La mente è un programma?

No. Un programma di calcolatore manipola simboli; il cervello vi annette un significato

di John R. Searle

Può una macchina pensare? Può avere pensieri coscienti esattamente nello stesso senso in cui li abbiamo noi? Se per «macchina» s'intende un sistema fisico capace di compiere certe funzioni (e che altro si potrebbe intendere?), allora gli esseri umani sono particolari macchine di tipo biologico e gli esseri umani sono in grado di pensare: quindi, ovviamente, le macchine sono in grado di pensare. E, per quanto ne sappiamo, sarebbe forse possibile costruire una macchina pensante servendosi di materiali del tutto diversi, per esempio di chip di silicio o di valvole termoioniche. Forse ciò potrebbe rivelarsi impossibile, ma certo non lo sappiamo ancora.

Negli ultimi decenni, tuttavia, il quesito se una macchina possa pensare ha ricevuto un'interpretazione del tutto diversa, anzi è stato sostituito dal quesito seguente: una macchina può pensare semplicemente in virtù del fatto che esegue un programma di calcolatore? Il programma è di per sé una componente del pensiero? Questo è un problema del tutto diverso, perché non riguarda le proprietà fisiche e causali di sistemi fisici attuali o potenziali, ma riguarda invece le proprietà computazionali astratte dei programmi formali di calcolatore che possono essere eseguiti in un qualunque supporto materiale, purché questo supporto sia in grado di svolgere il programma.

Un buon numero di ricercatori che si occupano di intelligenza artificiale (IA) crede che la risposta al secondo quesito sia affermativa; essi credono, cioè, di creare letteralmente delle menti allorquando scrivono i programmi giusti con gli ingressi giusti e le uscite giuste. Credono inoltre di possedere un criterio scientifico per stabilire il successo o il fallimento dell'impresa: il cosiddetto «test di Turing» ideato dal padre fondatore dell'intelligenza artificiale, Alan M. Turing. Il criterio di Turing, secondo l'interpretazione corrente, è semplicemente questo: se un calcolatore riesce a comportarsi in modo tale che un esperto non sia in grado di distinguere il suo comportamento da quello di un essere umano che possegga una certa capacità cognitiva - per esempio la capacità di fare le addizioni o di capire la lingua cinese - allora anche il calcolatore possiede questa capacità. Il fine è quindi quello di scrivere programmi capaci di simulare le capacità cognitive dell'uomo, in modo da superare il test di Turing. E, cosa più importante, un programma del genere non sarebbe solo un modello della mente: sarebbe una vera e propria mente, nello stesso senso in cui lo è la mente dell'uomo.

Certamente non tutti i ricercatori che si occupano di intelligenza artificiale condividono questa posizione estremistica. Un modo più moderato di avvicinarsi al problema consiste nel ritenere che i modelli basati sul calcolatore siano utili per studiare la mente così come sono utili per studiare le condizioni meteorologiche, i processi economici o i meccanismi della biologia molecolare. Per distinguere queste due impostazioni, chiamerò «IA forte» la prima e «IA debole» la seconda. È importante rendersi conto di quanto la posizione espressa dall'IA forte sia audace. Secondo l'IA forte il pensiero non è altro che la manipolazione di simboli formali, e questo è proprio quanto fa il calcolatore: manipola simboli formali. Questa posizione viene spesso riassunta con la frase: «la mente sta al cervello come il programma sta al calcolatore».

L'IA forte si discosta dalle altre teorie della mente sotto almeno due aspetti: può essere espressa in modo chiaro ed è soggetta a una confutazione semplice e decisiva. Questa confutazione può essere applicata da chiunque in prima persona. Eccola. Si consideri una lingua che l'individuo in questione non conosce. Io, per esempio, non conosco il cinese: ai miei occhi la scrittura cinese si presenta come una serie di scarabocchi senza significato. Supponiamo ora che io mi trovi in una stanza contenente scatole piene di ideogrammi cinesi e supponiamo che mi venga fornito un manuale di regole (scritto nella mia lingua) in base alle quali associare ideogrammi cinesi ad altri ideogrammi cinesi. Le regole specificano senza ambiguità gli ideogrammi in base alla loro forma e non richiedono che io li capisca. Le regole potrebbero essere di questo tipo: «Prendi uno scarabocchio dalla prima scatola e mettilo accanto a uno scarabocchio preso dalla seconda scatola».

