza... ma completamente incrociata!

E' probabile che questa corrispondenza incrociata sia la fonte principale della grande confusione relativa a *ciò che dovrebbero fare esattamente* i programmi dell'IA e le reti neurali dei connessionisti. Una *mente artificiale* dovrebbe, in qualche modo, "riaggiustare" la corrispondenza tra le caratteristiche della mente e delle macchine. Ma in che modo? E siamo sicuri che l'obiettivo sia costruire una *mente artificiale*?

Prima di entrare nel vivo della discussione, soffermiamoci brevemente sulle tipologie di "cervelli elettronici" oggi esistenti. Il *parallelismo* dell'architettura cerebrale può essere riprodotto in molti modi diversi, tra cui i principali sono:

- computer con hardware parallelo;
- reti neurali fisicamente realizzate;
- reti neurali simulate su computer.

Esiste un'intera gamma di *computer con hardware parallelo*. Anche il calcolatore sequenziale (architettura "di von Neumann") può essere incluso in tale gamma, come caso limite: esso ha un numero minimo di processori (uno) con una grande potenza di calcolo. All'altro estremo vi sono i calcolatori ad alto parallelismo, in cui un numero elevatissimo di processori di scarsa potenza compie contemporaneamente la stessa istruzione su dati differenti (architettura "SIMD": Stessa Istruzione, Molteplici Dati). Il più famoso computer di questo tipo è la Connection Machine di Daniel W. Hillis, dotata di ben 65536 unità di elaborazione [Hillis 1987]. A metà strada si trovano calcolatori con relativamente pochi processori, ciascuno con una potenza di calcolo piuttosto elevata. Di questo tipo sono i cosiddetti "elaboratori vettoriali", in cui i processori, che hanno accesso ai dati contenuti in una memoria comune, eseguono - contemporaneamente - istruzioni diverse (architettura "MIMD": Molteplici Istruzioni, Molteplici Dati). Ognuna di queste architetture ha vantaggi e svantaggi, in termini di velocità/potenza e costo/prestazioni [Corcoran 1991].

Per quanto riguarda le *reti neurali fisicamente realizzate*, la ricerca si sta muovendo in varie direzioni. La realizzazione fisica può avvenire per esempio mediante circuiti elettronici, con amplificatori operazionali come unità di elaborazione e cavi, resistori e condensatori come connessioni [Tank - Hopfield 1988]; oppure mediante sistemi ottici, con superfici semi-trasparenti ciascun punto delle quali rappresenta la connessione tra due neuroni logici [Abu-Mostafa - Psaltis 1987].

Infine, possiamo considerare le *reti neurali simulate su computer*. I computer possono essere ad hardware parallelo o seriale, ma la maggior parte delle simulazioni (e comunque tutte quelle del gruppo PDP) viene condotta sui tradizionali computer di von Neumann, con hardware non parallelo.

Nel presente contesto ci occuperemo esclusivamente di reti neurali simulate su computer[2]. Esse, come tutte le simulazioni, si collocano dunque sullo stesso piano "ontologico" delle macchine e dei mondi virtuali: *non vi è alcuna differenza*

[2]Da adesso in poi chiameremo «reti neuronali» le reti neurali naturali, neurobiologiche; chiameremo semplicemente «reti neurali» tutte le reti virtuali. *Costruire una rete neurale* significherà: scrivere un programma per creare al computer una rete neurale virtuale.

ontologica, per esempio, tra il mondo dei blocchi in cui si muove SHRDLU *e una rete neurale*.

Fatte tutte queste premesse, possiamo affrontare il problema degli obiettivi dell'IA. Poiché quelli dell'IA convenzionale sono piuttosto semplici da elencare, li vedremo per primi.

Una parte dei ricercatori dell'IA, quella che si occupa dell'IA "cognitiva", ritiene che programmando adeguatamente i computer si possano ottenere *modelli dell'attività mentale*. Nell'IA convenzionale questa idea si trova associata alla convinzione secondo cui tutto ciò che è cognitivamente rilevante avviene al di sopra della linea dei 100 millisecondi, per cui le nostre descrizioni di alto livello sono direttamente traducibili in *modelli dell'attività cognitiva*. Come abbiamo visto, tuttavia, questa convinzione è oggi giudicata un mero "sogno booleano".

Una posizione decisamente più sostenibile, ma non particolarmente interessante, è quella dei *descrittivisti*, secondo i quali i computer, adeguatamente programmati, possono fornirci modelli non tanto dell'attività cognitiva stessa, quanto piuttosto delle nostre descrizioni di essa.

Un terzo gruppo, che studia l'IA "tecnologica", si impegna a programmare i computer in modo da trasformarli in *sistemi esperti*; l'IA tecnologica, tuttavia, non riveste alcun interesse filosofico e non ne parleremo.

Infine, alcuni ricercatori - appartenenti a quella parte dell'IA convenzionale che John Searle ha battezzato IA "forte" [Searle 1980, 46] - aspirano a programmare i computer in modo da farli *pensare*. Poiché questa posizione è la più radicale, spenderemo alcune parole su di essa.

Consideriamo il seguente interrogativo: *il «programma è di per sé una componente del pensiero?»* [Searle 1990, 31]. Secondo Searle, i ricercatori che sostengono l'IA convenzionale forte ritengono che la risposta

> sia affermativa; essi credono, cioè, di creare letteralmente delle menti allorquando scrivono i programmi giusti con gli ingressi giusti e le uscite giuste [Searle 1990, 31].

Contro questa tesi Searle ha combattuto una battaglia decennale, alla quale ha dato inizio con il famoso articolo "Minds, Brains and Programs" (1980). Si tratta di uno di quei testi che dividono e provocano anche a distanza di molti anni; qui ci limitiamo a

proporne la seguente interpretazione: Searle ha avanzato contro l'IA forte un'obiezione più che legittima, basata sulla differenza tra fenomeni simulati e fenomeni effettivi, che chiameremo *obiezione di Searle*; egli ha però difeso la propria obiezione con un inaccettabile esperimento mentale, quello della "stanza cinese", basato su un uso decisamente improprio di quelle che Dennett chiama "pompe di intuizione" [cfr. Dennett 1980]. Probabilmente l'obiezione potrebbe essere sostenuta con buoni argomenti e senza esperimenti mentali (ma non è questa la sede per cercare di farlo).

L'obiezione di Searle consiste nella considerazione secondo cui nessuna simulazione su computer riproduce *effettivamente* tutte le caratteristiche del fenomeno simulato:

> Com'è possibile che qualcuno creda che una simulazione al calcolatore di un processo mentale possa coincidere con il processo mentale? In fine dei conti, per loro natura, i modelli presentano soltanto alcune caratteristiche degli oggetti che riproducono ed escludono le altre. Nessuno si aspetta di bagnarsi in una piscina piena di modelli di molecole d'acqua fatti con palline da ping pong. Allora perché dovremmo pensare che un modello informatico dei processi di pensiero debba pensare davvero? [Searle 1990, 36]

La risposta a questa domanda è che, di fronte alle simulazioni dei processi cognitivi, si scatena un irresistibile "effetto ELIZA":

> Nessuno pensa che una simulazione al calcolatore della digestione possa digerire qualcosa, ma quando si ha a che fare con i processi cognitivi si è disposti a credere in miracoli del genere perché non ci si rende conto che la mente è un fenomeno biologico al pari della digestione. La mente, si suppone, è qualcosa di astratto e formale, non fa parte di quella «roba umida e appiccicaticcia» contenuta nella nostra testa [*ibid.*].

Separare la mente dal cervello, però, equivale - osserva Searle - ad assumere una posizione radicalmente dualista:

> Nel campo dell'intelligenza artificiale, la letteratura polemica contiene di solito attacchi contro qualcosa che gli autori chiamano dualismo; ma questi autori non si rendono conto di peccare a loro volta di dualismo in forma forte: infatti a meno di non accettare l'idea che la mente sia del tutto indipendente dal cervello o da un altro sistema fisico particolare, non si può sperare di creare delle menti soltanto scrivendo programmi [Searle 1990, 36].

L'aspetto importante dell'obiezione di Searle è che essa si applica, indifferentemente, sia alla versione forte dell'IA convenzionale, sia ad un'eventuale interpretazione forte dell'IA connessionista: più in generale, si applica all'interpretazione "forte" di

> qualsiasi sistema computazionale. Non è possibile ottenere contenuti di pensiero semanticamente pregnanti tramite semplici computazioni formali, che siano eseguite in serie o in parallelo [Searle 1990, 33].

Tuttavia, anche se si accetta l'obiezione di Searle (e chi scrive è propenso a farlo), non vi è ragione di ritenere che un cervello artificiale - sia esso un computer con hardware parallelo oppure una rete neurale fisicamente realizzata - *non possa pensare* per ragioni di principio. E' invece probabile che l'unico ostacolo, lungo la strada che conduce ai cervelli artificiali *pensanti*, sia la complessità.

Se adottiamo la prospettiva della teoria gerarchica del reale, questa difficoltà può essere così esposta: il livello di complessità dei cervelli artificiali è lo stesso di tutti i sistemi di *media complessità*; il livello di complessità dei cervelli naturali, invece, è quello dei sistemi *molto complessi*: non solo si tratta di sistemi biologici, ma, tra i sistemi biologici, essi appartengono al sotto-ordine di quelli più complessi. Il salto di complessità tra un cervello artificiale e uno naturale, vivente e funzionante, è dunque "da brivido"; tuttavia, a parte questa differenza abissale di complessità, non vi è ragione di ipotizzare altri impedimenti alla realizzazione di macchine pensanti [Hebb 1980, 29; Braitenberg 1984, 71-5; Churchland - Churchland 1990, 42].

Si noti comunque la seguente difficoltà: un cervello *umano* artificiale, *benché pensante* - quindi complesso tanto quanto i sistemi biologici e in particolare tanto quanto quelli neurobiologici e plastici, dotato di dispositivi per ricevere gli stimoli ambientali (trasduttori), in grado di apprendere dall'esperienza e immerso da lungo tempo nell'ambiente, nonché, necessariamente, dotato dei mezzi sufficienti per comunicare i propri pensieri (per esempio: un sottosistema di produzione linguistica) - produrrebbe pensieri assai *diversi* da quelli prodotti da un cervello naturale normo-funzionante (probabilmente, nella migliore delle ipotesi, sarebbero simili a quelli di un grave handicappato mentale). La ragione è che gran parte dell'apparato emotivo umano è influenzato profondamente dal sistema endocrino e, come oggi è noto, la

cognizione è inscindibile dalle emozioni [Oliverio 1996]; dunque, per avere un sistema cognitivo normo-funzionante, il cervello non basta: occorre un sistema neuro-endocrino più o meno completo [von Foerster 1985, 139] [3].

