

“Computo, ergo sum”

Un'introduzione ai modelli computazionali della coscienza

Marco Colombetti

Dipartimento di elettronica e informazione, Politecnico di Milano
Facoltà di scienze della comunicazione, Università della Svizzera italiana

Viola Schiaffonati

Dipartimento di elettronica e informazione, Politecnico di Milano
Dottorato di ricerca in filosofia-filosofia della scienza, Università degli studi di Genova

Sommario

Nelle scienze cognitive prevale da un quarto di secolo il paradigma computazionale: si assume cioè che i processi mentali possano essere descritti e spiegati in quanto processi di computazione. Questo punto di vista non è privo di aspetti problematici, che diventano particolarmente evidenti quando l'oggetto di studio è la coscienza: la stessa locuzione *modello computazionale della coscienza* non è immediatamente chiara e ammette, nel migliore dei casi, interpretazioni diverse. In questo articolo cerchiamo innanzitutto di chiarire i termini, esaminando alcune posizioni sulla coscienza nell'ambito del computazionalismo e analizzando il significato di termini ingannevolmente simili, quali *computabile* e *computazionale*. L'analisi ci consente di collocare alcuni approcci significativi in un quadro più trasparente e di trarre alcune conclusioni: in particolare difenderemo l'inseparabilità delle funzioni della coscienza dalla sua dimensione soggettiva ed esperienziale, e sosteneremo che l'uso troppo spigliato del termine “coscienza” nell'ambito computazionale rende ancora più inestricabile un dibattito già di per sé impegnativo.

1. Introduzione

Nelle scienze cognitive domina da un quarto di secolo il paradigma computazionale: si assume cioè che i processi mentali possano essere descritti e spiegati in quanto processi di computazione. Questo punto di vista non è privo di aspetti problematici, il più evidente dei quali è costituito da ciò che potremmo chiamare il *dilemma della coscienza*: da un lato, infatti, è difficile aggirare l'ovvia intuizione che la coscienza giochi un ruolo centrale nei processi mentali; da un altro lato, tuttavia, non è per nulla chiaro che cosa significhi considerare la coscienza come un processo computazionale.

Fino a una decina di anni or sono, non si può dire che il dilemma della coscienza abbia turbato i sonni degli scienziati cognitivi: predominava, ci sembra, un atteggiamento vagamente riduzionista o eliminativista. La coscienza era considerata, nel migliore dei casi, come un fenomeno intrattabile dal punto di vista scientifico, da cui tenersi prudentemente lontani; nel peggiore dei casi, come una pura illusione dell'uomo della strada, destinata a dissolversi alla luce della *new wave* teorica, armata degli strumenti d'analisi sviluppati dalla teoria della computazione e dall'intelligenza artificiale. La situazione comincia a cambiare all'inizio degli anni novanta, sotto la duplice pressione dei lavori di vari filosofi, e in particolare di John Searle (1992, 1997), che riportano la coscienza al centro del dibattito, nonché dei progressi delle neuroscienze, che si pongono seriamente alla ricerca dei corrispettivi neurali della coscienza (*i neural correlates of consciousness* o *NCC*). Il risultato del mutato atteggiamento è che il dilemma della coscienza non può più essere semplicemente ignorato. La buona volontà, pur sostenuta da una cospicua dose di acume, non basta però a rendere trattabile un problema intrinsecamente ostico; è sufficiente una rapida scorsa alla letteratura specializzata per constatare che quando si parla di approccio

computazionale alla coscienza le difficoltà sorgono già al primo passo, ovvero al momento di definire quali siano le domande e gli spazi ammissibili per le risposte: che cosa intendiamo esattamente per *coscienza*? Che cos'è un *modello computazionale*? In che senso un modello computazionale potrebbe dar conto dei processi mentali consci?

Nel presente articolo esaminiamo alcune risposte possibili a queste domande, avendo cura di chiarire innanzi tutto i termini della questione. Per questo motivo nel paragrafo 2 esaminiamo la definizione del termine *coscienza* nell'ottica della sua computabilità; nel paragrafo 3 analizziamo il termine *modello*; e nel paragrafo 4 approfondiamo il significato del termine *computazionale* e di alcuni termini connessi. Sulla base di queste analisi, nel paragrafo 5 descriviamo alcune fra le posizioni più diffuse riguardo al problema della modellabilità della coscienza in termini computazionali. Le nostre conclusioni, infine, sono esposte nel paragrafo 6.

2. Coscienza e computabilità: lo scenario della discussione

Nulla come la coscienza è così familiare e intimamente connesso alla nostra sfera esperienziale, eppure allo stesso tempo così complesso e inspiegabile. Date la difficoltà e la vastità dell'argomento, questo paragrafo non ha l'obiettivo di presentare un'indagine completa dello stato dell'arte, né di prendere posizione nell'una o nell'altra direzione, bensì di illustrare schematicamente lo scenario della discussione per chiarire quali siano le prospettive che danno modo di parlare di modelli computazionali della coscienza.

2.1 Una premessa

Vista la centralità del concetto di coscienza nel definire le caratteristiche degli stati mentali è piuttosto sorprendente constatare che il tema della coscienza non è stato trattato in maniera esauriente nelle scienze cognitive, almeno fino a tempi molto recenti (Bechtel, 1994). Va notato che la maggior parte dei lavori in questo ambito è stata portata avanti con un preciso obiettivo progettuale: cercare di identificare le caratteristiche funzionali nell'elaborazione mentale delle informazioni e mostrare come queste possano essere riprodotte usando determinati modelli computazionali. Considerando questa enfasi progettuale, non desta stupore che il tema della coscienza non abbia trovato adeguato spazio.

Anche a costo di notevoli sforzi e di una maggiore attenzione al problema, la coscienza non può essere definita in maniera univoca: le opinioni sono diverse, la letteratura in proposito è assai vasta e manca un accordo sostanziale anche a proposito delle definizioni di base. Nonostante la differenza dei percorsi e delle soluzioni proposte, negli autori che si occupano di coscienza emergono alcuni temi fondamentali su cui essi ritengono essenziale confrontarsi. Nella maggior parte dei casi, però, l'obiettivo non è di proporre una soluzione, ma di chiarire i termini del dibattito; tra questi: l'intenzionalità della coscienza, la coscienza come esperienza qualitativa e soggettiva, la distinzione fra stati consci e processi mentali inconsci, l'esperienza soggettiva dell'autoconsapevolezza, la capacità di un sistema di accedere ai suoi stati interni e l'integrazione delle informazioni da parte di un sistema cognitivo.

Non tratteremo qui ciascuno di questi temi perché riteniamo che la distinzione fondamentale, per chiarire che cosa significhi proporre modelli computazionali della coscienza, sia innanzitutto quella fra coscienza (*consciousness*) e autocoscienza (*self-consciousness*), cui si aggiunge in alcuni casi anche un terzo concetto, ossia la consapevolezza (*awareness*) come distinta dalla coscienza. Che questa distinzione sia fondamentale non significa in realtà che tutti gli autori se ne siano occupati

o ne abbiano tenuto conto (in alcuni casi la confusione fra questi concetti è anzi piuttosto evidente), ma piuttosto che questo tema rappresenta un nodo centrale per comprendere alcune caratteristiche dei modelli computazionali della coscienza.

2.2 Coscienza e consapevolezza

Partiamo dalla distinzione, senza dubbio interessante anche se discutibile, proposta da David Chalmers (1995) fra esperienza conscia (*conscious experience*) e consapevolezza (*awareness*): la prima fa riferimento all'aspetto soggettivo ed esperienziale della coscienza, mentre la seconda ne riguarda gli aspetti funzionali. Il concetto di coscienza come esperienza soggettiva ha le sue radici in un famoso articolo di Thomas Nagel (1974), in cui si afferma che il fatto che un organismo abbia un'esperienza cosciente significa che a essere quell'organismo si prova qualcosa. Riflettere su che cosa si provi a essere un pipistrello (da cui il curioso titolo dell'articolo di Nagel) porta alla conclusione che non è possibile descrivere uno stato soggettivo a chi non sia in grado di provarlo in prima persona. Anche per Chalmers, spiegare l'esperienza soggettiva in prima persona è lo *hard problem* della coscienza; per la sua soluzione non è sufficiente risolvere l'*easy problem*, ovvero spiegare le capacità funzionali della consapevolezza, come per esempio la capacità di distinguere, categorizzare e reagire a stimoli ambientali o la possibilità di possedere stati mentali: un conto è caratterizzare i fenomeni che portano al possesso della coscienza, un altro è sperimentare soggettivamente la coscienza.