Supponiamo che fuori dalla stanza vi siano delle persone che capiscono il cinese e che introducano gruppetti di ideogrammi e che, in risposta, io manipoli questi ideogrammi secondo le regole del manuale e restituisca loro altri gruppetti di ideogrammi. Ora il manuale con le regole è il «programma di calcolatore», le persone che l'hanno scritto sono i «programmatori» e io sono il «calcolatore». Le scatole piene di ideogrammi sono la «base di dati», i gruppetti di ideogrammi che mi vengono forniti sono le «domande» e quelli che io restituisco sono le «risposte».

Supponiamo ora che le regole del manuale siano scritte in modo tale che le mie «risposte» alle «domande» non si possano distinguere da quelle di una persona di lingua madre cinese. Per esempio, gli individui situati al di fuori della stanza mi potrebbero passare degli ideogrammi il cui significato, a me sconosciuto, sia: «Qual è il colore che preferisci?» e, seguendo le regole, io potrei restituire loro degli ideogrammi il cui significato, a me del pari sconosciuto, sia: «Il colore che preferisco è l'azzurro, ma mi piace molto anche il verde.» Io supero così il test di Turing per la comprensione del cinese, eppure ignoro completamente questa lingua. E, nel sistema che ho descritto, non potrei in nessun modo giungere a capire il cinese, perché non avrei la possibilità di apprendere il significato di alcun simbolo. Come un calcolatore, io manipolo simboli, ma non annetto a questi simboli alcun significato.

Questo esperimento concettuale dimostra che se io non capisco il cinese per il solo fatto di eseguire un programma per la comprensione del cinese, allora non ci riesce alcun altro calcolatore digitale che si limiti a far girare un programma del genere. I calcolatori digitali si limitano a manipolare simboli formali secondo le regole contenute nel programma.

Ciò che vale per il cinese vale anche per altre attività cognitive. La sola manipolazione dei simboli non basta di per sé a garantire l'intelligenza, la percezione, la comprensione, il pensiero e così via. E poiché i calcolatori sono per loro natura dispositivi per operare sui simboli, la semplice operazione di far girare il programma non è garanzia sufficiente di attività cognitive.

Questo semplice argomento confuta in modo radicale le pretese dell'IA forte. La prima premessa dell'argomentazione asserisce semplicemente il carattere formale di un programma di calcolatore. I programmi sono definiti in termini di manipolazioni di simboli e i simboli sono enti puramente formali, cioè «sintattici». È il carattere formale del programma, per inciso, che rende così potenti i calcolatori: lo stesso programma può essere eseguito su una varietà illimitata di calcolatori e un certo complesso circuitale può far girare una gamma illimitata di programmi di calcolatore. Riassumerò questo «assioma» nel modo seguente:

Assioma 1. *I programmi di calcolatore sono formali (sintattici).*

Questo punto è talmente importante che merita una spiegazione più particolareggiata. Un calcolatore digitale elabora informazione codificandola in primo luogo nel simbolismo che esso usa e poi manipolando i simboli secondo un insieme di regole enunciate con precisione. Queste regole costituiscono il programma. Per esempio nella prima teoria dei calcolatori, enunciata da Turing, i simboli erano semplicemente 0 e 1 e le regole del programma erano del tipo: «Scrivi uno 0 sul nastro, spostalo a sinistra di una casella e cancella un 1.» La cosa straordinaria è che qualsiasi informazione che possa essere espressa in una lingua può essere anche codificata in un sistema del genere e che qualsiasi compito di elaborazione dell'informazione che possa essere risolto con regole esplicite può essere programmato.

Ancora due osservazioni importanti: primo, simboli e programmi sono nozioni puramente astratte; non esistono proprietà fisiche essenziali che li definiscano ed è possibile fornirne una realizzazione concreta con qualunque mezzo fisico. I simboli 0 e 1, per loro natura, non hanno alcuna proprietà fisica essenziale e *a fortiori* non hanno proprietà fisiche causali. Sottolineo questo punto perché è forte la tentazione di identificare i calcolatori con qualche tecnologia particolare - per esempio con la microelettronica del silicio - e pensare che le proprietà e i problemi di cui si parla riguardino la fisica dei chip di silicio, oppure pensare che la sintassi identifichi qualche fenomeno fisico che potrebbe avere poteri causali ancora sconosciuti così come hanno proprietà fisiche causali i fenomeni fisici veri e propri, come la radiazione elettromagnetica o gli atomi di idrogeno. La seconda osservazione è che sui simboli si opera senza riferimento ad alcun significato. I simboli del programma possono rappresentare qualunque cosa il programmatore o l'utente desideri. In questo senso il programma ha una sintassi, ma non ha una semantica.