Veniamo ora ai *possibili* obiettivi dell'IA connessionista; essi sono più o meno analoghi a quelli dell'IA convenzionale:

1. Costruire reti neurali che *possano pensare* (IA connessionista "forte");
2. Costruire reti da utilizzare come *sistemi esperti* (analoghi a quelli dell'IA tecnologica);
3. Costruire reti neurali che *simulino le reti neuronali* (IA connessionista "descrittiva");
4. Costruire reti neurali che *simulino le attività mentali* (IA connessionista "cognitiva").

Per quanto riguarda il primo obiettivo, abbiamo visto che nei suoi confronti si può sollevare l'obiezione di Searle (ma non è chiaro se tra i connessionisti esiste *davvero* un simile programma); il secondo è privo di qualsiasi rilevanza filosofica; pertanto, non ci occuperemo delle prime due strade di ricerca all'interno dell'IA connessionista. Le altre possibilità vanno invece commentate.

Secondo i descrittivisti, i quali ritengono che l'obiettivo da raggiungere sia quello di costruire *modelli neuronali*, le reti dovrebbero diventare sempre più ricche di dettagli, affinché il loro comportamento possa essere sempre più simile (in termini di input/output) a quello delle reti neurali naturali. In effetti, *questa strada sembra quella che oggi attrae il maggior numero di connessionisti*: grandi sforzi sono stati recentemente spesi per costruire reti neurali in grado di simulare, per esempio, il ripiegamento locale nella sanguisuga [Churchland - Sejnowski 1992, 499-516], la locomozione completa della lampreda di mare [Churchland - Sejnowski 1992, 567-81; Grillner 1996] e il riflesso vestibolo-oculomotore umano [Churchland - Sejnowski 1992, 517-54]. L'IA connessionista descrittiva ha un certo fascino, ma in questa sede non ci occuperemo neppure di essa, perché i descrittivisti - come tali - non si interessano alla mente.

[3] D'altra parte, non si vede come un sistema neuro-endocrino artificiale, completamente funzionante (!), potrebbe *sviluppare davvero una mente* senza diventare, come minimo, una specie di "mostro di Frankestein".

Il quarto obiettivo è quello filosoficamente più attraente: *simulare l'attività mentale riproducendo in forma semplificata, al computer, la struttura e il funzionamento delle reti neuronali*. Qui però sorge una perplessità. La perplessità è dovuta alle corrispondenze "incrociate" tra mente e computer: una rete neurale è una macchina virtuale implementata su computer; se anche la mente è una sorta di macchina virtuale implementata sul cervello, allora la simulazione della mente tramite una rete neurale dovrebbe essere la simulazione di una specie di macchina virtuale... *su un'altra macchina virtuale!* Ma che cosa vuol dire "simulare una macchina virtuale"? Una macchina virtuale è essa stessa una simulazione. Inoltre, farlo *utilizzando un'altra macchina virtuale* non è una complicazione eccessiva? Per analogia, sarebbe un po' come voler simulare un cervello utilizzando i programmi dell'IA convenzionale: ma in che modo, per esempio, si potrebbe utilizzare il programma ELIZA per simulare il *cervello* dello psicoterapeuta virtuale?

L'IA connessionista *cognitiva* si muove tra Scilla e Cariddi: da una parte ci sono i programmi tradizionali, che possono tradurre descrizioni di alto livello dell'attività cognitiva ma che non tengono conto delle caratteristiche reali dei cervelli; dall'altra parte ci sono le reti neurali dei descrittivisti, la cui architettura può essere resa più o meno simile a quella cerebrale ma che non presentano tratti psicologici. In altre parole: i programmi dell'IA convenzionale, considerati come descrizioni di basso livello, si rivelano biologicamente implausibili; le reti neurali dell'IA connessionista descrittiva, considerate come descrizioni di alto livello, sono apparentemente inutili. Parafrasando Andy Clark, il quale fa la parafrasi di Kant, possiamo dire che *i modelli della cognizione senza cervello sono vuoti, i modelli del cervello senza cognizione sono ciechi* [Clark 1989, 236; Kant 1787, 78].

8.2 Tra simboli vuoti e neuroni ciechi: i subsimboli

Qual è il legame tra le reti neurali (virtuali) dell'IA connessionista cognitiva e le capacità cognitive delle reti neuronali?

Questo è il problema cruciale, l'"occhio del ciclone" intorno al quale ruotano tutte le discussioni filosofiche sul connessionismo, ma che raramente viene affrontato

in modo diretto. In questa sede, invece, cercheremo di entrare proprio nell'occhio del ciclone. Per chiarezza espositiva, e data la complessità dell'argomento, anticipiamo subito a quale conclusione giungeremo: *l'obiettivo di simulare l'attività mentale è stato parzialmente raggiunto*; più in particolare, i connessionisti hanno raggiunto l'obiettivo di simulare l'attività mentale subcognitiva - ma non quella cognitiva. Non solo: tra le attività subcognitive, quella realmente simulata è quella di più basso livello, cioè il riconoscimento percettivo.

Nonostante questo risultato sia meno eclatante di quello, purtroppo ancora lontano, di simulare la cognizione, si deve riconoscere che l'IA connessionista cognitiva ha fatto luce su molti aspetti dell'attività mentale. Come vedremo, i connessionisti hanno probabilmente compreso qual è la modalità computazionale di base delle reti neuronali, cioè l'*elaborazione distribuita in parallelo*, e sono in grado di ricrearla sulle loro reti neurali; hanno dimostrato, simulandolo con successo, che il riconoscimento percettivo non è altro che semplice elaborazione distribuita in parallelo, da parte di una rete neuronale che abbia precedentemente codificato la rappresentazione di ciò che deve riconoscere; hanno anche compreso qual è il *meccanismo* utilizzato dalle reti per codificare tali rappresentazioni, cioè l'*immagazzinamento della conoscenza nelle connessioni*; infine, i connessionisti del gruppo PDP hanno proposto una *teoria della cognizione* che, pur essendo insoddisfacente, va nella direzione giusta.

Prima di tutto cominciamo a esaminare le reti neurali. A differenza dell'IA connessionista descrittiva, quella cognitiva produce modelli con il *minor numero possibile di dettagli* neurofisiologici e biochimici. Questo rientra perfettamente nello "spirito" delle reti di McCulloch e Pitts: eliminare dal modello tutto ciò che non è considerato indispensabile all'attività mentale. E' questa la ragione per cui, a proposito delle reti del gruppo PDP, si parla di modelli "neuralmente ispirati" [Rumelhart - McClelland 1986a, 180] piuttosto che di modelli "neurali" o "neuronali" (come sono invece quelli dei descrittivisti). I modelli neuralmente ispirati si collocano *a metà strada tra modelli neurali "ciechi" e modelli simbolici "vuoti"*.

Ora, è evidente che - se i connessionisti sono nel giusto - i dettagli presenti nelle simulazioni esauriscono *tutti i dettagli sufficienti a produrre, nelle reti neurali naturali, il pensiero*. Pertanto, "entriamo" in una rete neurale - come si può "entrare"

in un mondo virtuale - e cerchiamo di capire quali sono questi dettagli essenziali responsabili, nelle reti neuronali, dell'attività mentale.

Ogni neurone virtuale della rete viene immaginato come un'unità di elaborazione (processore), caratterizzata da un valore numerico chiamato *stato di attivazione* o *scarica*. Nel programma di simulazione della rete questo valore compare sotto forma di una variabile in funzione del tempo (o meglio, in funzione di una variabile *t* che rappresenta il tempo). Le connessioni tra le unità non vengono rappresentate direttamente, ma nel programma compare un'altra variabile, una per ogni connessione, chiamata *peso sulla connessione*[4]. I pesi non sono funzioni del tempo e si modificano solo durante una procedura chiamata *apprendimento* (che fa parte del programma di simulazione). I pesi, inoltre, possono essere positivi (connessione eccitatoria), negativi (connessione inibitoria) o nulli (nessuna connessione) [Rumelhart *et al.* 1986, 81-9].

Come viene elaborata l'informazione? Così *come* nel cervello, e *diversamente* dai computer di von Neumann, nelle reti neurali non ci sono memorie di dati a cui le unità di elaborazione accedono per "manipolare le informazioni": *l'unica "informazione" che si propaga è la scarica, opportunamente pesata nel passaggio dalle unità afferenti a quelle efferenti, e l'unica "elaborazione" che avviene è l'aggiornamento, nel tempo, dello stato di attivazione delle unità*.

Per quanto riguarda l'aggiornamento, il programma di simulazione deve contenere una regola che stabilisca, per ogni unità, quale sarà il suo stato di attivazione, e quindi la scarica, nell'istante successivo. Di solito, lo stato di attivazione di un'unità viene aggiornato in funzione della *somma pesata delle scariche dei suoi processori afferenti* ("ingresso-netto").

In una rete neurale non c'è assolutamente nient'altro. Ma è possibile che non serva altro? E' possibile che l'attività mentale si riduca a: stati di attivazione, pesi sulle connessioni, una regola di aggiornamento e un algoritmo di apprendimento? La risposta, secondo i connessionisti, è affermativa: l'essenziale è tutto lì. Ora, la letteratura relativa al connessionismo mostra che in questi anni il problema principale

[4]Nella teoria delle reti neurali si intende per *peso* un coefficiente, cioè un fattore moltiplicativo. *Pesare* un valore significa moltiplicare tale valore per il suo peso. Per esempio, la somma pesata di A più B è la seguente: $\alpha A + \beta B$, dove α è il peso di A e β è il peso di B.

dei teorici è stato interpretare ciò che si può ottenere da questi semplici "ingredienti", ovvero quale sia il modo *appropriato* di considerare le reti neurali [Smolensky 1988]. In effetti, sembra che nel trattamento teorico delle reti neurali sia incredibilmente facile confondere vari livelli di analisi. Esaminiamo questo problema.

Se accettiamo l'obiezione di Searle, non ascriveremo alle reti alcuna conoscenza, alcun pensiero, alcuna mente. Le reti neurali sono simulazioni: non hanno alcun punto di vista, né interiorità, né soggettività; in breve, per usare l'efficace formula nageliana, *le reti neurali non sanno cosa si prova a essere una rete neurale*. Tuttavia, come vedremo, esse simulano il riconoscimento percettivo. Come può succedere?

Una risposta ambigua, contro cui è bene spendere qualche parola, è la seguente: le reti neurali *apprendono* a riconoscere configurazioni (*pattern*). Con un'affermazione come questa si rischia di fare confusione tra livelli di analisi, perché *l'apprendimento fa parte del programma di simulazione e non della simulazione stessa*. Esso si pone, per così dire, a un livello più basso rispetto a quello della macchina virtuale, che è invece il livello della rete neurale.