Nonostante la profonda differenza qualitativa fra coscienza e consapevolezza esiste fra queste un'ovvia connessione, dato che con il termine 'consapevolezza' si fa riferimento ai fenomeni funzionali associati alla coscienza. In questo senso la consapevolezza è una nozione puramente funzionale, mentre la coscienza è esclusivamente soggettiva, anche se entrambe sono intimamente connesse: il *principio della coerenza strutturale* proposto da Chalmers serve proprio a stabilire una corrispondenza diretta e reciproca fra coscienza e consapevolezza, per cui ogni volta che si trovi l'una si trova anche l'altra. Una conseguenza importante di questa visione è che l'*easy problem*, ovvero la caratterizzazione della consapevolezza, potrebbe essere affrontato in termini computazionali. Chalmers avverte però che, in questo caso, non è sufficiente spiegare la funzione per dare conto dell'intero fenomeno: per spiegare l'esperienza non bastano i consueti metodi delle scienze cognitive, ma occorre un approccio completamente nuovo. L'*ingrediente extra* per una teoria dell'esperienza conscia è, secondo Chalmers, postulare l'esperienza stessa come entità fisica fondamentale, nello stesso modo in cui sono fondamentali la massa, la carica elettrica e lo spazio-tempo; da ciò sarebbe possibile derivare poi i principi psicofisici che connettono le proprietà dei processi fisici con quelle dell'esperienza e che servirebbero a spiegare in che modo l'esperienza emerga dal mondo fisico.

La distinzione fra la coscienza come funzione (l'aver coscienza di qualcosa) e la coscienza come esperienza (lo sperimentare la coscienza di qualcosa) sembra costituire una linea di demarcazione fra gli approcci funzionali e quelli che potremmo definire *esperienziali* per sottolineare la primarietà giocata in essi dall'esperienza. Postulando questa separazione netta, solo gli approcci funzionali sembrerebbero aprire le porte al trattamento del tema della coscienza attraverso modelli computazionali. Tuttavia, occorre osservare che anche nel caso dei modelli computazionali la centralità della nozione soggettiva ed esperienziale della coscienza non viene dimenticata; semmai il tentativo è di trattare l'esperienza in modo da farla rientrare nel modello proposto, allo scopo di eludere gli esiti riduzionistici più estremi.

Accanto alla soluzione di Chalmers, quindi, che rimane neutra sulla possibilità di racchiudere la coscienza nelle sue varie accezioni in un modello computazionale, altri autori cercano di includere anche l'esperienza soggettiva in un modello computazionale della coscienza. Harvey (2002), per esempio, pur accettando la distinzione fra le funzioni della coscienza e l'esperienza della coscienza posta da Chalmers, propone una soluzione che risiede in un cambio di atteggiamento da parte dell'osservatore: sono gli osservatori, secondo Harvey, che attribuiscono la presenza della coscienza, intesa nel senso di esperienza soggettiva, alle entità, laddove queste entità sembrano relazionarsi con il mondo attraverso una struttura autonoma di obiettivi e bisogni. Per questo motivo, le entità artificiali prodotte nel contesto di un approccio evuzionistico alla robotica, che sono in grado di sviluppare in modo autonomo bisogni e desideri, paiono i candidati più plausibili per l'attribuzione della coscienza intesa in questo senso (pur ammettendo che debbano passare anni per arrivare al livello di sviluppo adeguato).

La distinzione fra coscienza e consapevolezza secondo le linee proposte da Chalmers viene postulata anche da quegli autori (come ad esempio Mathis e Mozer, 1996) che, pur riconoscendone l'esistenza, non si preoccupano di spiegare l'aspetto soggettivo dell'esperienza, ma si limitano a valutare quali siano le differenze fra un sistema cognitivo che elabori informazioni e ne sia conscio e uno che non lo sia. Nell'articolo citato, in particolare, l'obiettivo è trovare un corrispettivo computazionale della coscienza, per proporre poi un'architettura cognitiva che ne dia conto. In questo modo il problema dell'esperienza soggettiva sembra essere tagliato alla radice, anche se può destare un certo stupore la facilità con cui i due aspetti della coscienza (funzionale ed esperienziale) vengono scissi e trattati separatamente, senza alcuna spiegazione preliminare di se e come nella coscienza umana essi si trovino indissolubilmente connessi.

2.3 Coscienza e autocoscienza

Oltre alla distinzione fra coscienza e consapevolezza, sembra che nei lavori esplicitamente impegnati a identificare dei corrispettivi computazionali della coscienza acquisti un ruolo decisivo anche l'*autoconsapevolezza*, che permette di tenere traccia delle azioni e dei processi compiuti, modulando le attività future sulla base di quelle passate. È proprio in questo contesto che diventa importante distinguere fra coscienza e autocoscienza o fra consapevolezza e autoconsapevolezza, tenendo presente che in molti casi le analisi tralasciano del tutto la distinzione fra coscienza e consapevolezza.

All'interno di un quadro computazionale, e partendo dal presupposto che la mente possa essere considerata un sistema biologico che elabora informazioni, Jackendoff (1987) ritiene importante distinguere fra *consapevolezza primaria* e *consapevolezza secondaria*. La consapevolezza primaria è relativa agli oggetti esterni dell'esperienza, siano essi gli oggetti percepibili del mondo, le esperienze dei nostri corpi o gli atti di immaginazione; sebbene questi ultimi possano apparire "irreali", essi formano in realtà una parte della nostra esperienza in un determinato momento come oggetti "reali" e sono quindi oggetti della consapevolezza primaria. Della consapevolezza secondaria o riflessiva, invece, fa parte un secondo piano di esperienza, che conduce alla consapevolezza di se stessi e delle proprie interazioni con gli oggetti della consapevolezza primaria. In altre parole, 'arriva un cane' è un resoconto di consapevolezza primaria, 'vedo arrivare un cane' è invece un resoconto di consapevolezza secondaria, laddove il passaggio dal primo al secondo resoconto risiede nell'ingresso di colui che ha l'esperienza.

Molto simile è la distinzione proposta da Lloyd (1995) fra *coscienza* o *consapevolezza primaria* e *coscienza* o *consapevolezza riflessiva*: la prima diretta verso il mondo (che comprende anche il proprio corpo), la seconda verso i propri stati di consapevolezza. Naturalmente la coscienza primaria è centrale da un punto di vista sia costitutivo sia esplicativo: senza coscienza primaria non è possibile una coscienza riflessiva, che non avrebbe nulla su cui riflettere; inoltre, una volta che la coscienza primaria sia stata compresa è più facile afferrare anche la coscienza riflessiva come caso particolare. Lo scenario di queste riflessioni è un modello computazionale di tipo connessionista, all'interno del quale si vuole mantenere la differenza fra i due livelli della coscienza. Quindi se gli stati della coscienza coincidono, in ultima analisi, con stati cerebrali, si deve poter scorgere anche in questi ultimi una differenza fra gli stati cerebrali corrispondenti alla coscienza primaria e gli stati corrispondenti alla coscienza riflessiva.

Un'altra soluzione adottata consiste nell'assimilare la *conoscenza conscia* alla *conoscenza introspettiva*, ossia all'informazione conscia che è distinta da quella inconscia per il fatto di essere osservabile, come suggerisce per esempio McCarthy (1999). Qui l'attenzione è posta sulla necessità da parte di un robot di utilizzare la conoscenza di tipo introspettivo per operare all'interno del mondo del senso comune: un certo grado di conoscenza di se stessi e dei propri stati mentali viene reputato essenziale per portare a termine compiti che richiedono un'intelligenza di tipo umano. Sebbene McCarthy tenti anche una distinzione fra consapevolezza (come insieme degli enunciati disponibili per il ragionamento) e coscienza (come insieme degli enunciati disponibili per l'osservazione), la prospettiva progettuale-ingegneristica del suo approccio lo induce a fare coincidere le due nozioni. Se le capacità introspettive sono la condizione necessaria per svolgere attività intellettuali di alto livello, sempre la prospettiva ingegneristica lo induce ad affermare che un robot cosciente può essere progettato in modo da includere anche capacità introspettive non presenti negli esseri umani.