L'assioma successivo serve solo a ricordarci il fatto ovvio che gli atti di pensiero, di percezione, di comprensione e così via hanno un contenuto mentale. Grazie al loro contenuto possono concernere oggetti e situazioni del mondo esterno. Se il contenuto coinvolge una lingua, accanto alla semantica ci sarà una sintassi, ma la comprensione linguistica richiede almeno un ambiente semantico. Se, per esempio, penso alle ultime elezioni presidenziali, la mia mente sarà attraversata da certe parole, ma esse riguardano le elezioni solo perché a queste parole annetto significati particolari in conformità con la mia conoscenza della lingua in cui mi esprimo. Sotto questo profilo esse sono per me ben diverse dagli ideogrammi cinesi. Riassumerò così questo assioma:

Assioma 2. *La mente umana ha contenuti mentali (una semantica).*

Voglio ora aggiungere quanto è stato dimostrato dall'argomento della stanza cinese. Il possesso dei soli simboli, della sola sintassi, non è sufficiente per possedere la semantica. Le semplici manipolazioni dei simboli non bastano per garantire la conoscenza del loro significato. Riassumerò tutto ciò nel seguente assioma:

Assioma 3. *La sintassi di per sé non è condizione essenziale, né sufficiente, per la determinazione della semantica.*

A un certo livello questo principio è vero per definizione. Naturalmente sarebbe possibile definire in modo diverso i termini sintassi e semantica. Il fatto è che c'è una differenza fra gli elementi formali, che non hanno un significato o un contenuto intrinseco, e i fenomeni che hanno contenuto intrinseco. Da queste premesse segue la

Conclusione 1. *I programmi non sono condizione essenziale né sufficiente perché sia data una mente.*

Questo è solo un modo diverso per affermare che l'IA forte è falsa.

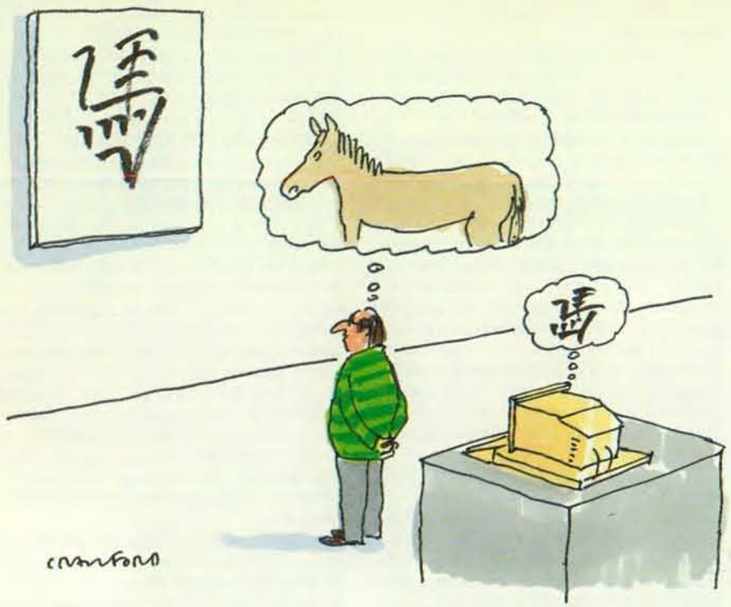
È importante capire che cosa esattamente dimostra e che cosa non dimostra questo ragionamento.

In primo luogo, non ho cercato di dimostrare che «un calcolatore non può pensare». Poiché tutto ciò che può essere simulato per via computazionale può essere descritto come un calcolatore e poiché il nostro cervello può a certi livelli essere simulato, ne segue ovviamente che il nostro cervello è un calcolatore ed esso può certo pensare. Ma dal fatto che un sistema possa essere simulato per mezzo della manipolazione di simboli e dal fatto che esso sia pensante non ne segue che pensare equivalga all'esecuzione di manipolazioni formali su simboli.

In secondo luogo, non ho cercato di dimostrare che solamente i sistemi a base biologica, come il nostro cervello, possono pensare. Per il momento questi sono gli unici sistemi conosciuti che di fatto pensino, ma potremmo scoprirne altri nell'universo capaci di produrre pensieri coscienti e potremmo addirittura riuscire a costruire in futuro sistemi artificiali in grado di pensare. Ritengo che questo problema sia aperto a ogni soluzione.