Consideriamo, per chiarire meglio il punto, un computer programmato per creare una rete virtuale. Nell'analisi del software di questo sistema possiamo distinguere due livelli: al livello più basso vi è il programma di simulazione, cioè la sequenza di istruzioni che il programmatore ha introdotto al momento della programmazione, e l'algoritmo di apprendimento si trova a questo livello; al livello più alto vi è invece la rete neurale, una macchina virtuale come tutte le altre, la quale non è altro che il modo in cui l'utente descrive le grandezze trattate dal programma di simulazione e le relazioni tra queste grandezze[5]. Le reti neurali "apprendono" seguendo un algoritmo presente nel programma di simulazione: non è questo il *modo* in cui apprendono le reti neuronali. Se l'apprendimento delle reti neurali facesse parte della *simulazione* (e non solo del *programma di simulazione*), esso dovrebbe avvenire *nello stesso modo* in cui apprendono le reti neuronali: da sole, per auto-modificazione

[5]Si pensi, per analogia, a un computer programmato per elaborare testi: il programma (*word processor*) non ha nulla a che fare con le macchine da scrivere elettroniche - si limita a manipolare variabili numeriche e stringhe di caratteri; tuttavia, l'utente descrive il computer così programmato *come se* fosse una macchina da scrivere elettronica. Il programmatore e l'utente del programma, in altre parole, interagiscono con il computer a due livelli completamente diversi, ciascuno con un proprio linguaggio.

sinaptica, non appena vengono immerse nell'ambiente (con dei trasduttori). Poiché queste modalità di apprendimento non vengono simulate, concludiamo che *la simulazione inizia solo dopo la fase di apprendimento*.

Ma, se le reti neurali non *apprendono a riconoscere* configurazioni, è vero che esse *riconoscono* configurazioni. Esse eseguono perfettamente questo compito (*pattern-matching*), reso possibile dal fatto che *le reti posseggono rappresentazioni codificate delle configurazioni da riconoscere*. La domanda sorge spontanea: che tipo di rappresentazioni possono essere codificate nelle reti neurali? Rappresentazioni simboliche, come quelle immagazzinate e manipolate dai sistemi simbolici, o rappresentazioni di tipo totalmente diverso?

La risposta "ufficiale", relativamente complicata, viene generalmente attribuita a un membro del gruppo PDP, Paul Smolensky. Nell'enigmatico articolo "On the proper treatment of connectionism" (1988), Smolensky illustra la propria interpretazione di quella che abbiamo chiamato l'IA connessionista cognitiva e che egli chiama «*paradigma subsimbolico*» [Smolensky 1988, 60]. Data l'importanza di questo contributo, ad esso dedicheremo un po' di spazio.

Prima di tutto Smolensky affronta il problema dei livelli di analisi. A quali livelli è possibile descrivere un sistema cognitivo? Egli distingue i seguenti:

- livello *concettuale*: è quello delle descrizioni formulabili nel linguaggio ordinario ed eventualmente traducibili in linguaggi di programmazione di alto livello [Smolensky 1988, 62];
- livello *neurale*: è quello delle descrizioni neurobiologiche, utilizzate per parlare del cervello o delle reti neuronali, o anche dei singoli neuroni [Smolensky 1988, 76-8];
- livello *subconcettuale*: è quello di tutte le descrizioni *non formulabili nel linguaggio naturale*; la descrizione subconcettuale di un sistema fisico, per esempio, è quella microscopica (formulata usando le leggi della meccanica quantistica) [Smolensky 1988, 84].

Mentre l'analisi concettuale adopera i concetti del linguaggio naturale (*tazza*, *caffè*, *pieno*, *vuoto*, ecc.), quella subconcettuale fa riferimento a entità che Smolensky definisce *subsimboli*. Che cosa sono i subsimboli? Per rispondere occorre prima

stabilire cos'è un simbolo. Abbiamo già visto che cosa sono i simboli secondo l'approccio cognitivista classico (che Smolensky chiama «*paradigma simbolico*» [Smolensky 1988, 61]): rappresentazioni interpretabili semanticamente, cioè in corrispondenza biunivoca con i concetti del linguaggio naturale utilizzato per descrivere il sistema; allora, i subsimboli "classici", se così si vogliono chiamare, non possono che essere i *costituenti simbolici* di un simbolo molecolare, ciascuno in corrispondenza con i concetti che, a loro volta, costituiscono il "concetto-molecolare" corrispondente al simbolo molecolare.

Che cos'è invece un simbolo secondo Smolensky? E' una rappresentazione interpretabile semanticamente, *ma in un certo contesto*. Anche un simbolo "connessionista", *fissato il contesto*, è dunque in corrispondenza con qualche concetto del linguaggio naturale. Tuttavia, le sue parti costitutive (o le parti delle parti) non sono in corrispondenza biunivoca con le parti costitutive del concetto corrispondente al simbolo, perché *possono anche corrispondere a parti costitutive di concetti relativi al contesto*.

Come esempio analizziamo, a livello rigorosamente concettuale, una rete neurale che rappresenta il simbolo "connessionista" «caffè» nel contesto della «tazza»; potremmo interpretare semanticamente la codificazione compiuta da (estesi) gruppi di processori di questa rete mettendola in corrispondenza con concetti come: «*liquido marrone con superficie piatta, liquido marrone con lati curvi e superficie inferiore, liquido marrone a contatto con porcellana, liquido caldo, contenitore eretto con maniglia, odore di bruciato, e così via*» [Smolensky 1988, 97]. Molti di questi concetti non sarebbero pertinenti se il contesto di «caffè» non fosse «tazza»: il concetto di *liquido marrone a contatto con la porcellana*, per esempio, non potrebbe essere messo in corrispondenza con l'attività di alcun gruppo di processori se la rete rappresentasse il caffè nel contesto di «lattina» [Clark 1989, 258]. Cade così la condizione di *ricorrenza* dell'interpretazione semantica: il simbolo "connessionista" di «caffè» non ricorre ogni volta che la rete neurale rappresenta il caffè, perché lo rappresenta sempre in maniera dipendente dal contesto.

Ciò che vale per l'attività di grandi gruppi di unità, com'è evidente, vale a maggior ragione per i singoli processori, la cui attività corrisponde appunto a ciò che Smolensky chiama *subsimbolo* (o sottosimbolo). La differenza tra simboli "classici" e

simboli "connessionisti" dipende dunque dalla diversa *rappresentazione simbolica* del contesto:

> nel paradigma simbolico il contesto di un simbolo si mostra attorno ad esso, e consiste di altri simboli; nel paradigma subsimbolico il contesto di un simbolo si mostra dentro di esso, e consiste di sottosimboli [Smolensky 1988, 98].

I subsimboli "connessionisti", a differenza dei subsimboli "classici", non sono a loro volta simboli: l'unica interpretazione che possiamo darne è *numerica* e il numero in questione è precisamente il valore dello stato di attivazione (ovvero della scarica) dei singoli processori. Secondo la definizione di Smolensky, infatti, i subsimboli

> corrispondono all'attività delle unità di elaborazione individuali nelle reti connessioniste. Entità che nel paradigma simbolico sono tipicamente rappresentate da simboli, vengono rappresentate nel paradigma subsimbolico da un grande numero di subsimboli. A questa distinzione semantica si affianca una distinzione sintattica. I subsimboli non vengono trattati tramite manipolazione simbolica: essi hanno una natura computazionale di tipo numerico, non simbolico [Smolensky 1988, 62].

Poiché la natura computazionale dei subsimboli è numerica, non sarebbe esatto dire che l'attività dei neuroni reali - la cui natura è neurobiologica - è subsimbolica; tuttavia, poiché l'eccitazione di un singolo neurone non corrisponde ad alcun concetto, essa non è neppure simbolica. Per analogia, almeno finché l'unica "interpretazione" che viene data a un singolo neurone riguarda il suo stato di eccitazione, possiamo allora considerare *subsimbolico* il formato in cui il cervello codifica le rappresentazioni del mondo.

Il fatto che consista di molti subsimboli fa sì che una rappresentazione, in una rete neurale, sia codificata in forma *distribuita*. Vi è un'intera gamma di possibilità a questo proposito: dal caso limite in cui si può dare un'interpretazione semantica all'attività delle singole unità di elaborazione; all'estremo opposto in cui tutte le unità delle reti partecipano alla codificazione subsimbolica di tutte le informazioni [van Gelder 1991]. Naturalmente l'utilità di un tipo di rappresentazione piuttosto di un altro dipende dal tipo di informazione che si vuole rappresentare. Per esempio, supponiamo che l'informazione da codificare sia una cifra scritta a mano su un campo recettivo [Hinton 1992, 119]. Il tipo di codificazione distribuita più utilizzato per affrontare questo tipo di problemi è la "macro-codificazione" (*coarse-coding*): il

campo recettivo viene suddiviso in zone più piccole che si sovrappongono parzialmente in modo sistematico; ogni unità di elaborazione è collegata a una zona e codifica i tratti contenuti in essa. Dallo studio della macro-codificazione risulta che:

> L'idea intuitiva che zone più grandi comportino rappresentazioni meno accurate è del tutto sbagliata, poiché le rappresentazioni distribuite sono più efficienti di quelle locali nel conservare le informazioni. Sebbene ogni unità attiva sia meno specifica nel suo significato, la combinazione delle unità attive lo è assai di più. Si noti pure che nel caso della macrocodificazione la fedeltà è proporzionale al numero delle unità, mentre nel caso delle codificazioni locali è proporzionale solo alla […] radice di tale numero [Hinton *et al.* 1986, 135].

Quali sono i vantaggi delle rappresentazioni distribuite? Detto in breve, i vantaggi sono costituiti da tutte quelle caratteristiche che hanno reso famose le reti neurali: il degrado graduale, l'assegnazione per difetto e le cosiddette rappresentazioni "prototipiche".

Il *degrado graduale* è una caratteristica ovvia: il malfunzionamento delle unità subsimboliche danneggia le informazioni codificate nella rete in misura molto minore rispetto al malfunzionamento di un simbolo in un sistema simbolico classico. In quest'ultimo, la perdita di un simbolo è irreparabile e le conseguenze sono disastrose (a questo problema, di solito, fa fronte la *ridondanza* dei simboli rappresentati); nelle reti neurali, invece, c'è una proporzionalità diretta tra la gravità del danno e il degrado della prestazione [McClelland *et al.* 1986, 60-2].