Come abbiamo visto la distinzione fra coscienza e autocoscienza fa emergere certe caratteristiche importanti del dibattito sulla coscienza e permette di individuare alcuni tratti salienti all'interno dell'approccio che focalizza l'attenzione sulla coscienza come fenomeno funzionale. Il carattere soggettivo dell'esperienza, che può essere interpretato in vari modi e dare luogo a una varietà di approcci, è riconosciuto come una delle difficoltà maggiori qualora si voglia dare una spiegazione esauriente della coscienza. Queste difficoltà aumentano quando la spiegazione ambisce a rendere conto della coscienza nella sua totalità; tuttavia, ciò non impedisce che anche tra chi propone modelli computazionali della coscienza ci sia chi riconosca l'importanza di fare rientrare nel modello anche l'aspetto soggettivo. Nella realtà dei fatti, non si riscontra quindi una netta scissione fra chi tratta la coscienza come esperienza soggettiva e chi invece la ritiene un fenomeno puramente funzionale; piuttosto fra chi propone modelli computazionali appare evidente il tentativo di fare i conti anche con l'aspetto soggettivo, o trasformandolo in un fenomeno oggettivo reale oppure dissolvendolo in quanto pseudoproblema¹.

Se l'opposizione fra esperienza e funzione è stata prevalentemente messa in luce riflettendo sulle differenze fra consapevolezza e coscienza, le caratteristiche utilizzate per distinguere fra coscienza e autocoscienza permettono di comprendere alcune sfumature interne all'approccio funzionale e, in particolare, a quello che abbiamo chiamato approccio progettuale-ingegneristico. In questo

¹ Nel dibattito relativo viene usato il termine *qualia* per riferirsi a tutto quanto può essere accessibile da un punto di vista introspettivo, ossia agli aspetti fenomenici delle nostre vite mentali. Sfortunatamente il termine è spesso mal definito e utilizzato in modo inappropriato; qui lo riportiamo in quanto si tratta di una parola chiave della discussione relativa alla coscienza come fenomeno soggettivo.

caso è possibile distinguere due diversi atteggiamenti: il primo si concentra sul tentativo primario di costruire una teoria della coscienza umana che, proprio in quanto espressa in termini computazionali, possa dar luogo anche a una riproduzione artificiale; il secondo, invece, mette in primo piano l'obiettivo di proporre una nozione di coscienza adatta a un sistema artificiale, ritenendo peraltro che questa enfasi progettuale possa offrire una nuova prospettiva anche al problema di comprendere cosa sia e come funzioni la coscienza umana. Si tratta, in ultima analisi, di una differente gerarchia di priorità, che può tuttavia portare a esiti teorici piuttosto distanti.

Avendo a questo punto delineato in generale lo scenario della discussione, è opportuno ora soffermarci sui concetti di modello e di computazione, per vedere come possa essere intesa la nozione di modello computazionale della coscienza.

3. Il concetto di modello

Nell'accezione in cui usiamo il termine, un *modello* è una costruzione astratta dotata di proprietà che stanno in relazione uno a uno con determinate caratteristiche dell'oggetto modellato. Per dare concretezza alle nostre considerazioni ci baseremo su un semplice esempio.

3.1 Un modello di clessidra

Supponiamo di voler definire un modello di clessidra. Il primo passo consiste nella selezione degli aspetti che consideriamo rilevanti: se ciò che vogliamo modellare è la capacità di misurare il tempo, possiamo trascurare il materiale di cui la clessidra è fatta, il suo peso e così via, per concentrarci soltanto sugli aspetti che rendono possibile la misurazione del tempo. Analizzando il funzionamento del nostro oggetto di studio possiamo giungere alle seguenti conclusioni:

- Una *clessidra* è un contenitore costituito da due *camere* collegate fra loro, che chiameremo A e B . Ad ogni istante di tempo, n_A oggetti fisici, che chiameremo *grani*, sono contenuti nella camera A ed n_B oggetti dello stesso tipo sono contenuti nella camera B . La somma $n = n_A + n_B$ è una proprietà invariante nel tempo e caratteristica di ogni singola clessidra (in altre parole, è un *parametro* della clessidra).
- Ad ogni istante di tempo, una clessidra può assumere due *posizioni*, che chiameremo rispettivamente $A|B$, ovvero A su B , e $B|A$, ovvero B su A . In ogni momento si può far passare la clessidra da una posizione all'altra eseguendo un'azione dall'esterno (*girare* la clessidra); assumeremo, per semplicità, che il tempo necessario per girare la clessidra sia nullo.
- Quando la clessidra si trova nella posizione $A|B$, se la camera A è vuota non succede nulla; se invece A non è vuota, il valore di n_A diminuisce di uno ogni volta che trascorrono τ secondi (anche τ è un parametro della clessidra). Lo stesso vale, *mutatis mutandis*, quando la clessidra si trova nella posizione $B|A$.

Sperando di non aver abusato della pazienza del lettore, vediamo ora in che modo questo modello dia conto del fatto che la clessidra è utilizzata per misurare intervalli di tempo. Supponiamo che a un istante di tempo generico, diciamo $t = 0$, la clessidra si trovi nello stato $A|B$, con $n_A(0)$ grani nella camera A e $n_B(0) = n - n_A(0)$ grani nella camera B . Se non si gira la clessidra, a un istante di tempo t_1 (pari al prodotto di $n_A(0)$ per τ) il valore di n_A diviene zero, il valore di n_B diviene n , e questo stato viene poi mantenuto indefinitamente. La clessidra è ora pronta a misurare un intervallo: basterà girarla e attendere che la camera B si svuoti. Non è

difficile vedere che questo processo richiede un numero costante di secondi, pari al prodotto di n per τ : questo è appunto l'intervallo che la clessidra può misurare.

La costruzione matematica che abbiamo presentato è un possibile *modello di clessidra*. Va sottolineato che non abbiamo modellato una clessidra specifica, bensì la classe di tutte le clessidre; d'altra parte possiamo ottenere il modello di una clessidra specifica fissando i valori di n e di τ : è proprio quest'aspetto di generalità e di specificabilità che rende i modelli così utili nelle discipline scientifiche e tecnologiche.

3.2 Modelli descrittivi e modelli esplicativi

Il modello che abbiamo definito descrive, ma non spiega, il comportamento di una clessidra. Ad esempio, ci siamo limitati ad assumere che esattamente un grano passi dalla camera superiore alla camera inferiore ogni τ secondi: ma *perché* ciò debba avvenire all'interno di quei dispositivi di vetro e sabbia che ben conosciamo non l'abbiamo certo spiegato.

Per molte applicazioni un modello descrittivo è più che sufficiente; ad esempio, se un costruttore conosce il parametro τ del suo prodotto, può fornire su ordinazione una clessidra che misuri un intervallo desiderato, agendo sul numero n di grani che vengono introdotti all'atto della produzione. Tuttavia, i modelli possono giocare un ruolo importante non solo nella descrizione di un fenomeno, ma anche nella sua spiegazione. È ragionevole supporre che il comportamento di una clessidra possa essere spiegato dalla fisica, e più specificamente dalla meccanica: nel comportamento dei grani entrano in gioco la forza di gravità, la rigidità dei corpi fisici, l'attrito. Il modello che abbiamo descritto è però così rozzo da non consentire un'applicazione interessante delle leggi della meccanica. Dovremmo quindi definire un modello più raffinato, che tenga conto della forma dei grani e delle pareti del contenitore: potremmo ad esempio assumere che i grani siano modellabili come sfere di raggio determinato, e così via; a questo punto potremmo dimostrare come conseguenza delle leggi della meccanica, e non più assumere per ipotesi, che un grano impiega in media τ secondi per passare dalla camera superiore alla camera inferiore².