In terzo luogo, la tesi dell'IA forte non sostiene che, per quanto si sappia, calcolatori con programmi giusti potrebbero pensare e avere proprietà psicologiche finora non osservate; la tesi sostiene che essi vanno considerati pensanti

Supero il test di Turing per la comprensione del cinese



I programmi di calcolatore sono formali (sintattici). La mente umana ha contenuti mentali (ha una semantica)

perché il pensare si riduce a questo.

In quarto luogo, ho cercato di confutare l'IA forte definita in questo modo. Ho cercato di dimostrare che il programma di per sé non è condizione necessaria per il pensiero, perché il programma si limita a eseguire operazioni formali sui simboli e sappiamo per altra via che le manipolazioni sui simboli non sono in sé sufficienti a garantire la presenza di significati. Questo è il principio su cui è basato l'argomento della stanza cinese.

Ho sottolineato questi punti anche perché mi sembra che i Churchland (si veda *Può una macchina pensare?* di Paul M. Churchland e Patricia Smith Churchland a pagina 22) siano convinti che secondo l'IA forte i calcolatori potrebbero dimostrarsi capaci di pensare e che io neghi questa possibilità per motivi di buon senso. Ma non è questo che sostiene l'IA forte, né la mia confutazione ha niente a che vedere con il buon senso.

Riprenderò più avanti il discorso sulle loro obiezioni, ma ora vorrei osservare che, a differenza di ciò che credono i Churchland, l'argomento della stanza cinese confuta anche tutte le ipotesi di tipo IA forte costruite a proposito delle nuove tecnologie parallele ispirate alle reti neuroniche. A differenza del calcolatore tradizionale di von Neumann, che lavora passo passo, questi sistemi hanno molte unità di elaborazione che funzionano in parallelo e interagiscono fra loro secondo regole che si ispirano alla neurobiologia. Per quanto i risultati siano ancora modesti, questi modelli a «elaborazione parallela distribuita», chiamati anche «connessionistici», sollevano interrogativi utili sulla funzione che possono avere certi sistemi a rete complessi e paralleli come quelli del cervello nella produzione di un comportamento intelligente.

La natura parallela, «cerebrale», dell'elaborazione, tuttavia, è irrilevante per gli aspetti puramente computazionali del processo. Se una funzione può essere calcolata su una macchina parallela, può essere calcolata anche su una macchina seriale. Anzi, poiché le macchine parallele sono ancora rare, di solito i programmi connessionistici sono eseguiti su macchine seriali tradizionali. Pertanto l'elaborazione parallela non costituisce una via d'uscita nei confronti dell'argomento della stanza cinese.

Anzi, il sistema connessionistico è soggetto di per sé a una variante dell'obiezione costituita dall'argomentazione originaria della stanza cinese. Immaginiamo di avere, invece di una stanza cinese, una palestra cinese: una sala contenente molti uomini che parlino solo l'inglese. Questi uomini effettuerebbero le stesse operazioni svolte dai nodi e dalle sinapsi dell'architettura connessionistica descritta dai Churchland, e il risultato sarebbe lo stesso che se a manipolare gli ideogrammi secondo un prontuario di regole ci fosse una sola persona. Nella palestra nessuno sa una parola di cinese e il sistema nel suo complesso non ha modo di apprendere il significato di alcuna parola cinese. E tuttavia, con opportune modifiche, il sistema può fornire risposte corrette a domande in cinese.

Ho accennato prima che le reti connessionistiche posseggono interessanti proprietà che consentono loro di simulare i processi cerebrali in modo più preciso delle tradizionali architetture seriali. Ma i vantaggi che l'architettura parallela presenta per l'IA debole sono affatto irrilevanti per dirimere i contrasti fra la prova della stanza cinese e l'IA forte.

In merito a questo punto, i Churchland mancano il bersaglio quando sostengono che una palestra cinese sufficientemente grande potrebbe avere caratteristiche mentali di livello superiore che derivano dalla dimensione e dalla complessità del sistema, proprio come un cervello completo ha caratteristiche mentali che non sono possedute dai singoli neuroni. Può darsi che sia così, ma ciò ha ben poco a che fare con le capacità di calcolo: da questo punto di vista, sistemi seriali e paralleli sono del tutto equivalenti. Infatti, ogni calcolo che possa essere eseguito nei sistemi paralleli può esserlo anche in quelli seriali. Se l'uomo nella stanza cinese è equivalente, dal punto di vista computazionale, a entrambi i tipi di sistemi, allora, se egli non è in grado di capire il cinese in virtù del solo fatto che computa, neppure i sistemi lo sono.