Per quanto riguarda l'*assegnazione per difetto*, essa dipende dal fatto che tra i subsimboli si manifesta una sorta di "cooperazione" per attivare le informazioni rappresentate in forma distribuita; in particolare, se queste ultime sono state codificate in classi, i gruppi di unità che rappresentano elementi appartenenti alla stessa classe si inibiscono reciprocamente, mentre le connessioni tra le classi sono eccitatorie. Così, se l'ingresso in una rete riguarda solo informazioni codificate in alcune classi ma non in altre (*difetto*), le unità attive della rete attiveranno, tramite le connessioni più forti, anche i processori che hanno codificato le informazioni relative alle altre classi (*assegnazione*) [McClelland *et al.* 1986, 62].

Le *rappresentazioni "prototipiche"*, in una rete neurale, sono una diretta conseguenza del fatto che i simboli non sono memorizzati direttamente, ma vengono

codificati in formato subsimbolico. Quindi decodificare le informazioni precedentemente codificate equivale a *ricostruirle* partendo dai subsimboli (cioè dall'attivazione delle singole unità di elaborazione). In questa ricostruzione, la rete non può limitarsi rigorosamente alle informazioni che ha ricevuto: se vi sono regolarità reali, la rete è strutturalmente obbligata, per così dire, a "generalizzare" - naturalmente senza abbandonare il contesto creato dalle informazioni ricevute. Benché questa caratteristica sia spesso considerata la qualità più affascinante delle reti neurali, la si può anche vedere come l'*incapacità* di distinguere ciò che è stato effettivamente codificato da ogni sua ricostruzione plausibile [McClelland *et al.* 1986, 62-3]. In altre parole, la rete *non può evitare* di trarre "micro-inferenze" tra subsimboli: attivata un'unità, è impossibile che non si attivino anche tutte quelle efferenti. Per questa ragione, le reti neurali più semplici tendono a ricostruire non tanto le singole informazioni, quanto un'unica rappresentazione (il "prototipo") che "media" tra le varie informazioni. In questo senso, più che "generalizzare", è corretto dire che le reti possono "genericizzare". Il concetto di generalizzazione, infatti, è legato alla capacità di riconoscere regolarità *inter*contestuali - cosa *assolutamente impossibile* per una rete neurale. La capacità di *genericizzare* è sempre, esclusivamente, *intra*contestuale.

Facciamo il punto della situazione. Le reti neurali riconoscono configurazioni (*pattern*) grazie alla propria capacità di codificare rappresentazioni simboliche *complesse* - cioè corrispondenti a concetti complessi del linguaggio ordinario - in forma di informazioni distribuite tra numerosi subsimboli privi di significato, ciascuno corrispondente all'attività di una singola unità di elaborazione. Realizzando reti con tale capacità, i connessionisti hanno dimostrato di aver individuato un *meccanismo* che traduce il codice dei simboli, interpretabili semanticamente, in quello dei subsimboli, interpretabili solo numericamente (cioè, secondo Smolensky, sintatticamente). Il meccanismo di traduzione individuato dai connessionisti è, in pratica, *l'immagazzinamento della conoscenza nelle connessioni*. Tutto ciò che una rete codifica diventa una configurazione (una matrice) di pesi sulle connessioni.

Aver individuato questo meccanismo è un risultato che non va sottovalutato: qualunque modello dell'attività mentale umana deve probabilmente partire da una simile tecnica di traduzione. Inoltre, le prestazioni sbalorditive delle reti neurali nel riconoscere configurazioni dimostrano ampiamente che il meccanismo individuato è

ottimo. E' il medesimo utilizzato dalle reti neuronali? Non possiamo dare una risposta definitiva a questa domanda: è certo, però, che non sappiamo neppure immaginare come avvicinarci maggiormente al meccanismo di traduzione reale. Quest'ultimo, per quanto oggi ne sappiamo, si basa sulla plasticità delle sinapsi: cioè sulle modificazioni presinaptiche (che riguardano la quantità e il tipo di trasmettitori liberati oppure la biochimica dei canali nella membrana) e su quelle postsinaptiche (che riguardano il numero e il tipo di recettori sbloccati oppure la depolarizzazione della membrana o ancora la morfologia della spina dendritica). La plasticità sinaptica è assai complessa [Churchland - Sejnowski 1992, 377-435]. Ma, in una rete neurale, i pesi sulle connessioni, i quali devono simulare la qualità e la forza delle sinapsi, possono assumere *qualsiasi* valore: così, benché non sia possibile *quantificare* il tipo e la forza delle sinapsi reali, si ha quanto meno la certezza di non confinare la variabilità sinaptica in uno spettro di valori ristretto. In effetti, fino ad oggi, i ricercatori non hanno ancora avuto la sensazione che esistano dei limiti alle capacità di riconoscimento di configurazioni delle reti neurali. Per dirlo con altre parole: la "conoscenza" codificata nei pesi sulle connessioni si è dimostrata *sufficiente* a risolvere tutti i problemi di riconoscimento finora sottoposti alle reti neurali.

Possiamo finalmente tornare alla domanda da cui siamo partiti: come si legano le reti neurali alle *capacità cognitive* delle reti neuronali? Non abbiamo risposto a questo interrogativo, però abbiamo fatto il primo passo. Siamo infatti in grado di rispondere a una domanda più semplice: qual è il legame tra le reti neurali e le *capacità percettive* (quindi, *subcognitive*) delle reti neuronali? La risposta a questa seconda domanda è la seguente.

Una rete neurale codifica, in forma distribuita (subsimbolica), rappresentazioni interpretabili semanticamente. Quando una configurazione viene fornita come ingresso nella rete, questa informazione viene *elaborata sulla base delle rappresentazioni codificate* e la rete fornisce come informazione di uscita una configurazione interpretabile semanticamente come una *discriminazione*. L'elaborazione delle informazioni realizzata dalla rete, chiamata *elaborazione distribuita in parallelo*, consiste nell'aggiornare contemporaneamente (*in parallelo*) lo stato di attivazione di molte unità; inoltre, l'aggiornamento avviene in funzione delle scariche pesate e,

proprio nel pesare le scariche, la rete utilizza le rappresentazioni *distribuite* che ha precedentemente immagazzinato nei pesi sulle connessioni.

Ora, secondo i connessionisti, *l'elaborazione distribuita in parallelo è la modalità computazionale su cui si basa il riconoscimento percettivo delle reti neuronali*. Le reti neurali, in altre parole, elaborano le informazioni *nello stesso modo* in cui lo fanno le reti neuronali quando sono impegnate in un compito di riconoscimento percettivo.

Ma allora, poiché il riconoscimento di configurazioni, in una rete neurale, consiste semplicemente nell'elaborazione stessa (purché la rete possegga già la rappresentazione subsimbolica delle configurazioni da riconoscere), si può affermare che *il riconoscimento di configurazioni simula il riconoscimento percettivo*. Questo è esattamente il legame tra le reti neurali e le capacità percettive "di basso livello" delle reti neuronali.

8.3 La mente come emergenza secondo il gruppo PDP

I membri del gruppo PDP sono, tra i connessionisti, quelli che si sono maggiormente occupati dei problemi teorici legati alla possibilità di estrapolare alle reti neuronali i risultati della ricerca sulle reti neurali. Il problema del passaggio *dalla percezione alla cognizione* è quello su cui il gruppo PDP si è concentrato specialmente. Prima di tutto, che cosa rende meno trattabile il problema della cognizione rispetto a quello della percezione? Non lo si può semplicemente simulare con le reti neurali?

Noi diciamo che una rete neurale ha simulato il fenomeno del riconoscimento percettivo perché essa ha elaborato una configurazione di ingresso *nello stesso modo* in cui riteniamo che le reti neuronali elaborino le informazioni provenienti dai trasduttori (stimoli percettivi) e ha poi fornito una configurazione di uscita *interpretabile semanticamente* come una discriminazione. Ora, quali requisiti sono richiesti per attribuire alla rete neurale la capacità di simulare la formazione di un pensiero cosciente? Requisiti analoghi a quelli di prima: modalità di elaborazione uguali a quelle cerebrali e una configurazione di uscita interpretabile come pensiero

cosciente. Supponiamo pure che il tipo di elaborazione coinvolto nella formazione di un pensiero, da parte di una rete neuronale, sia ancora l'elaborazione distribuita in parallelo. Ma il problema della configurazione di uscita *interpretabile come pensiero cosciente* appare insormontabile: che tipo di configurazione potrebbe andare bene? Nessuno sa rispondere (probabilmente *non si può* rispondere).

Benché Rumelhart, McClelland e colleghi abbiano sempre ammesso che le reti neurali sono "naturalmente adatte" solo per affrontare i processi subcognitivi e le conoscenze implicite in genere, essi si dicono

> peraltro convinti che questi modelli sono ugualmente applicabili ai processi cognitivi di livello superiore e nello stesso tempo consentono di capire meglio questi fenomeni. Dobbiamo però essere chiari sul fatto che non possiamo attenderci che i modelli PDP trattino i processi di ragionamento sequenziali, estesi e complessi come un unico assetto di una rete parallela. Noi riteniamo che i modelli PDP descrivano la microstruttura dei processi di pensiero, e i meccanismi attraverso cui questi processi giungono, con l'esercizio, a fluire più rapidamente, e a svolgersi in modo integrato tra loro [Rumelhart - McClelland 1986a, 197].

Secondo i membri del gruppo PDP, dunque, la microstruttura e i processi alla base del pensiero astratto sono sempre gli stessi: quelli delle reti neurali. Ma cosa dicono della cognizione? *La cognizione*, secondo loro, *è un fenomeno emergente*. Con questa risposta, *i connessionisti abbandonano il campo della ricerca empirica ed entrano in quello della filosofia della mente*.

Se la cognizione è un fenomeno emergente, la questione del rapporto tra processi cognitivi, o simbolici, e processi subsimbolici va considerata nei termini in cui va trattata quella del rapporto tra il livello di analisi dei fenomeni emergenti e quello del sistema ad essi associato (inteso come tutto o come parti).

Secondo il gruppo PDP, questa

> relazione, in sostanza, è analoga a quella tra il livello di analisi della dinamica dei fluidi ed il livello di descrizione soggiacente che chiama in causa la meccanica statistica. E' spesso utile teorizzare di turbolenze e di differenti tipi di turbolenze, e una descrizione del genere andrà bene per molti scopi. Tuttavia, possiamo spesso imbatterci in fenomeni cui le nostre descrizioni di alto livello non sono adeguate, sicché è necessario, se si vuole comprendere il comportamento del sistema, descriverlo nei termini dei processi soggiacenti. Alla luce di questo esempio, possiamo illustrare anche la nostra interpretazione delle proprietà emergenti. La turbolenza non è predicibile a partire dalla conoscenza degli elementi del sistema; essa dipende dalle interazioni tra questi elementi. Analogamente, non crediamo che il livello di analisi appropriato sia quello dell'attività delle singole unità. Le proprietà della rete «emergono» dall'interazione tra gli elementi [Rumelhart *et al.* 1986b, 311-2].