Sulla base di queste osservazioni, ci sembra che la distinzione fra *modelli descrittivi* e *modelli esplicativi* possa essere vista nei termini seguenti: un modello descrittivo è una costruzione astratta che, partendo da un certo numero di assunzioni, rappresenta certi aspetti del comportamento di una classe di oggetti; un modello esplicativo è invece una costruzione astratta che, sempre partendo da alcune assunzioni, consente di dimostrare, sulla base di una teoria precostituita, che una classe di oggetti si comporta in un certo modo. Naturalmente, il potere esplicativo non è intrinseco al modello, ma deriva direttamente dalla teoria cui si ricorre per le dimostrazioni; il modello, tuttavia, gioca un ruolo importante e ineliminabile, perché costituisce per così dire il ponte o l'interfaccia fra la teoria e la realtà: non possiamo applicare direttamente la meccanica ai grani di sabbia, per il semplice fatto che la meccanica parla non di grani di sabbia, ma di entità astratte quali punti materiali, solidi geometrici e così via. Per applicare la meccanica a un oggetto reale dobbiamo prima "vedere" i grani come sfere: dobbiamo appunto farci un modello dell'oggetto reale. Possiamo concludere che l'essere descrittivo o esplicativo, per un modello, non è in ultima

² Il nostro modello di clessidra non si basa su alcuna proprietà *fisica* delle entità coinvolte, bensì soltanto su proprietà *ontologiche*. Più precisamente, nel modello sono coinvolte un'ontologia degli oggetti individuali, nonché un'ontologia dello spazio sufficiente a definire i concetti di camera e di passaggio da una camera all'altra. Ad esempio, la conservazione del numero totale dei grani è coerente con l'assunzione che i singoli grani siano individui che mantengono la loro identità quando passano da una camera all'altra.

analisi una proprietà intrinseca, ma piuttosto una relazione con il mondo delle teorie note e accettate: un fatto importante, questo, da tenere presente nel resto della trattazione.

3.3 Adeguatezza di un modello

Stiamo per lanciare un sasso, con una certa angolatura e una certa velocità iniziale, e vogliamo sapere a che distanza andrà a cadere. Forti di buoni ricordi scolastici, sappiamo che la risposta ci può venire da semplici principi di meccanica galileiana, che comportano che la traiettoria di un grave soggetto alla gravità sia una parabola. Fatti i nostri calcoli lanciamo il sasso e misuriamo la distanza raggiunta – e probabilmente finiamo con il concludere che la meccanica galileiana è falsa, o più umilmente (ma forse *troppo* umilmente) che abbiamo sbagliato i calcoli. In realtà, abbiamo senz'altro scoperto una discrepanza fra le nostre previsioni e la realtà: ma c'è un terzo luogo, oltre alla teoria e ai calcoli, dove l'errore si può annidare, e si tratta del modello. Il lettore perdoni la ripetizione se sottolineiamo ancora una volta che le teorie non si applicano direttamente alla realtà: la meccanica non parla di sassi ma di entità astratte, ad esempio di *punti materiali* (punti della geometria euclidea dotati di massa). Presumibilmente, per applicare la meccanica al lancio del sasso abbiamo modellato un sasso come un punto materiale e abbiamo poi condotto i calcoli di conseguenza. Ma un sasso è un corpo esteso e come tale, al contrario di un punto, nell'atmosfera terrestre subisce la resistenza dell'aria; conseguentemente la traiettoria del sasso si discosta da una parabola.

La migliore delle teorie non può che fallire se viene applicata alla realtà con la mediazione di un modello inadeguato. Ma, a sua volta, l'adeguatezza di un modello non può che essere relativa a uno *scopo*: se ci è sufficiente una previsione piuttosto approssimativa, il modello di un grave come punto materiale può essere adeguato; non lo è invece quando occorre una precisione elevata, come già sapevano gli artiglieri di un tempo, ben prima che le bombe diventassero “intelligenti”.

4. Il concetto di computabilità

Per comprendere che cosa sia un modello computazionale della coscienza è ora necessario chiarire il significato del termine *computazionale*. La questione è resa particolarmente delicata dal fatto che tre termini interconnessi, ma non equivalenti, sono spesso utilizzati in modo poco accurato: ci riferiamo agli attributi *essere computabile*, *essere computazionale* ed *essere un computer*, che vogliamo ora analizzare e mettere a confronto.

4.1 Essere computabile

Nella teoria della computabilità, l'attributo *computabile* si applica a entità astratte ben precise, ovvero a *funzioni* (nel senso matematico di corrispondenza univoca fra gli elementi di un insieme e gli elementi di un altro insieme). Più precisamente, una funzione f si dice computabile quando esiste un procedimento di calcolo puramente meccanico (un *algoritmo*) che consente di determinare il valore $f(x)$ in relazione a qualunque argomento x che faccia parte del campo di definizione di f . Ovvii casi di funzioni computabili sono ad esempio la somma e la sottrazione fra numeri interi³: i relativi algoritmi di calcolo si imparano alla scuola elementare, e tutti sanno che per sommare due numeri interi si devono eseguire certe operazioni in modo meccanico, senza esercitare né intelligenza né creatività.

³ Somma e sottrazione sono funzioni a due argomenti, ma non è difficile estendere la definizione di funzione computabile alle funzioni con più di un argomento.

La definizione precedente presuppone che sia ben definito il significato dell'espressione *procedimento di calcolo meccanico*. È noto che i matematici hanno convissuto con un'idea puramente intuitiva di questo concetto fino a quando Church e Turing ne hanno proposto indipendentemente una definizione rigorosa (Church, 1936a,b; Turing, 1936). Nei lavori citati i due autori forniscono una risposta (la stessa, s'intende) al problema della decisione per la logica dei predicati, il celebre *Entscheidungsproblem* posto esplicitamente da Hilbert pochi anni prima: si trattava di determinare se esiste un procedimento meccanico che fosse in grado di stabilire, per ogni enunciato della logica dei predicati, se tale enunciato è o non è dimostrabile. La risposta a questo problema, naturalmente, non era nota quando il problema della decisione venne formulato: è vero che nessun procedimento del genere era stato ancora inventato, ma molti matematici, incluso presumibilmente Hilbert, erano fiduciosi nei progressi futuri della logica matematica.

Per risolvere positivamente il problema della decisione sarebbe stato sufficiente esibire un procedimento che svolgesse il compito richiesto in modo puramente meccanico. Ma come si sarebbe potuto, invece, risolvere lo stesso problema negativamente? Come dimostrare, una volta per tutte, che un procedimento del genere, non ancora inventato, non potrà essere inventato neppure in futuro? Una dimostrazione di questo tipo è possibile solo se il concetto di procedimento meccanico viene delimitato in modo preciso e rigoroso; questo appunto fecero Church e Turing, escogitando tecniche diverse (note come il *lambda-calcolo* di Church e la *macchina di Turing*) ma in definitiva equivalenti. Sia la tecnica di Church sia la tecnica di Turing consentono di dare risposta negativa alla domanda posta da Hilbert: non esiste, né potrà essere inventato in futuro, un procedimento meccanico in grado di decidere se un enunciato arbitrario della logica dei predicati è o non è dimostrabile.