I Churchland sono nel giusto quando affermano che l'argomentazione originale della stanza cinese fu ideata pensando all'intelligenza artificiale tradizionale, ma hanno torto quando pensano che l'architettura connessionistica sia immune da quella confutazione. L'argomento si applica a qualsiasi sistema computazionale. Non è possibile ottenere contenuti di pensiero semanticamente pregnanti tramite semplici computazioni formali, che siano eseguite in serie o in parallelo. Questa è la ragione per cui l'argomento della stanza cinese confuta l'intelligenza artificiale forte in qualunque sua forma.

Molti di coloro su cui questo ragionamento fa presa restano nondimeno perplessi circa le differenze tra persone e calcolatori. Se gli esseri umani sono, almeno in senso banale, calcolatori e se gli esseri umani dispongono di una semantica, perché allora non dovremmo essere in grado di fornire una semantica agli altri calcolatori? Perché non potremmo programmare un Vax o un Cray in modo che anch'essi abbiano pensieri e sentimenti? O, ancora, perché non potrebbe una qualche nuova tecnologia informatica superare il divario tra forma e contenuto, tra sintassi e semantica? Quali sono, insomma, le differenze tra il cervello degli animali e i sistemi informatici per cui la prova della stanza cinese funziona contro i calcolatori ma non contro i cervelli?

La differenza più evidente è che i processi che trasformano un qualcosa in un calcolatore - le procedure di calcolo - sono completamente indipendenti da qualunque riferimento a un tipo specifico di attuazione. In linea di principio si potrebbe costruire un calcolatore legando insieme con filo di ferro vecchie lattine di birra e fornendo energia con un mulino a vento.

Ma quando si considera il cervello, anche se gli scienziati sanno ben poco di come esso produca gli stati mentali, si è colpiti dall'estrema specificità della sua anatomia e fisiologia. Nei casi in cui si sa qualcosa sul modo in cui i processi cerebrali producono i fenomeni mentali - per esempio il dolore, la sete, la vista, l'olfatto - è chiaro che sono in gioco processi neurobiologici specifici. La sete, almeno la sete di un certo tipo, è provocata da certe scariche neuroniche nell'ipotalamo, a loro volta causate dall'azione di un peptide specifico, l'angiotensina II. Il rapporto causa-effetto procede dal basso verso l'alto, nel senso che i fenomeni mentali di livello superiore sono causati da processi neuronici di livello più basso. Anzi, per quanto ne sappiamo, ogni evento «mentale», dalla sensazione di sete al pensiero di un teorema matematico ai ricordi d'infanzia, è causato dalle scariche di specifici neuroni in architetture nervose ben determinate.

Ma perché è importante questa specificità? In fin dei conti le scariche dei neuroni potrebbero essere simulate su calcolatori costruiti con materiali del tutto diversi, per proprietà fisiche e chimiche, da quelli del cervello. La risposta è che il cervello non rappresenta solo un'esemplificazione concreta di una configurazione formale o di un programma (è anche questo, certo), ma *causa* anche eventi mentali grazie a processi neurobiologici specifici. Il cervello è un organo biologico specifico e le sue proprietà biochimiche specifiche gli consentono di causare la coscienza e altri tipi di fenomeni mentali. Le simulazioni al calcolatore dei processi cerebrali forniscono modelli degli aspetti formali di questi processi, ma la simulazione non va confusa con la riproduzione. Il modello computazionale dei processi mentali non è più reale di quello di qualsiasi altro fenomeno naturale.

Si può immaginare una simulazione al calcolatore precisa fino all'ultima sinapsi dell'azione dei peptidi nell'ipotalamo. Ma si può del pari immaginare una simulazione al calcolatore dell'ossidazione degli idrocarburi in un motore d'automobile o dei processi di digestione in uno stomaco alle prese con una pizza. Nel caso del cervello la simulazione non è più reale che nel caso dell'automobile o dello stomaco. A meno che non avvenga un miracolo, non potremmo far marciare la nostra macchina grazie a una simulazione al calcolatore dell'ossidazione della benzina né potremmo digerire la pizza eseguendo il programma che simula tale digestione. Sembra altrettanto ovvio che la simulazione di un processo cognitivo non produca gli stessi effetti della neurobiologia di quel processo cognitivo.