L'applicazione alle reti neurali del concetto di emergenza discende indubbiamente dalla concezione di John Hopfield, ma qui diventa qualcosa di più:

> In generale, noi crediamo che i fenomeni cognitivi emergano dall'interazione di grandi insiemi di unità. Possiamo dunque considerare il livello di analisi simbolico come un'approssimazione del sistema soggiacente. In molti casi, queste approssimazioni si riveleranno utili; in altri, esse saranno sbagliate, e per comprendere il comportamento del sistema saremo costretti a passare al livello delle unità [Rumelhart *et al.* 1986b, 312].

Si potrebbe pensare che l'applicazione alla mente del concetto di emergenza sia, per autori abituati a trattare i fenomeni mentali in termini di algebra vettoriale, formale e sofisticata. Fortunatamente (o sfortunatamente?) non è così.

Rumelhart e McClelland presentano la loro adesione all'emergentismo con le seguenti parole, in una pagina così densa che non lascia "il tempo per respirare":

> Noi crediamo certamente ai fenomeni emergenti, nel senso di fenomeni che non potrebbero essere mai capiti o predetti da uno studio degli elementi inferiori isolati. Questi fenomeni sono funzioni dei particolari tipi dei raggruppamenti delle unità elementari. In generale, è utile un vocabolario nuovo per parlare di fenomeni aggregati, piuttosto che di caratteristiche di elementi isolati. Ciò vale in molti campi. Per esempio, potremmo non conoscere i diamanti partendo dallo studio di atomi isolati; potremmo non capire la natura dei sistemi sociali, partendo dallo studio di individui isolati; e potremmo non capire il comportamento delle reti di neuroni, partendo dallo studio di neuroni isolati. Delle caratteristiche come quella della durezza dei diamanti non sono comprensibili attraverso le interazioni degli atomi di carbone e del modo in cui si allineano. Il tutto è differente dalla *somma* delle parti. Tra queste vi sono interazioni non lineari. Ciò comunque non indica che la natura degli elementi di livello inferiore non sia rilevante per i livelli superiori di organizzazione; al contrario, il livello superiore, a quanto crediamo, va capito in primo luogo attraverso lo studio delle interazioni tra le unità di livello inferiore. I modi in cui le unità interagiscono non sono predicibili dagli elementi del livello inferiore, studiati isolatamente. Quello che è comunque predicibile è *se* parte del nostro studio implica le interazioni tra queste unità di livello inferiore. *Possiamo* capire perché i diamanti sono duri, non come fatto isolato, ma perché capiamo come gli atomi di carbonio possono allinearsi a formare un reticolo perfetto. E' questa una caratteristica dell'aggregato, non dei singoli atomi, ma le caratteristiche degli atomi sono necessarie per capire il comportamento dell'aggregato. Sinché non capiamo questo, rimaniamo con l'enunciato insoddisfacente che i diamanti sono duri, punto e basta. Il che è un fatto utile, ma non è una spiegazione. Analogamente, a livello sociale le organizzazioni non possono essere capite se non si capiscono gli individui che le compongono. Avere conoscenze sugli individui ci dice poco sulla struttura dell'organizzazione, ma noi non possiamo *capire* la struttura delle organizzazioni di livello superiore senza sapere molto degli individui, e di come funzionano. Questo è il senso dell'emergenza che ci fa sentire a nostro agio. Riteniamo che sia del tutto coerente con la concezione PDP dei processi cognitivi [Rumelhart - McClelland 1986a, 177-8].

Questa pagina sembra nata da una collaborazione tra tutti gli emergentisti, da Morgan a Bechtel e Richardson. Proviamo a ripercorrerne gli elementi principali.

Innanzitutto, il concetto di emergenza utilizzato da Rumelhart e McClelland è perfettamente compatibile con la definizione "unificata" che abbiamo proposto in §4.6: nessuna delle due versioni di E. Nagel, per intenderci, andrebbe altrettanto bene. Vediamo perché: i fenomeni emergenti sono funzioni di *particolari* tipi di raggruppamenti delle unità elementari, quindi sono *indipendenti* da queste ultime e inoltre *non potrebbero essere predetti* a partire dallo studio delle unità elementari considerate isolatamente. Vista la precisazione: *particolari* tipi di raggruppamenti, i fenomeni emergenti non potrebbero essere predetti neppure a partire dalle unità elementari considerate in *qualsiasi* raggruppamento; poiché viene usato il condizionale (*non potrebbero*) e data la presenza dell'avverbio *mai*, se ne deduce che l'impredicibilità vale in linea di principio: tutto questo ci riporta immediatamente alla definizione di Broad. Tuttavia, poiché non si parla di indeducibilità, possiamo escludere l'interpretazione di Nagel e Stephan e considerare l'impredicibilità di Rumelhart e McClelland più vicina al concetto di imprevedibilità usato da Morgan che non all'irriducibilità nomologica. La definizione connessionista fa poi riferimento all'utilità di un *vocabolario nuovo*: si ha novità qualitativa, dunque, e occorre un modello che utilizzi un linguaggio "più alto" rispetto al linguaggio utilizzato dal modello che descrive gli elementi isolati. Con ciò, *tutti e soli i requisiti richiesti dalla definizione di emergenza che abbiamo proposto sono richiesti anche dalla definizione connessionista*.

Inoltre, nella pagina citata, si trovano numerosi riferimenti impliciti alla teoria gerarchica del reale ed è importante osservare bene in che modo i coordinatori del gruppo PDP utilizzano tale teoria. Il riferimento più illuminante è il seguente:

> potremmo non conoscere i diamanti partendo dallo studio di atomi isolati; potremmo non capire la natura dei sistemi sociali, partendo dallo studio di individui isolati; e potremmo non capire il comportamento delle reti di neuroni, partendo dallo studio di neuroni isolati [Rumelhart - McClelland 1986a, 177].

In questo passo, gli autori richiamano l'attenzione su almeno due gerarchie di livelli di organizzazione: la prima di sistemi semplici (atomi, cristalli, ecc.), la seconda di

sistemi complessi (cellule, sistemi di cellule, individui completi, sistemi sociali); la struttura dell'anafora indica però l'attenzione per i sistemi neurobiologi dotati di neuroni organizzati in reti, cioè sistemi nervosi probabilmente molto complessi (forse umani), per i quali si potrebbe pensare di riservare un sotto-ordine di complessità specifico. Ma Rumelhart e McClelland ci dicono ancora di più: essi ci dicono che - focalizzata l'attenzione su un sistema (sia esso diamante, rete neuronale, o sistema sociale) - esso va studiato anche a un livello *inferiore* di analisi (atomi isolati, neuroni isolati, individui isolati). Se poi ammettiamo che i fenomeni emergenti vadano studiati a un livello *superiore*, rispetto a quello del sistema come totalità, ritroviamo qui i tre livelli di Morin: organizzazione-globalità-emergenza. Come per facilitare l'associazione con la teoria di Morin, inoltre, Rumelhart e McClelland asseriscono che il tutto è *differente* dalla somma delle parti (si ricordi che Morin sostiene che il tutto è più, *ma anche meno*, di tale somma).

Come se non bastasse, nel "manifesto" emergentista del gruppo PDP si dice che, tra le parti di un sistema a cui è associato qualche fenomeno emergente (come si presume dal contesto), vi sono *interazioni non lineari*. Secondo Bechtel e Richardson, come si ricorderà, l'emergenza è una conseguenza delle interazioni non lineari tra le parti di un sistema; è forse forzato ritrovare qui la stessa idea? Decisamente no: secondo Rumelhart e McClelland, infatti, le «*proprietà emergenti si riscontrano ogni qual volta abbiamo delle interazioni non lineari*» [Rumelhart - McClelland 1986a, 179].

Queste sono solo alcune delle idee presenti nel passo citato; non sono le uniche a essere interessanti, ma per ragioni di spazio dobbiamo fermarci[6]. Ci sembra tuttavia di aver argomentato a sufficienza a favore della seguente tesi: l'applicazione del concetto di emergenza alla mente e l'adozione della teoria gerarchica del reale, da parte del gruppo PDP, sono il risultato di una *profonda vicinanza epistemologica tra i*

[6]Osserviamo solo che i connessionisti sembrano voler tracciare una distinzione tra due modalità conoscitive, più o meno *soddisfacenti* (per esempio, la semplice constatazione che i diamanti sono duri è insoddisfacente), e l'esigenza di una simile distinzione è stata avvertita, forse per ragioni diverse, anche da chi scrive. Nel passo riportato, per la verità, tale distinzione non è molto limpida: Rumelhart e McClelland si limitano a utilizzare due verbi diversi per la conoscenza dei sistemi analizzati a differenti livelli (*capire*, per i livelli superiori, e *sapere*, per quelli inferiori). Altrove, tuttavia, la spiegazione viene esplicitamente distinta dalla comprensione: «*Il vero carattere della scienza cognitiva consiste nel tentativo di spiegare i fenomeni mentali attraverso la comprensione dei meccanismi che sono alla loro base*» [Rumelhart - McClelland 1986a, 168].

connessionisti e gli emergentisti. Inoltre, poiché il connessionismo - come tutte le discipline della terza cultura che studiano il cervello e la mente - condivide la prospettiva evoluzionistica, riteniamo corretto affermare che la filosofia della mente adottata dai connessionisti, o almeno dal gruppo PDP, è proprio quella *emergentista*.

Stabilito che i processi cognitivi sono fenomeni emergenti, i connessionisti hanno anche affrontato il problema di *come* questa emergenza possa rendere le reti neuronali capaci di quelle attività mentali di alto livello, come il ragionamento sequenziale e conscio, che permettono agli esseri umani di risolvere problemi nei quali devono essere soddisfatte esplicite regole logico-matematiche. A questo proposito il gruppo PDP ha formulato una teoria che è, da un lato, basata sull'interpretazione del comportamento effettivo delle reti neurali e, dall'altro lato, estremamente speculativa e svincolata «*da qualsiasi particolare modello*» [Rumelhart *et al.* 1986b, 303].

La prima parte di questa teoria, che chiameremo semplicemente "teoria della cognizione", afferma che gli esseri umani sono in grado di *percepire* la soluzione di qualsiasi problema logico, purché ridotto ai minimi termini. Prima di vedere il resto della teoria, cerchiamo di renderne trasparente questa prima parte.