Church e Turing hanno potuto raggiungere il loro scopo definendo un modello dei procedimenti di calcolo meccanici. Come per dimostrare una proprietà delle clessidre non ci si può basare sulle clessidre concrete ma si deve utilizzare un modello di clessidra, allo stesso modo non è possibile dimostrare una proprietà dei procedimenti di calcolo meccanici basandosi sui procedimenti concreti: solo un modello di tali procedimenti consente di raggiungere l'obiettivo. Il lambda-calcolo e la macchina di Turing sono dunque modelli dei procedimenti di calcolo meccanici e, come ogni modello, si basano su assunzioni che, per quanto ragionevoli, non sono prive di un certo grado di arbitrarietà. Non è quindi insensato chiedersi se un modello diverso potrebbe portare, ad esempio, a conclusioni differenti per quanto concerne il problema della decisione. Oggi la stragrande maggioranza dei matematici e degli informatici ritiene che i modelli di Church e di Turing siano perfettamente adeguati, in ciò confortata dal fatto che ogni modello di procedimento meccanico definito fino ad oggi è risultato equivalente al lambda-calcolo e alla macchina di Turing. Che la questione però non sia del tutto peregrina è mostrato, ad esempio, dal filosofo Jack Copeland, che in diversi lavori analizza i "miti" sorti da cattive letture del lavoro di Turing (di gran lunga più citato di Church, forse per l'aspetto più concreto del suo modello). Dato che non ci è possibile qui approfondire questa tematica, rimandiamo il lettore interessato ai lavori di Copeland, molti dei quali scaricabili dal sito web dell'autore (vedi in particolare 2000).

Ritorniamo ora alle nostre domande iniziali. Abbiamo chiarito che l'asserzione X è *computabile* sottostà alle seguenti condizioni:

- perché l'asserzione sia sensata, X dev'essere una funzione (in senso matematico);
- l'asserzione è vera se il valore della funzione, corrispondente a un argomento x qualsiasi del suo campo di definizione, può essere calcolato con un procedimento meccanico.

Ma se questo è il significato del termine *computabile*, che cosa vuol dire allora *computazionale*?

4.2 Essere computazionale

Per comprendere il significato del termine *computazionale* vale la pena di chiedersi in che contesti questo termine sia effettivamente utilizzato dalla comunità scientifica. L'uso più comune, ci sembra, è come attributo di una disciplina: esiste, ad esempio, una disciplina denominata *geometria computazionale*, che studia i problemi geometrici risolubili con procedimenti meccanici (nel senso chiarito sopra). Altrettanto nota è la *linguistica computazionale*, che ha l'obiettivo di sviluppare procedimenti meccanici per l'analisi e la generazione di testi in una lingua naturale. Analogamente, parte della psicologia cognitiva contemporanea potrebbe essere chiamata *psicologia computazionale*, dato che si pone l'obiettivo di definire e convalidare modelli computazionali di processi mentali.

Come è naturale, i modelli definiti nell'ambito di queste discipline sono denominati *modelli computazionali*. A questo punto non è difficile rispondere alla domanda che ci siamo posti all'inizio del paragrafo:

- l'asserzione X è *computazionale* ha senso, in particolare, quando X è un modello, e in tal caso significa che il modello si basa su procedimenti di calcolo meccanici;
- per estensione, si denomina *computazionale* una disciplina il cui obiettivo è sviluppare modelli computazionali.

Come si può verificare facilmente, il modello di clessidra che abbiamo definito nel paragrafo 3.1 è computazionale: basta notare che coinvolge soltanto operazioni aritmetiche eseguibili con una semplice calcolatrice, e quindi ovviamente meccaniche⁴. Volendo, la definizione del modello può essere vista come un semplice esercizio di *fisica computazionale*.

Forti di queste definizioni, sembrerebbe a questo punto possibile affrontare il punto nodale di questo lavoro: che cos'è un modello computazionale della coscienza? Ovviamente si dovrebbe trattare di un modello di qualche aspetto del pensiero cosciente, definito (il modello) in termini puramente meccanici. Le cose però non sono così semplici, come cercheremo di mostrare nel prossimo paragrafo.

4.3 Essere un computer

Il fatto che un sistema o un processo ammettano un modello computazionale non significa naturalmente che quell'entità o quel processo *sia un computer*. Ad esempio, il nostro modello della clessidra è computazionale, ma ciò non significa che una clessidra di vetro e sabbia sia un computer (e in effetti non lo è). In altre parole, l'essere computazionale è una proprietà del modello, non dell'oggetto modellato – e confondere le due cose significa commettere la più subdola e pericolosa delle fallacie, ovvero confondere la mappa con il territorio. Tutt'al più, l'esistenza di un modello computazionale della clessidra ci autorizza a dire che la clessidra si è rivelata, dopo tutto, essere un *sistema computabile*, ovvero un sistema che si può modellare adeguatamente in termini di funzioni computabili.

Che cosa succederebbe, allora, se una situazione analoga si verificasse nello studio della cognizione umana? Ovvero, se i processi cognitivi si lasciassero modellare adeguatamente in

⁴ Per la verità, il modello di clessidra è computazionale solo se il valore di τ può essere rappresentato in modo "finito": per questo è sufficiente assumere che τ sia un numero positivo intero o al più razionale.

modo computazionale? Anche qui potremmo dire che i processi cognitivi sono computabili. Come William Rapaport correttamente sottolinea (1998, p. 403), “computationalism is—or ought to be—the thesis that cognition is computable.” Molti scienziati cognitivi, tuttavia, sembrano sottoscrivere un’affermazione molto più forte; ma prima di affrontare questo punto dobbiamo premettere alcune osservazioni sull’importante distinzione fra *computer* e *programma*.

Nel paragrafo 4.1 abbiamo detto che una funzione f è computabile quando il valore $f(x)$ può essere calcolato, per un x arbitrario, seguendo un procedimento meccanico. Ora, per calcolare il valore $f(x)$ sono necessarie due cose:

- una descrizione del procedimento da seguire, e
- un agente (un essere umano o un dispositivo automatico) in grado di seguire tale descrizione e di eseguire effettivamente i calcoli prescritti.

Nella terminologia informatica, la descrizione del procedimento da seguire è detta *programma*, mentre il termine *computer* è riservato all’agente che esegue il programma. Ritorniamo ora alla posizione che vogliamo analizzare. Come dicevamo, molti scienziati cognitivi (come ad esempio Pylyshyn, 1984) non si limitano a ipotizzare che i processi mentali siano computabili, ma affermano piuttosto che la mente è ciò che è perché *il cervello è un computer e la mente è il suo programma*. Ma che cosa significa dire che il cervello è un computer?

Nei discorsi quotidiani considereremmo vera l’affermazione x è un *computer* se x è un Apple Powerbook, e falsa se è una teiera o una pianta di gerani. Ma che cosa rende un Apple Powerbook un computer, che manchi alla teiera e al geranio? Come abbiamo visto, per dimostrare che un determinato sistema è un computer non è sufficiente dimostrare che è computabile (ovvero descrivibile adeguatamente con un modello computazionale). Se così fosse, nel terzo paragrafo avremmo dimostrato che una clessidra è un computer; peggio ancora, sarebbe difficile trovare un sistema fisico che non lo sia: l’asserzione che qualcosa è un computer diventerebbe così praticamente vuota e irrilevante nel discorso scientifico.

In effetti diversi autori, fra cui in particolare John Searle (1990, 1992), ritengono vuota di significato l’affermazione che un sistema fisico naturale, quale ad esempio il cervello, sia un computer. Searle considera l’affermazione che qualcosa sia un computer come un’attribuzione di funzione, analoga a ciò che facciamo quando diciamo che qualcosa è un coltello o uno sgabello: un coltello è un arnese fatto per tagliare, uno sgabello è un oggetto fatto per sedersi e un computer è un apparecchio fatto per svolgere calcoli automaticamente. Le attribuzioni di funzione, però, sono sempre relative a un osservatore: la struttura oggettiva di un oggetto non è mai sufficiente, di per sé, a determinarne la funzione; ad esempio, potremmo decidere di utilizzare un tronco tagliato come sgabello, ma non avrebbe senso dire che un tronco tagliato è, di per sé e indipendentemente dal nostro punto di vista, uno sgabello. Da ciò deriva che non ci può essere nulla nella struttura fisica di un dispositivo che lo renda, oggettivamente e indipendentemente da un osservatore, un computer; ma se è così, affermare che il cervello è un computer è privo di contenuto: da questa affermazione non può derivare nessuna conseguenza significativa su come il cervello effettivamente opera e sulle ragioni per cui può sostenere processi mentali, perché tali ragioni – se esistono – devono essere trovate nella struttura intrinseca del cervello e non nell’attribuzione di una funzione da parte di un osservatore.