Tutti i fenomeni mentali vengono, dunque, causati da processi neurofisiologici che avvengono nel cervello; da questo deriva:

Assioma 4. *Il cervello causa la mente.*

Insieme con quanto è stato dedotto in precedenza, si ricava subito un'ovvia conseguenza:

Conclusione 2. *Qualunque altro sistema in grado di causare una mente dovrebbe possedere poteri causali (almeno) equivalenti a quelli del cervello.*

È come dire che se un motore elettrico è capace di far marciare un'automobile alla stessa velocità di un motore a benzina, esso deve avere (almeno) una potenza d'uscita equivalente. Questa conclusione non ci dice nulla sui meccanismi. In realtà i processi cognitivi sono un fenomeno biologico: gli stati e i processi mentali sono causati dai processi cerebrali. Ciò non implica che solo un sistema biologico possa pensare, ma implica che ogni sistema alternativo, fatto di silicio, lattine di birra o che altro, dovrebbe possedere capacità causali pertinenti equivalenti a quelle del cervello. Posso così ricavare la:

Conclusione 3. *Qualunque sistema artificiale in grado di produrre fenomeni mentali, ossia qualunque cervello artificiale, dovrebbe essere in grado di riprodurre gli stessi poteri causali specifici del cervello umano e non potrebbe farlo soltanto svolgendo un programma formale.*

Posso inoltre ricavare un'importante conclusione sul cervello umano:

Conclusione 4. *Il modo in cui il cervello umano produce effettivamente i fenomeni mentali non può ridursi solamente allo svolgimento di un programma al calcolatore.*

Pubblicai per la prima volta la storia della stanza cinese nel 1980 sulle pagine di «Behavioral and Brain Sciences», dove comparve, com'è abitudine della rivista, insieme ai commenti di altri esperti del settore, che in questo caso erano ben ventisei. Francamente ritengo che il succo dell'articolo fosse chiaro, ma con mia sorpresa la pubblicazione fu seguita da un'altra valanga di obiezioni che, cosa ancora più sorprendente, continua ancor oggi. La stanza cinese ha, evidentemente, toccato qualche nervo sensibile.

La tesi dell'IA forte è che un sistema arbitrario, sia esso fatto di lattine di birra, di chip di silicio o di carta igienica, purché attui il programma giusto, con gli ingressi giusti e le uscite giuste, non solo può avere pensieri e sentimenti, ma *deve* avere pensieri e sentimenti. Ebbene, questa è un'opinione fortemente antibiologica, e si potrebbe pensare che gli esperti di IA sarebbero lieti di abbandonarla. Molti di loro, specie quelli della generazione più giovane, sono d'accordo con me, ma sono sbalordito dal numero e dalla veemenza di quanti difendono questo punto di vista. Ecco alcune delle obiezioni più comuni:

a. La persona che sta nella stanza cinese in realtà capisce il cinese, anche se non se ne rende conto. Dopo tutto, è possibile capire qualcosa senza sapere che la si capisce.

b. La persona nella stanza non capisce il cinese, ma in essa vi è un sottosistema (inconscio) che lo capisce. In fin dei conti è possibile avere stati mentali inconsci e non c'è motivo perché la comprensione del cinese non debba essere del tutto inconscia.

c. La persona non capisce il cinese, ma la stanza nel suo complesso sì. La persona è proprio come un singolo neurone del cervello e, come un singolo neuro-

Quale semantica esprime ora il sistema?

ne da solo non può capire, ma può contribuire alla comprensione del sistema complessivo, così la persona non capisce, ma il sistema complessivo sì.

d. La semantica non esiste comunque: esiste solo la sintassi. È un'illusione prescientifica supporre che nel cervello vi siano certi misteriosi «contenuti mentali», «processi di pensiero» o una «semantica». Tutto ciò che avviene nel cervello è lo stesso tipo di manipolazione di simboli sintattici che avviene nei calcolatori, e nient'altro.