8.4 Dal riconoscimento percettivo alla soluzione di problemi

Per esaminare nel modo migliore la teoria della cognizione, è opportuno considerare le reti neurali come *sistemi dinamici complessi* [Smolensky 1988, 68-9]. Questo modo di vedere è esattamente lo stesso con cui Kauffman e i suoi colleghi del Santa Fe Institute guardano alle *proprie* reti (di lampadine, di agenti chimici, ecc.).

Le stesse reti di Kauffman possono essere considerate come *reti neurali preventivamente addestrate* - non differiscono molto, in particolare, dalle reti neurali di Hopfield: gli stati di attivazione (cioè le scariche) dei processori assumono solo *due valori*, esattamente come nelle reti di Hopfield; inoltre, le regole di logica booleana in base alle quali i processori assumono i loro stati di attivazione non sono altro che altrettante *regole di aggiornamento*: nelle reti di Kauffman un'unità inattiva si attiva se, per esempio, quelle ad essa connesse sono tutte attive (congiunzione logica); analogamente, nelle reti di Hopfield, un'unità inattiva si attiva se, per esempio, il suo

ingresso-netto, cioè la somma pesata delle scariche delle unità ad essa afferenti, supera una certa soglia.

Cosa significa considerare le reti neurali sistemi dinamici complessi? Significa, molto semplicemente, *concepire l'elaborazione distribuita in parallelo come una traiettoria nello spazio degli stati*. Lo spazio degli stati (o delle fasi) di una rete neurale con N unità di elaborazione è uno spazio matematico a N dimensioni, in cui ogni punto specifica lo stato di attivazione di tutte le unità. L'elaborazione delle informazioni fornite alla rete, come sappiamo, corrisponde all'aggiornamento, nel tempo, dello stato di attivazione delle unità. Allora ogni cambiamento (piccolo o grande) nello stato di attivazione di qualche unità può essere immaginato come un tratto (piccolo o grande) di traiettoria nello spazio degli stati.

Ora, lo stato iniziale del sistema è imposto dallo sperimentatore: di solito si assegnano certi valori (interpretabili semanticamente) allo stato di attivazione di alcune unità (chiamate *unità di ingresso*) e si assegnano stati di attivazione casuali a tutte le altre. Questo equivale a stabilire il punto di partenza della traiettoria nello spazio delle fasi. La rete comincia subito a elaborare le informazioni così ricevute e la traiettoria descrive un tracciato; infine, quando l'evoluzione si arresta, lo sperimentatore rileva i valori di scarica di determinate unità (dette *unità di uscita*) e li interpreta semanticamente. Questa fase viene descritta dicendo che la rete si è "assestata" su una soluzione. A questo punto le domande che sorgono sono quasi scontate: quali regole governano l'evoluzione del sistema? Ci sono degli *attrattori* nello spazio degli stati di una rete neurale? Sono interpretabili semanticamente? E ancora: qualunque zona dello spazio delle fasi è accessibile?

Il gruppo PDP si è occupato a fondo di questi quesiti (che sono ovviamente gli stessi che si pongono i ricercatori del Santa Fe). Per ragioni di spazio, possiamo affrontare solo la prima delle suddette domande: quali regole governano l'evoluzione del sistema? Smolensky ha individuato due classi di condizioni che influenzano la traiettoria della rete nel suo peregrinare attraverso lo spazio degli stati: i *vincoli morbidi* e i *vincoli rigidi*.

I vincoli *morbidi* e i vincoli *rigidi* sono esattamente analoghi a ciò che Morin chiama rispettivamente vincoli *materiali* (parti → tutto) e vincoli *organizzativi* (tutto

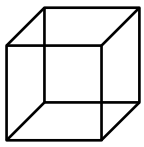
→ parti). La differenza è che, mentre i vincoli di cui parla Morin hanno lo statuto di concetti filosofici, i vincoli di Smolensky sono valori numerici (quelli morbidi) e possono essere equazioni (quelli rigidi).

I *vincoli morbidi* (o *sfumati*) sono, senza possibilità di equivoco, i pesi sulle connessioni. In essi, secondo i connessionisti, risiede tutta la conoscenza della rete, che è stata codificata e immagazzinata in tale formato (subsimbolico) durante la procedura di apprendimento:

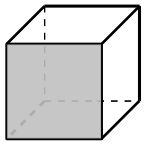
> Una connessione positiva (eccitatoria) fra l'unità *a* e l'unità *b* rappresenta un vincolo sfumato che comporta che, se *a* è attiva, debba essere attivata anche *b*. Una connessione negativa (inibitoria) rappresenta il vincolo opposto. Il valore numerico di una connessione rappresenta la forza del vincolo [Smolensky 1988, 103].

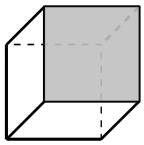
I *vincoli rigidi* sono invece le condizioni affinché il risultato dell'elaborazione, interpretato semanticamente, possa essere giudicato corretto dallo sperimentatore. Per esempio, se la rete ha codificato la rappresentazione del "cubo impossibile" di Necker, essa - indipendentemente dall'ingresso - è *vincolata rigidamente* ad assestarsi su una delle due configurazioni interpretabili semanticamente come "cubo possibile" [Rumelhart *et al.* 1986b, 257-62]:

Cubo impossibile di Necker:



Cubi possibili:





Un vincolo rigido può anche essere un'equazione matematica: in questo caso l'interpretazione semantica delle scariche delle unità di uscita deve essere una soluzione dell'equazione. Un esempio di questo tipo è fornito dallo stesso Smolensky. Egli ha costruito una rete che simula il funzionamento di un circuito elettrico. I vincoli rigidi del problema sono le leggi dei circuiti: la legge di Ohm, la legge di Kirchoff, ecc. [Smolensky 1988, 106-7].

Se per uno stesso problema ci sono *più* vincoli rigidi, diremo che la rete si è assestata su una soluzione "locale" se l'interpretazione semantica della configurazione di uscita soddisfa *alcuni* vincoli rigidi. Diremo invece che la rete si è assestata sulla soluzione "globale" se sono soddisfatti *tutti* i vincoli rigidi del problema.

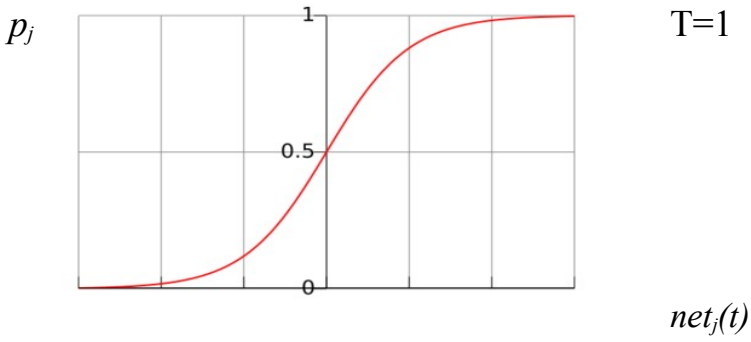
Si può dimostrare che alcune reti neurali sono in grado di assestarsi sulla soluzione globale di qualsiasi problema di riconoscimento, purché *ben posto* (i vincoli rigidi devono essere non-contradditori)[7]. Tali reti neurali sono i cosiddetti "modelli stocastici", caratterizzati dal fatto che la regola di aggiornamento non stabilisce, per ogni unità, quale sarà il suo stato di attivazione, e quindi la scarica, nell'istante successivo; stabilisce invece qual è la *probabilità di scarica*, nell'istante successivo, di ogni unità.

Data l'importanza dei modelli stocastici, è bene essere un po' più precisi. Tali reti hanno le seguenti caratteristiche:

- le unità di elaborazione possono assumere solo i valori 1 e 0 (come nelle reti di Kauffman e in quelle di Hopfield); per semplicità le penseremo come lampadine (accese/spente);
- le reti sono caratterizzate da un parametro T, regolabile a piacere dallo sperimentatore, chiamato *temperatura computazionale*;
- indicando con $net_j(t)$ l'ingresso-netto ricevuto dalla j-ima lampadina, la probabilità p_j che all'istante t essa risulti accesa è

$$p_j = 1 / (1 + e^{- net_j(t) / T})$$

il cui grafico è il seguente [cfr. Rumelhart *et al.* 1986, 108]:



[7]In realtà, quello che si dimostra è che *la probabilità* che la rete si assesti sulla soluzione globale *tende* a 1.

Al crescere di T la probabilità tende a 0.5, indipendentemente dal valore dell'ingresso-netto; invece, quando T tende a 0, la probabilità assume un andamento a gradino: se l'ingresso-netto è positivo, la lampadina è sicuramente accesa - altrimenti è sicuramente spenta.

Ora, come abbiamo detto, questo tipo di reti - completata la fase di apprendimento - è in grado di assestarsi sulla soluzione globale di qualsiasi problema di riconoscimento[8].

Com'è possibile, in generale, che una rete raggiunga stati interpretabili semanticamente come soluzioni di un problema rigidamente vincolato? La risposta è relativamente semplice: durante la fase di apprendimento, *i vincoli morbidi "incorporano" quelli rigidi*. Questa, fondamentalmente, è la ragione per cui l'algoritmo di apprendimento delle reti è al centro dell'attenzione in tutta la letteratura connessionistica, nonostante il fatto che esso non faccia parte della simulazione vera e propria.

I connessionisti hanno inventato molti algoritmi di apprendimento, ciascuno dei quali "obbliga" i vincoli morbidi a incorporare quelli rigidi. Per esempio, nella retropropagazione dell'errore (o regola delta generalizzata), è la forza degli esempi a creare la giusta "pressione" sui vincoli morbidi. Naturalmente, sostenere che i vincoli morbidi incorporano quelli rigidi è solo *un altro modo di dire* che la "conoscenza" della rete viene immagazzinata nelle connessioni. L'unica differenza è che, nel caso dei problemi vincolati, la "conoscenza" non riguarda rappresentazioni, bensì regole logico-formali. Ora, l'elaborazione delle informazioni non differisce nei due casi: si tratta comunque di elaborazione distribuita in parallelo. Tuttavia, se la conoscenza riguarda rappresentazioni, allora la configurazione di uscita è interpretabile come una discriminazione e quindi la rete ha simulato un riconoscimento percettivo; invece, se la conoscenza riguarda regole logico-formali, allora la configurazione di uscita è interpretabile come soluzione di un problema vincolato.

[8]Questo eccezionale risultato, però, può essere raggiunto solo se lo sperimentatore segue scrupolosamente la seguente procedura (che "simula" la *tempratura* dei metalli): l'elaborazione deve iniziare con una temperatura computazionale elevata, condizione nella quale tutte le unità di elaborazione hanno, grosso modo, uguale probabilità di essere attivate. Poi lo sperimentatore abbassa gradualmente la temperatura; se il raffreddamento computazionale è *molto* lento, la rete si assesta sulla soluzione globale, altrimenti si assesterà su qualche soluzione locale.