È importante notare che dall’argomentazione di Searle non segue l’impossibilità di un modello computazionale della mente: è ancora possibile che tutti i processi mentali siano descrivibili come

funzioni computabili⁵. Ciò che viene negato è che abbia senso affermare che il cervello sia un computer, e che i processi mentali siano l'esecuzione di un programma da parte del cervello.

In un articolo interessante quanto complesso, Jack Copeland (1996) difende invece un punto di vista diverso: secondo Copeland, infatti, si possono identificare certe proprietà strutturali oggettive che fanno sì che un sistema sia un computer. Essere un computer è quindi una proprietà indipendente dall'osservatore; ciò comporta, fra l'altro, che l'asserzione che il cervello è un computer sia sensata, e che la sua verità o falsità sia una questione empirica, una domanda come tante altre, cui rispondere nell'ambito della normale ricerca scientifica. In sostanza, ciò che Copeland afferma è che l'asserzione *X è un computer* non significa che *X* può essere descritto in termini di funzioni computabili (possono essere descritti in questo modo moltissimi sistemi che non sono computer, come la nostra clessidra), ma piuttosto che *X* può essere adeguatamente modellato come un dispositivo che esegue un programma: la principale difficoltà sta nel precisare che cosa significhi in questo caso *adeguatamente*. L'idea di Copeland è che un sistema, naturale o artificiale, può essere considerato un computer se, e solo se, è possibile spiegarne la dinamica identificando nella struttura del sistema due componenti interconnesse, interpretabili rispettivamente come la rappresentazione di un programma e un esecutore di programmi. A un primo esame, la trattazione di Copeland sembra fornire una definizione esauriente di che cosa significhi essere un computer. Diciamo "sembra" perché è fin troppo frequente, in questo campo, che un'argomentazione apparentemente impeccabile venga confutata poco tempo dopo la sua formulazione; a tutt'oggi, tuttavia, non ci risulta che nulla di simile sia avvenuto.

Se Copeland è nel giusto, l'affermazione che il cervello è un computer non è priva di senso. Ciò non significa però che sia vera, né che sia plausibile sulla base di ciò che oggi si sa dei processi cerebrali; a nostro avviso, anzi, nulla di ciò che è stato scoperto finora sul cervello sembra puntare in questa direzione.

5. Modelli computazionali della coscienza

Nel paragrafo precedente per spiegare il concetto di computabilità abbiamo distinto tre concetti: l'essere computabile, l'essere computazionale e l'essere un computer. Un sistema è computabile se ammette un modello computazionale e un modello è computazionale se può "girare" su un computer. Un computer è necessariamente un sistema computazionale, ma in generale non vale l'inverso: un sistema può essere computazionale senza per questo essere un computer. Per comprendere ora cosa sia un modello computazionale della coscienza dobbiamo considerare questi concetti in relazione alla nozione di coscienza.

5.1 Due diverse interpretazioni

Secondo la definizione proposta nel paragrafo 4.2 del termine computazionale, possiamo dire che un modello della coscienza (o di qualche aspetto del pensiero cosciente) è computazionale se può essere definito in termini puramente meccanici; in altre parole, l'uso di un siffatto modello della coscienza richiede soltanto procedimenti di calcolo meccanici per dare conto del fenomeno della coscienza. Se la definizione di un modello computazionale non desta alcun problema nel caso di

⁵ Va sottolineato che a tutt'oggi non sappiamo se i processi mentali possono essere descritti come processi di calcolo. Alcuni autori sembrano ritenere che la computabilità della mente sia implicata dai lavori di Church e Turing, ma che questa posizione sia insostenibile è mostrato, ad esempio, da Copeland (1997).

una clessidra, come abbiamo illustrato in precedenza, nel caso invece della coscienza appare problematica a causa dell'ancora scarsa conoscenza della sua struttura e dei suoi processi.

L'idea di utilizzare modelli computazionali per comprendere e spiegare la coscienza e il suo funzionamento ha origine nel *computazionalismo*, corrente trasversale alle scienze cognitive, alla filosofia, all'intelligenza artificiale e alla psicologia, che afferma in generale che i processi cognitivi sono computabili⁶. Il computazionalismo (Haugeland, 1981; Johnson-Laird, 1983; Pylyshyn, 1984, per citare alcune delle posizioni più significative) può essere declinato in molti modi diversi e con esiti più o meno forti: dall'idea che i processi cognitivi umani siano computabili e quindi siano delle funzioni che possono essere computate da programmi, all'affermazione che gli esseri umani *in toto* siano dei computer. Non per tutti gli autori in questo settore è però scontato il passaggio dello scenario computazionale dal caso dei processi cognitivi al caso della coscienza: che la coscienza sia computabile non è infatti una conseguenza necessaria del fatto che i processi cognitivi siano computabili, anche se in alcuni casi (Llyod, 1995) viene esplicitamente affermato che quando i modelli computazionali diventano descrizioni adeguate dei processi cognitivi umani, essi sono parimenti in grado di modellare la coscienza, ammettendo con ciò che i due processi siano indissolubilmente connessi.

Naturalmente, parlare di modelli computazionali della coscienza non è come parlare di modelli computazionali di una clessidra o, per cambiare scala, del sistema solare: se, come abbiamo visto, la stessa definizione di che cosa sia la coscienza rappresenta ancora uno dei grandi interrogativi da risolvere, diventa assai difficile valutare l'adeguatezza dei modelli che vengono proposti per spiegare i processi coscienti. In cerca di una buona base su cui poggiare, e in seguito all'esplosione dell'interesse nelle neuroscienze, molti degli autori attualmente impegnati in questo settore tentano di aggirare il problema attraverso la ricerca dei corrispettivi neurali della coscienza (Chalmers, 2000); in altre parole, si individua nel cervello una catena di eventi che sia plausibile ritenere responsabile del fenomeno della coscienza e, solo dopo questo passaggio, si cerca di modellare questa catena in termini computazionali. Ovviamente questo approccio non risolve una volta per tutte il problema di capire se i modelli siano del tutto adeguati, ma fornisce una base scientifica all'analisi, evitando in parte gli estremismi di alcune speculazioni del passato.

Ritorniamo ora ai modelli computazionali della coscienza e cerchiamo di capire come la definizione generale che abbiamo data possa essere declinata secondo due diverse interpretazioni, ricordando che l'essere computazionale è una proprietà del modello, mentre l'essere computabile e l'essere un computer sono proprietà dell'oggetto modellato. Le due interpretazioni derivano dalle diverse accezioni della relazione fra coscienza e computabilità e sono compatibili con la nostra definizione di modello computazionale:

- *La coscienza è computabile.* Affermare che la coscienza è computabile significa affermare che c'è un programma – o meglio una collezione di programmi interconnessi – che la computa, ossia computa la funzione o le funzioni che costituiscono il pensiero cosciente. Ne segue che l'attività mentale cosciente può essere simulata perfettamente da una particolare macchina di Turing, che ne costituisce un modello computazionale. Per comodità, chiameremo questa posizione *computazionalismo estrinseco*.

⁶ Qui non vogliamo entrare nel merito se l'esatta definizione del computazionalismo sia "cognition is computable" (Rapaport, 1998) oppure "cognition is computation" (Bringsjord e Zenzen, 1997), ma solo offrire una caratterizzazione generale.

- *La coscienza è un computer.* Per affermare che la coscienza è un computer non è sufficiente dimostrare che la coscienza è computabile. In accordo con quanto discusso nel paragrafo 4.3, per stabilire che la coscienza è un computer o si attribuisce alla coscienza la funzione di essere un computer (ricordando che le attribuzioni di funzione sono sempre relative a un osservatore e non solo alla struttura oggettiva di un oggetto), o si stabilisce che la coscienza ha certe proprietà strutturali oggettive e, in particolare, che può essere adeguatamente modellata come un dispositivo che esegue un programma. In opposizione al caso precedente, chiameremo questa posizione *computazionalismo intrinseco*.