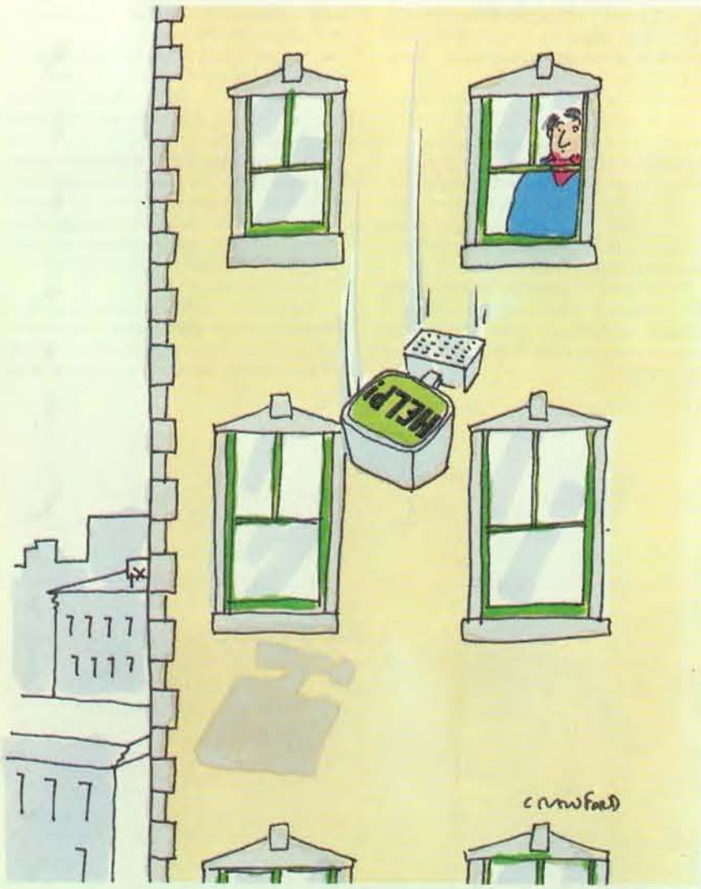
e. La persona nella stanza in realtà non svolge un programma di calcolatore, crede soltanto di svolgerlo. Quando un agente conscio esegue le istruzioni del programma, non si tratta più dell'esecuzione di un programma.

f. I calcolatori avrebbero una semantica e non solo una sintassi se i loro ingressi e le loro uscite fossero posti in un'opportuna relazione causale con il resto del mondo. Immaginiamo di inserire il calcolatore in un robot, di applicare alla testa del robot una telecamera, di installare dei trasduttori per inviare al calcolatore le immagini televisive e di far muovere le braccia e le gambe del robot in base alle uscite del calcolatore. Allora il sistema complessivo avrebbe una semantica.

g. Se il programma simulasse il funzionamento del cervello di una persona capace di parlare cinese, allora capirebbe il cinese. Supponiamo di simulare il cervello di un cinese a livello neuronico. Allora certamente questo sistema capirebbe il cinese bene quanto il cervello di qualunque cinese.

E così via.

Queste obiezioni hanno in comune una caratteristica: sono tutte difettose perché in realtà non affrontano l'argomento della stanza cinese. Questo si basa sulla distinzione fra le manipolazioni formali sui simboli compiute dal calcolatore e il contenuto mentale del cervello, distinzione che ho ricondotto - spero in modo non equivoco - alla distinzione tra sintassi e semantica. Non ripeterò qui le mie risposte a tutte queste obiezioni, ma gioverà alla chiarezza illustrare le debolezze dell'obiezione più frequente, l'argomento c, che io chiamo la risposta del sistema. (Anche l'argomento g, la risposta del simulatore cerebrale, è molto frequente, ma di essa mi sono già occupato nel paragrafo precedente.)



Com'è possibile che qualcuno creda che la simulazione al calcolatore di un processo mentale possa coincidere con il processo mentale?

La risposta del sistema afferma che naturalmente la persona nella stanza non capisce il cinese, ma il sistema complessivo - la persona, la stanza, il prontuario delle regole, gli scatoloni pieni di ideogrammi - lo capisce. Quando per la prima volta venni a conoscenza di questa spiegazione, chiesi a uno dei suoi sostenitori: «Vuol dire che la stanza capisce il cinese?» e lui rispose di sì. È una mossa ardita, ma a parte la sua scarsa plausibilità, non si regge a lume di logica. L'essenza dell'argomentazione originale era che di per sé il rimescolamento dei simboli non dà accesso al loro significato. Ma questo è vero per la stanza nel suo complesso quanto per la persona che vi sta dentro. Questo lo si capisce estendendo l'esperimento concettuale. Immaginiamo che la persona impari a memoria il contenuto delle scatole e del prontuario e faccia tutti i calcoli a mente. Si può anche immaginare che essa lavori all'aperto. Non c'è nulla nel «sistema» che non sia anche nella persona e poiché la persona non capisce il cinese, non lo capisce neanche il sistema.