Si noti che i vincoli rigidi, essendo incorporati in quelli morbidi, acquistano a loro volta una notevole "morbidezza"; per esempio, nella rete di Smolensky che simula il funzionamento di un circuito elettrico, il

> sistema può contravvenire alla legge di Ohm se deve farlo, ma se non ne ha bisogno questo non succede. Al di fuori del dominio idealizzato dei problemi ben posti e di un tempo di elaborazione illimitato, il sistema offre comunque prestazioni notevoli. Non è fragile come i sistemi di inferenza simbolici. Se al sistema viene presentato un problema mal posto, esso cerca di soddisfare il maggior numero di vincoli possibile. Se gli viene fornita informazione incoerente, esso non crolla e non deduce semplicemente qualsiasi cosa. Se gli viene data una quantità di informazione insufficiente, non si blocca senza dedurre niente. Dato un tempo di elaborazione limitato, le prestazioni diminuiscono gradualmente. Tutte queste caratteristiche emergono «gratis» come conseguenza automatica della produzione di inferenze in un sistema subsimbolico; non vengono aggiunti meccanismi ulteriori per gestire le deviazioni dalle circostanze ideali [Smolensky 1988, 107].

Come si applica tutto questo alle reti neuronali? Secondo i connessionisti, queste osservazioni suggeriscono l'ipotesi secondo cui le reti neuronali *apprendono a percepire le soluzioni dei problemi vincolati*, esattamente come apprendono il riconoscimento percettivo degli stimoli ambientali. Per esempio, consideriamo l'operazione 3×2: si tratta di un problema rigidamente vincolato alle regole della moltiplicazione; un computer lo risolverebbe attraverso un'opportuna manipolazione simbolica (il simbolo «3×2» diventa «+ 3» ripetuto 2 volte, poi ciascun «+ 3» diventa «+1» ripetuto 3 volte). Secondo i membri del gruppo PDP, però, un bambino che va a scuola impara a risolvere il problema in modo diretto: impara, cioè, che *vicino a* «3×2» *bisogna scrivere* «6». Le reti neuronali del suo cervello, in altre parole, si assestano sulla soluzione globale, «6», dopo aver elaborato l'informazione in ingresso, «3×2», *nello stesso modo* in cui avrebbero elaborato un qualsiasi altro stimolo percettivo. Non c'è alcuna manipolazione simbolica, né applicazione di regole logiche: le reti neuronali *percepiscono* la soluzione del problema [Rumelhart *et al.* 1986b, 299].

8.5 Conclusioni e osservazioni critiche

Quella che abbiamo visto nel paragrafo precedente è la prima parte della teoria della cognizione. La seconda parte, come avevamo anticipato, è estremamente speculativa e, in verità, anche piuttosto vaga:

> L'idea di fondo è che la nostra capacità di risolvere problemi logici non è tanto basata sull'uso della logica, quanto sulla riduzione dei problemi che vogliamo risolvere a problemi che siamo capaci di risolvere. Gli esseri umani sembrano possedere tre fondamentali capacità che consentono loro di pervenire a conclusioni logiche senza essere logici [Rumelhart *et al.* 1986b, 297].

Queste tre capacità sono le seguenti:

1. Capacità di riconoscimento percettivo;
2. Capacità di elaborare delle aspettative;
3. Capacità di manipolare l'ambiente.

Grazie alla prima capacità, come abbiamo visto, gli esseri umani sono in grado di *percepire le soluzioni* dei problemi semplici. Con la seconda capacità gli esseri umani interiorizzano le proprie esperienze; ciò consente loro di compiere, successivamente, quelle *simulazioni mentali* indispensabili per la sopravvivenza di tutti gli organismi per i quali l'apprendimento svolge un ruolo cruciale (gli animali dotati di sistema nervoso plastico). La terza capacità consente agli esseri umani di creare *rappresentazioni esterne* dei problemi complessi: è questa l'abilità necessaria per risolvere i problemi complessi (le soluzioni dei quali non sono immediatamente percepibili). Nelle parole dei membri del gruppo PDP:

> In sostanza, la nostra concezione è questa: noi siamo capaci di «percepire» le soluzioni dei problemi. Sfortunatamente, non c'è un meccanismo universale per risolvere i problemi e pensare; piuttosto, diventando più esperti, diventiamo più abili nel ridurre i problemi a compiti di corrispondenza tra pattern [...]. Così, un esperto giocatore di scacchi, di fronte ad una scacchiera, può «vedere» la mossa giusta. Questo, a nostro giudizio, è un problema del tutto simile a quello di percepire qualcosa. [...] Resta comunque vero che non tutti i problemi possono essere risolti «vedendo» immediatamente la risposta. Così, pochi [...] sono in grado, guardando una moltiplicazione a tre cifre (come 343 per 822), di vederne la soluzione. [...] Diventa critica qui la nostra capacità di manipolare l'ambiente [Rumelhart *et al.* 1986b, 298-9].

Manipolando l'ambiente (capacità 3), ad esempio scrivendo su carta i due numeri, uno sotto l'altro (*rappresentazione esterna del problema*), possiamo ridurre il problema ad una serie di operazioni ciascuna alla portata della capacità di riconoscimento percettivo del nostro cervello. Cioè possiamo *percepire* (capacità 1) che sotto la colonna del 3 e del 2 bisogna scrivere «6»:

$$\begin{array}{r} 343 \times \\ 822 = \\ \hline 6 \end{array}$$

Poi, possiamo *percepire* che a sinistra del 6 bisogna scrivere «8». E così via, continuando a modificare la rappresentazione esterna del problema, fino alla sua soluzione "globale". L'esperienza di questa operazione viene in qualche modo *interiorizzata* aggiornando la nostra rappresentazione interna del mondo (capacità 2) e rendendo la moltiplicazione, concretamente risolta, disponibile per future simulazioni mentali:

> Via via che facciamo esperienza del mondo creato dalle nostre (e dalle altrui) azioni, noi formiamo dei modelli interni di queste rappresentazioni esterne. Possiamo così immaginare di essere di fronte ad una moltiplicazione e possiamo immaginare di eseguirla. Se l'operazione è sufficientemente semplice, possiamo portarla a termine nella nostra immaginazione [Rumelhart *et al.* 1986b, 300].

Secondo il gruppo PDP, le tre capacità appena esaminate spiegano *perché gli esseri umani* - nonostante il fatto che l'elaborazione che avviene nel loro cervello, invece di essere l'applicazione rigorosa di regole logiche, corrisponde al banale sviluppo di una traiettoria attraverso lo spazio degli stati - *sono stati in grado di produrre la logica, la matematica e le altre discipline della nostra cultura teoretica*.

Questa è, in breve, la teoria della cognizione del gruppo PDP. Nel presente paragrafo faremo alcune osservazioni sia su questa teoria, sia sull'approccio connessionista al problema della percezione (dal quale inizieremo immediatamente).

Per quanto riguarda l'approccio connessionista alla percezione, dobbiamo metterne in evidenza la debolezza maggiore: la questione dell'interpretazione semantica. Questa questione non va confusa con la cosiddetta "questione semantica", che è un problema legato all'"ingombrante" presenza dello sperimentatore che interpreta semanticamente le configurazioni di ingresso e uscita delle reti neurali: le

reti neuronali non sembrano richiedere alcun osservatore che dia significato alla loro attività (casomai la loro attività stessa *è* un osservatore); in altre parole, mentre una rete virtuale elabora informazioni *interpretabili semanticamente* (da un osservatore), una rete naturale sembra poter elaborare informazioni *dotate di significato* (senza osservatore). Tuttavia, la questione semantica, così presentata, non è ben posta. E' la mente, e non le reti neuronali, che "elabora" informazioni dotate di significato indipendentemente da qualsiasi osservatore. Le reti, siano esse biologiche o virtuali, ricevono e forniscono informazioni a cui *può essere assegnato* un significato - ma per fare questo ci vuole, appunto, una mente.

La questione dell'interpretazione semantica è comunque un problema diverso. Come sappiamo, le configurazioni di ingresso e uscita di una rete neurale sono interpretabili semanticamente. Il significato che attribuiamo a tali configurazioni, che di solito sono vettori numerici, è del tutto *arbitrario*. L'unica condizione a cui l'interpretazione semantica deve sottostare è la coerenza: l'interpretazione di ingressi e uscite, durante la simulazione, deve essere la medesima adottata durante la precedente fase dell'apprendimento. Esclusa la condizione della coerenza, l'interpretazione semantica è a completo arbitrio dello sperimentatore: questo fatto stabilisce una differenza tra l'interpretazione semantica delle reti neurali e quella delle reti neuronali, differenza che non è quella evidenziata dalla questione semantica.

La differenza tra reti neurali e neuronali consiste nel fatto che *l'interpretazione semantica delle reti neuronali non è affatto arbitraria*. Lo sperimentatore scopre, e non lo decide, il "significato" dell'attività dei neuroni reali. E lo scopre empiricamente, stimolando le reti neuronali oppure osservandone l'attività mediante tecniche di visualizzazione. Si pensi, per esempio, alle famose scoperte del neurochirurgo Wilder Penfield: i suoi esperimenti lo condussero *necessariamente* a certe interpretazioni semantiche. Le reazioni dei suoi stessi pazienti, dei quali egli studiò la corteccia cerebrale, fornivano il significato che egli *doveva* associare all'attività della corteccia. Penfield, per esempio, individuò il famoso *homunculus* senso-motorio, che non è affatto associabile a un'interpretazione semantica *arbitraria*: tutti gli esperimenti "obbligano" ad interpretare l'attività di tale parte della corteccia come una mappa topografica senso-motoria della superficie corporea.

Una rete neurale, al contrario, non fornisce allo sperimentatore alcun significato. Consideriamo, per esempio, una rete neurale collegata al sonar di un sottomarino, in grado di percepire la differenza tra configurazioni interpretabili come «mine» e configurazioni interpretabili come «rocce». Che cosa ci potrebbe dire che essa percepisce *proprio tale differenza*, se non lo sapessimo? La rete è muta, è passiva. In che modo possiamo "farla parlare"?