Avendo stabilito che la locuzione *modelli computazionali della coscienza* può essere interpretata in due modi diversi, analizziamo ora più in dettaglio alcune posizioni significative all'interno di ciascuno dei due approcci: naturalmente, l'analisi è condotta dal punto di vista che abbiamo delineato nei paragrafi precedenti, e non è quindi escluso che in alcuni casi possa verificarsi una lieve forzatura delle posizioni considerate; un difetto, questo, che riteniamo ampiamente compensato dal fatto che il nostro schema riesce a chiarire aspetti importanti solitamente trascurati.

Gli autori che appoggiano un computazionalismo estrinseco generalmente pongono l'enfasi sull'importanza degli strumenti computazionali per comprendere e modellare la coscienza umana. Ciò dà luogo a diverse soluzioni: si va dall'idea che i sistemi connessionistici rappresentino i modelli più adeguati per la comprensione della coscienza, in quanto sono in grado di spiegare aspetti salienti della nostra esperienza (Lloyd, 1989); alla discussione dettagliata di quali siano i possibili contenuti della coscienza, individuati negli stati di una rete di moduli computazionali temporalmente persistenti (Mathis e Mozer 1995); fino a modellare l'attività cosciente come insieme di concetti e capacità molto differenti tra loro, che possono essere presenti o assenti in diverse combinazioni (Sloman, 2002).

Chi invece sostiene una posizione di computazionalismo intrinseco assume che la coscienza sia una sorta di proprietà emergente del cervello, e che tale proprietà possa emergere proprio perché il cervello è un computer. Anche in questo caso è possibile distinguere fra diverse posizioni: alcune affermano che i candidati artificiali più adatti per l'emergenza della coscienza siano le reti neurali (Brockmeier, 1997); altre che un approccio evuzionistico alla robotica è ciò che permette di rendere conto della coscienza come computer (Harvey, 2002); altre ancora che la coscienza è un sistema di elaborazione di alto livello che coordina i processi di livello più basso (Johnson-Laird, 1996).

È significativo però notare che in entrambi i casi, e quindi indipendentemente da come la coscienza sia interpretata in quanto oggetto di analisi (se come processo computabile o come computer), i modelli computazionali verso i quali oggi si propende sono di tipo connessionistico, proprio a sottolineare che solo modelli di questo tipo sono in grado di rendere conto della complessità dei processi che sottostanno al fenomeno della coscienza.

5.2 Modelli descrittivo-esplicativi e modelli di ispirazione

Il panorama che abbiamo delineato secondo le due diverse interpretazioni non può dirsi completo senza considerare un'altra prospettiva significativa. Quando si parla di modelli computazionali della coscienza non ci si riferisce solo a modelli che descrivono o spiegano la coscienza umana; ci si può riferire infatti anche al caso in cui questi modelli costituiscono un motivo di ispirazione per lo sviluppo di architetture informatiche innovative. Così come la struttura neurale del cervello umano è stata un motivo di ispirazione nell'ideazione delle reti neurali artificiali, e la struttura del DNA ha ispirato l'invenzione degli algoritmi genetici, anche la

struttura e le caratteristiche della coscienza umana possono costituire un possibile motivo ispiratore a livello progettuale.

È interessante notare che l'adozione di questo punto di vista è compatibile con ambedue le forme di computazionalismo, sia estrinseco sia intrinseco. Ci sono infatti autori, come Aaron Sloman, che pensano che la coscienza sia computabile e propongono un'architettura per una macchina cosciente e altri, come Inman Harvey, che sostengono che la mente sia un computer e analizzano come sia possibile produrre la coscienza in una macchina. In entrambi i casi la prospettiva è, come abbiamo già fatto notare, anche progettuale-ingegneristica, oltre che descrittivo-esplicativa: l'obiettivo quindi non è solo valutare se la coscienza umana venga adeguatamente compresa attraverso un modello computazionale (obiettivo che peraltro viene considerato realizzato in molti casi), ma anche capire quale sia la via più efficace per la progettazione di entità artificiali coscienti. Ciò che viene dato per scontato è che capire come funziona la coscienza umana e come essa si articoli possa costituire un motivo di ispirazione per la progettazione di una forma artificiale di coscienza.

Una delle caratteristiche della coscienza umana (anzi, per essere più precisi, dell'autocoscienza) che colpisce immediatamente per le sue potenzialità progettuali è la capacità del sistema cosciente di modellare riflessivamente ciò che sta avvenendo all'interno di se stesso; a McCarthy (1999), ad esempio, questa capacità introspettiva sembra di grande importanza per lo sviluppo di robot intelligenti. Un altro esempio è dato dall'architettura distribuita del sistema cognitivo umano postulata da Johnson-Laird (1988), con la coscienza che assolve la funzione di sistema centrale, preposto all'integrazione delle informazioni e all'armonizzazione del funzionamento delle componenti periferiche. Caratteristiche del genere forniscono un motivo di ispirazione nella progettazione e nella realizzazione di entità artificiali che esibiscano comportamenti apparentemente coscienti.

6. Discussione

La trattazione che abbiamo proposto è ben lungi dall'aver esaurito l'argomento dei modelli computazionali della coscienza. Riteniamo, tuttavia, di aver fatto chiarezza su alcuni termini del dibattito, lasciando il lettore non con una soluzione definitiva, ma con alcuni concetti chiariti e una serie di spunti per la riflessione. Per concludere la nostra analisi proponiamo ora qualche considerazione personale.

Uno dei nodi centrali emersi – relativamente sia al tema della coscienza in generale sia al tema dei modelli computazionali della coscienza – è la distinzione fra l'aspetto esperienziale e l'aspetto funzionale della coscienza. Come si ricorderà, il primo rende conto della coscienza come fenomeno qualitativo e soggettivo, strettamente connesso all'esperienza dell'aver coscienza di qualcosa, mentre il secondo è la controparte funzionale, e quindi modellabile in termini oggettivi, della coscienza come fenomeno soggettivo. Ora, dire che la coscienza può essere descritta sotto due differenti aspetti è ben diverso dal sostenere che tali aspetti possono essere considerati indipendenti l'uno dall'altro. Alla base di questa separazione vi è la convinzione che l'aspetto funzionale possa essere spiegato secondo le categorie scientifiche tradizionali, rappresentando un punto di partenza per comprendere la coscienza nella sua totalità: partendo dall'aspetto funzionale, l'*easy problem*, si passerebbe poi con una serie di passi all'aspetto esperienziale, la vera essenza della coscienza, connesso in qualche modo al primo e quindi comprensibile a partire da quello. Una distinzione del genere ci sembra implausibile proprio perché postula una separazione

netta e artificiosa tra due aspetti che nella realtà sono indissolubilmente connessi: infatti è ragionevole ritenere che le funzioni della coscienza siano determinate proprio dalle caratteristiche dell'esperienza soggettiva, e siano quindi le funzioni a dover essere spiegate in termini di qualità dell'esperienza, e non viceversa.

Se la separazione degli aspetti funzionali dall'esperienza soggettiva ambisce a far sparire un mistero, ne introduce però un altro di non minore complessità: come possiamo rendere conto delle connessioni fra i due livelli? Diventa così difficile evitare una forma di dualismo fra entità fisiche ed entità mentali, ricadendo nel classico problema mente-corpo, ossia nel problema di connettere il mentale al fisico e viceversa. Sorge poi un'altra difficoltà: se si ammette che l'aspetto funzionale sia separato dall'aspetto esperienziale, e contemporaneamente si considera la coscienza come fenomeno fondamentalmente biologico (in accordo ad esempio con Searle, 1992, 1997), diventa problematico spiegare la coscienza in termini evuzionistici: se infatti le funzioni della coscienza potessero essere realizzate in assenza dei loro correlati soggettivi, quale sarebbe il vantaggio selettivo fornito dall'esperienza?