Se lo sperimentatore modifica l'interpretazione semantica - e può farlo arbitrariamente in qualunque momento - la rete deve essere "riaddestrata", cioè il programma di simulazione deve eseguire nuovamente l'algoritmo di apprendimento. In generale, se una rete viene riaddestrata, la nuova distribuzione di pesi sulle connessioni sarà diversa da quella precedente; *questa differenza corrisponde al cambiamento dell'interpretazione semantica* e, di solito, è piuttosto radicale - ma può anche succedere che sia minima. La questione dell'interpretazione semantica si traduce nel fatto che non sappiamo nulla di questa corrispondenza; la constatiamo, punto e basta. Constatiamo, in altre parole, che ogni

> rappresentazione subsimbolica (cioè connessionista) incorpora naturalmente un tipo di *metrica semantica* (devo questo termine ad Andler [...]), che alimenta le caratteristiche distintive di generalizzazione, degrado graduale e così via. Il modo migliore di intendere la metrica semantica è di intenderla come un arrangiamento spaziale delle unità, in uno spazio multidimensionale organizzato in modo tale che elementi semanticamente correlati siano codificati da unità-di-tratto spazialmente correlate [Clark 1989, 257].

Il profilo della metrica semantica incorporata in una rete neurale può essere rivelato da una tecnica, detta *analisi dei raggruppamenti*, che mostra in quali sottospazi si è suddiviso lo spazio degli stati della rete [Churchland 1989a, 134]. Questa analisi è molto interessante e ci permette di stabilire, per esempio nel caso della rete "mina/roccia", che essa riconosce la differenza tra *due* classi di elementi. Ma *quali* siano le due classi, non lo si può sapere.

I connessionisti probabilmente hanno *compreso* (cioè hanno capito in maniera intuitivamente soddisfacente) i processi su cui si basa la percezione, ma non hanno *spiegato* la percezione - e non avranno una spiegazione (cioè una conoscenza completa e completamente soddisfacente) di essa finché non sapranno stabilire *che cosa* percepisce effettivamente una rete addestrata. Ma, per raggiungere questo

obiettivo (sempre che sia possibile), può darsi che occorrano modelli del tutto diversi da quelli odierni. Quel che è certo è che attualmente nulla ci permette di capire che cosa percepiscono i modelli dei connessionisi, ovvero *non c'è nulla che conferisca necessità ad una interpretazione semantica*. Quella dei connessionisti, pertanto, non dovrebbe essere considerata una *spiegazione* della percezione (neppure di quella "di basso livello", come il mero riconoscimento percettivo).

Le cose non vanno meglio con la teoria della cognizione. Secondo il gruppo PDP, le capacità di riconoscimento percettivo, simulazione mentale e manipolazione ambientale sono direttamente responsabili di ogni attività mentale cognitiva. Ora, poiché la capacità di simulazione mentale si basa essenzialmente sulla *costruzione di rappresentazioni interne* del mondo e poiché la manipolazione ambientale consiste nella *costruzione di rappresentazioni esterne* del mondo, possiamo modificare le tre capacità di cui sopra in:

1'. Capacità di riconoscimento percettivo;
2'. Capacità di costruire rappresentazioni interne;
3'. Capacità di costruire rappresentazioni esterne (!).

Come abbiamo visto, la prima capacità si basa sulla modalità computazionale di elaborazione distribuita in parallelo, mentre la seconda corrisponde al meccanismo di immagazzinamento della conoscenza nelle connessioni; a rigore, solo la prima capacità viene *simulata* con le reti neurali, perché le rappresentazioni non vengono codificate dalle reti neurali *nello stesso modo* in cui lo fanno le reti neuronali (cioè il *meccanismo* è il medesimo, ma differiscono le *modalità*). Inoltre, nessuna rete neurale, sia chiaro, simula qualcosa che si avvicini (anche vagamente) alla terza capacità.

Cercando di immaginare un cervello con queste tre capacità, gli emergentisti devono chiedersi: *che tipo di mente potrebbe emergere da un simile cervello*? Rispondere è un po' imbarazzante, perché, per quanto riguarda le prime due capacità, possiamo forse immaginare un cervello con la stessa complessità funzionale delle reti neurali del gruppo PDP, ma la terza capacità richiede un salto di complessità immenso. Ciò diventa evidente se ripensiamo all'evoluzione della mente: le prime due capacità sono alla portata di menti semplici, come quella episodica delle antropomorfe o quella dei neonati prima ancora della ridescrizione rappresentazionale. La capacità

di costruire rappresentazioni esterne, invece, costituisce il culmine dello sviluppo cognitivo dei membri della specie *Homo sapiens sapiens* - è, cioè, la capacità che più di ogni altra richiede menti *estremamente* complesse.

Le tre capacità elencate dal gruppo PDP sono, allo stesso tempo, troppe e troppo poche: se desideriamo comprendere le menti *più semplici*, la terza capacità è fuori luogo; ma se cerchiamo di comprendere la *nostra* mente (quella di adulti *sapiens sapiens*), allora nell'elenco mancano molte capacità fondamentali. Queste ultime sono quelle che, nel corso della filogenesi, hanno probabilmente caratterizzato la mente dell'*Homo erectus* (capacità mimica) e che, nel corso dell'ontogenesi, trasformano la conoscenza implicita in conoscenza esplicita (capacità di ridescrizione rappresentazionale).

Non avrebbe senso, alla luce di quanto sopra, interpretare la teoria della cognizione come il tentativo (gravemente lacunoso) di comprendere la *nostra* mente. *Il modello connessionista va considerato un eccellente passo in avanti verso la comprensione delle menti più semplici*. Questa interpretazione è ben sintetizzata da Merlin Donald:

> Al pari di un sistema nervoso primitivo, una rete connessionista costruisce la propria versione percettiva del mondo senza fare affidamento su un sistema simbolico assegnatole da un operatore umano. Attualmente questi modelli sono piuttosto rudimentali, ma in linea di principio potrebbero essere resi molto più efficienti e in futuro potrebbero forse comprendere qualcosa di tanto complesso quanto la percezione di eventi sociali, cioè la più elevata acquisizione della mente episodica, dove gli oggetti in giustapposizione sono scomposti in una «situazione» percepita. Tuttavia non è evidente come una rete connessionista potrebbe gestire il livello di astrazione e di variabilità richiesto dalla percezione di eventi sociali, e quindi questa possibilità mi pare molto remota [Donald 1991, 424].

Non vi sono ragioni per essere pessimisti sul futuro dei modelli connessionisti, almeno finché le aspettative non superano l'obiettivo di simulare l'attività delle menti *semplici*:

> Se il connessionismo viaggia così lontano si avvicinerà al confine del mondo episodico, e a questo punto potremmo forse verificare se sia possibile affrontare qualcosa di ancora più complesso, come la rappresentazione mimica. La mente mimica è intelligente - davvero molto intelligente, se la confrontiamo con i suoi antecedenti - eppure non possiede né parole né qualcosa di equivalente. [...] Il primo modello formale di rappresentazione realmente umana, in un lontano futuro, dovrà affrontare il problema della mimica prima di considerare quello dell'invenzione linguistica. Senza dubbio ciò si rivelerà molto difficile, ma non vi è alcuna

ragione aprioristica per disperare. Se la rappresentazione di eventi può essere spiegata entro una struttura connessionistica, anche la rappresentazione mimica potrebbe essere spiegabile in modo analogo [Donald 1991, 424].

D'altra parte, conclude Donald, *considerare* «*realistica tale possibilità richiede una buona dose di fede (di cui, evidentemente, molti studiosi partecipi del sogno connessionista sono già dotati)*» [Donald 1991, 425].

Del tutto equivalenti sono le considerazioni di Karmiloff-Smith. La ricercatrice ha esaminato una delle più potenti reti neurali esistenti, costruita da J. L. Elman, la quale può codificare la rappresentazione di gran parte della grammatica inglese. Ecco le sue conclusioni, decisamente convincenti:

Tutta questa conoscenza grammaticale, che pare davvero impressionante, è soltanto implicita nelle rappresentazioni interne del sistema. Ciò non significa che essa non sia rappresentata. Come nel caso dell'apprendimento nella prima infanzia, direi che è rappresentata in un formato di livello I. Ma siamo noi, che teorizziamo dall'esterno, a usare i formati di livello E per etichettare le traiettorie [...] come nomi o verbi, soggetti o oggetti, transitivi o intransitivi, plurali o singolari, ecc. Di per sé, la rete non va mai oltre la formazione dell'equivalente di rappresentazioni stabili di livello I. In altre parole, la rete non oltrepassa spontaneamente quella padronanza comportamentale che le consente di fornire prestazioni efficienti, né ridescrive le rappresentazioni che sono registrate nelle sue traiettorie di attivazione. A differenza del bambino, la rete non si «appropria» spontaneamente della conoscenza (relativa alle diverse categorie linguistiche) che essa rappresenta, non può usare conoscenze di livello superiore, più astratte, per nessun altro scopo diverso da quello per cui è stata progettata, né può impegnarsi in un trasferimento della conoscenza da rete a rete. [...] Il concetto di nome resta sempre implicito nel sistema dinamico della rete. Analogo è l'apprendimento *iniziale* del bambino. Ma i bambini procedono poi a una ridescrizione spontanea della loro conoscenza, e questo processo pervasivo di ridescrizione rappresentazionale dà luogo alla manipolabilità e flessibilità del sistema rappresentazionale umano [Karmiloff-Smith 1992, 261-2].

In breve: non rendersi conto del fatto che le prestazioni delle reti neurali sono attualmente ancorate al riconoscimento percettivo significherebbe, molto semplicemente, ignorare un dato di fatto.

Il "sogno connessionista" di cui parla Donald ci ricorda il "sogno booleano" del cognitivismo classico: entrambi sono dovuti a una *sottovalutazione della complessità della mente e del cervello*. Forse, il risveglio dal "sogno connessionista" consisterà proprio nel capire finalmente quale sia il livello di tale sbalorditiva complessità.

Le reti neuralmente ispirate dell'IA connessionista cognitiva si collocano *a metà strada* tra i modelli neuronali dei descrittivisti e i modelli simbolici dell'IA convenzionale. Nel prossimo capitolo incontreremo un'altra classe di modelli *a metà strada* - a metà strada tra i modelli stocastici dell'IA connessionista cognitiva e i modelli simbolici dell'IA convenzionale - cioè i programmi ad *architettura emergente* creati dai membri del FARG. Ecco la loro sfida:

> Il «sogno connessionistico» (quello di modellare tutta quanta la cognizione impiegando un'architettura subsimbolica) è plausibile dal punto di vista neurologico, ma forse un po' troppo ambizioso, dato il livello di sviluppo attuale delle scienze cognitive. Se c'è una speranza di capire il modo in cui l'intelligenza emerge da miliardi di neuroni, o anche solo il modo in cui può emergere dalle reti connessionistiche, essa si basa sulla comprensione della natura dei *concetti*, entità fondamentali i cui princìpi operativi sembrano collocarsi tra quelli delle reti neurali a forte parallelismo e quelli della cognizione seriale altamente simbolica [Mitchell - Hofstadter 1993, 317-8].