L'idea della base biologica della coscienza ci pare interessante e in linea con la tendenza attuale della neurobiologia, cui abbiamo accennato nel paragrafo 5 parlando dei correlati neurali della coscienza. Searle sostiene che i processi cerebrali *causano* la coscienza; tuttavia, questo naturalismo biologico non ci fa ricadere nel dualismo: infatti la coscienza causata in questo modo non è una sostanza o un'entità di natura non fisica, bensì una caratteristica di alto livello dell'intero sistema biologico, che deriva dal livello inferiore formato dagli elementi neuronali. Naturalmente, se la derivazione degli stati coscienti dai processi cerebrali è, secondo Searle, un dato di fatto, i dettagli di questo processo di causazione ci sono ancora sconosciuti, ed è probabile che per arrivare a capire esattamente come il cervello causi la coscienza sia necessaria una vera e propria rivoluzione nelle neuroscienze. Ciò che ci appare estremamente interessante in questa proposta è la centralità del carattere soggettivo della coscienza, che viene mantenuta nonostante l'origine biologica. Pur riconoscendo che la soggettività della coscienza la rende unica rispetto agli altri fenomeni biologici, l'idea forte è che essa possa comunque essere spiegata in modo oggettivo e scientifico: è possibile infatti ricercare una spiegazione epistemologicamente oggettiva di una realtà ontologicamente soggettiva.

Una posizione di apertura verso la spiegazione della coscienza in termini biologici, pur nella consapevolezza di essere lontani da una soluzione definitiva, riconosce comunque la validità dell'approccio che abbiamo denominato "modelli computazionali d'ispirazione". In questo caso i modelli computazionali non ambiscono a *spiegare* la coscienza, ma solo a trasferire nei modelli computazionali alcuni aspetti salienti della coscienza: e non c'è dubbio che l'informatica, e in particolare l'intelligenza artificiale, si sia spesso ispirata ai processi mentali coscienti con l'idea di realizzare macchine e programmi con prestazioni sempre più adeguate. Va tenuto presente, tuttavia, che il metro di valutazione di queste realizzazioni è stato ed è, nella maggior parte dei casi, un criterio puramente pragmatico: se qualcosa funziona, e funziona bene, allora quella è una strada da seguire, indipendentemente dal fatto che i processi coscienti, o alcuni di essi, siano più o meno ben compresi dal punto di vista computazionale. Proprio per questo motivo riteniamo che, all'interno di questo approccio, non sia corretto né plausibile asserire in senso letterale che i sistemi così realizzati possiedano una coscienza. Ciò che in definitiva viene fatto è isolare alcuni aspetti della coscienza ritenuti importanti per il comportamento intelligente e modellarli in termini computazionali: che i prodotti di questo sforzo possano essere chiamati coscienti è del tutto arbitrario, dato che abbiamo postulato il carattere soggettivo e unitario della coscienza. Per

questo motivo riteniamo non abbia senso parlare di coscienza quando in realtà ci si ispira solo ad alcuni aspetti dei processi mentali coscienti ritenuti comprensibili in termini funzionali, siano questi la consapevolezza, la capacità di introspezione o l'autocoscienza. Non occorre scomodare Confucio per sapere che i nomi hanno un'importanza essenziale: una terminologia suggestiva ma scorretta può dare luogo a tali ambiguità e fraintendimenti da rendere praticamente incomprensibile anche il più serio dei dibattiti.

Dedica e ringraziamenti

Questo lavoro è dedicato alla memoria di Marco Somalvico, amico e maestro. Desideriamo inoltre ringraziare Claudia Arrighi per i suoi validi suggerimenti.

Riferimenti bibliografici

- Bechtel, W., 1994. Consciousness: Perspectives from symbolic and connectionist AI. *Neuropsychologia*, 33, 1075–1086.
- Bringsjord, S., Zenzen, M., 1997. Cognition is not computation. *Synthese*, 113, 285–320.
- Brockmeier, S., 1997. Computational architecture and the creation of consciousness. *The Dualist, Undergraduate Journal of Philosophy, Stanford University*, 4, <http://www.stanford.edu/group/dualist/vol4/> (ultima visita 5 novembre 2003).
- Chalmers, D., 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 200–219.
- Chalmers, D., 2000. What is a neural correlate of consciousness? In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, a cura di T. Metzinger, MIT Press, Cambridge, MA, 17–40.
- Church, A., 1936a. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58, 345–363.
- Church, A., 1936b. A note on the Entscheidungsproblem. *Journal of Symbolic Logic*, 1, 40–41.
- Copeland, B.J., 1996. What is computation? *Synthese*, 108, 335–359.
- Copeland, B.J., 1997. The broad conception of computation. *American Behavioral Scientist*, 40, 690–716.
- Copeland, B.J., 2000. The Church-Turing thesis. *The Stanford Encyclopedia of Philosophy (Fall 2002 Edition)*, a cura di E.N. Zalta, <http://plato.stanford.edu/archives/fall2002/entries/church-turing> (ultima visita 31 ottobre 2003).
- Harvey, I., 2002. Evolving robot consciousness: The easy problem and the rest. In *Evolving consciousness*, a cura di J. Fetzer, Advances in Consciousness Research Series, John Benjamins, Amsterdam, 205–219.
- Haugeland, J., 1981. *Artificial Intelligence: The very idea*. MIT Press, Cambridge, MA. Trad. it., *Intelligenza artificiale: il significato di un'idea*, Bollati Boringhieri, Torino, 1988.
- Jackendoff, R., 1987. *Consciousness and the computational mind*. MIT Press, Cambridge, MA. Trad. it., *Coscienza e mente computazionale*, Il Mulino, Bologna, 1990.
- Johnson-Laird, P. N., 1983. *Mental models*. Harvard University Press, Cambridge, MA. Trad. it., *Modelli mentali: verso una scienza cognitiva del linguaggio, dell'inferenza e della coscienza*, Il Mulino, Bologna, 1988.
- Johnson-Laird, P. N., 1988. *The computer and the mind: An introduction to Cognitive Science*. Harvard University Press, Cambridge, MA. Trad. it., *La mente e il computer: introduzione alla scienza cognitiva*, Il Mulino, Bologna, 1990.

- Johnson-Laird, P. N., 1996. A computational analysis of consciousness. In *Consciousness in contemporary science*, a cura di E. Bisiach e A. Marcel. Oxford University Press, Oxford, 357–368.
- Lyod, D., 1989. *Simple minds*. MIT Press, Cambridge, MA.
- Lloyd, D., 1995. Consciousness: A connectionist manifesto. *Minds and Machines*, 5, 161–185.
- Mathis, D.W., Mozer, M.C., 1995. On the computational utility of consciousness. In *Advances in Neural Information Processing Systems 7*, a cura di G. Tesauro, D.S. Touretzky, T.K. Leen, MIT Press, Cambridge, MA, 11–18.
- Mathis, D.W., Mozer, M.C., 1996. Conscious and unconscious perception: A computational theory. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, a cura di G. Cottrell, Erlbaum, Hillsdale, NJ, 324–328.
- McCarthy, J., 1999. Making robots conscious of their mental states. In *Machine Intelligence 15*, a cura di K. Furukawa, D. Michie, S. Muggleton, Oxford University Press, Oxford, 3–17.
- Nagel, T., 1974. What is it like to be a bat? *The Philosophical Review*, 83, 435–450.
- Pylyshyn, Z.W., 1984. *Computation and cognition: Towards a foundation for cognitive science*. MIT Press, Cambridge, MA.
- Rapaport, W.J., 1998. How minds can be computational systems. *Journal of Experimental and Theoretical Artificial Intelligence*, 10, 403–419.
- Searle, J.R., 1990. Is the brain's mind a computer program? *Scientific American*, 262, 20–25.
- Searle, J.R., 1992. *The rediscovery of the mind*. MIT Press, Cambridge, MA. Trad. it., *La riscoperta della mente*, Bollati Boringhieri, Torino, 1994.
- Searle, J.R., 1997. *The mystery of consciousness*, New York Review of Books, New York. Trad. it., *Il mistero della coscienza*, Raffaello Cortina, Milano, 1998.
- Sloman, A., 2002. Architecture-based conceptions of mind. In *The scope of Logic, Methodology, and Philosophy of Science (Vol. II)*, a cura di P. Gärdenfors, K. Kijania-Placek, J. Woléński. Kluwer, Synthese Library, 316, 403–427.
- Turing, A.M., 1936. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, Series 2, 42, 230–265.