Nell'articolo che accompagna questo, i Churchland propongono una variante della risposta del sistema immaginando una divertente analogia. Supponiamo che qualcuno affermi che la luce non può avere natura elettromagnetica perché se si agita un magnete a barra in una stanza buia, il sistema non emette luce visibile. Ora, chiedono i Churchland, l'argomento della stanza cinese non è esattamente uguale? Non dice semplicemente che se si agitano i simboli cinesi in una stanza dove regni il buio semantico essi non emettono la luce della comprensione del cinese? Ma come, in seguito a ulteriori indagini, si scoprì che la luce è costituita soltanto da radiazione elettromagnetica, non potrebbe allo stesso modo accadere che, in seguito a ulteriori indagini, si scoprisse che la semantica è costituita soltanto da sintassi? Non è questo un problema che merita ulteriori ricerche scientifiche?

Si sa che i ragionamenti per analogia sono deboli; perché il ragionamento tenga, infatti, si deve accertare che i due casi siano davvero analoghi. E qui credo che non lo siano. La spiegazione della luce in termini di radiazione elettromagnetica ha carattere esclusivamente causale: è una spiegazione causale della fisica della radiazione elettromagnetica. L'analogia con i simboli formali non sussiste perché i simboli formali non hanno alcun potere causale fisico. L'unico potere che essi hanno in quanto simboli è quello di causare il passo successivo del programma quando la macchina è in funzione. E non si tratta certo di aspettare che ulteriori ricerche rivelino le proprietà fisiche causali delle cifre 0 e 1. Le uniche proprietà rilevanti di queste cifre sono proprietà computazionali astratte, che sono già ben conosciute.

I Churchland mi accusano di «dare per dimostrata la tesi» quando affermo che i simboli formali non interpretati non coincidono con i contenuti mentali. È vero che non ho dedicato molto tempo alla dimostrazione di questo punto, perché lo considero una verità logica. Come accade per qualunque verità logica, si vede subito che è vera perché se si cerca di immaginare il contrario si ottengono delle contraddizioni. Proviamo. Supponiamo che nella stanza cinese si svolga davvero qualche pensiero cinese non rilevabile. Che cosa esattamente dovrebbe trasformare la manipolazione degli elementi sintattici in contenuti di pensiero specificamente cinesi? Ebbene, io suppongo che i programmatori sappiano il cinese e che abbiano programmato il sistema in modo da fargli elaborare informazione in cinese.

Benissimo. Ma ora immaginiamo che il nostro uomo, che sta nella stanza cinese a mescolare gli ideogrammi cinesi, si stufi di mescolare soltanto questi simboli, che per lui sono privi di significato. Supponiamo dunque che l'uomo decida di interpretare i simboli secondo le mosse di una partita a scacchi. Quale semantica esprime ora il sistema? Esprime la semantica del cinese, la semantica degli scacchi o entrambe? Supponiamo che vi sia un altro uomo che guarda dentro la stanza da una finestra e che costui decida che queste manipolazioni dei simboli possano essere interpretate come previsioni dell'andamento della borsa. E così via: non c'è limite al numero delle interpretazioni semantiche che si possono assegnare ai simboli perché, ripeto, essi sono puramente formali. Essi non posseggono una semantica intrinseca.

C'è qualche modo per salvare dall'incoerenza l'analogia dei Churchland? Ho detto prima che i simboli formali non posseggono proprietà causali. Ma naturalmente il programma sarà sempre eseguito in circuiti di qualche tipo, e questi circuiti avranno poteri fisici causali specifici. E qualunque calcolatore reale sarà sede di svariati fenomeni. I miei calcolatori, per esempio, emanano calore ed emettono un ronzio e talvolta uno scricchiolio. C'è allora qualche motivo dotato di cogenza logica per cui essi non debbano emanare anche coscienza? No. Sotto il profilo scientifico il problema non si pone neppure, ma non si tratta di qualcosa che l'argomento della stanza cinese debba confutare e non è qualcosa che un seguace dell'IA forte desidererebbe sostenere, perché un fenomeno del genere dovrebbe scaturire dalle caratteristiche fisiche del mezzo con cui è costruita la macchina. Ma la premessa di fondo dell'IA forte è che le caratteristiche fisiche del mezzo sono affatto irrilevanti. Ciò che conta sono i programmi e i programmi sono puramente formali.

L'analogia dei Churchland tra sintassi ed elettromagnetismo si trova dunque di fronte a un dilemma: o la sintassi è costruita per via semplicemente formale in termini delle sue proprietà matematiche astratte oppure non lo è. Se lo è, l'analogia non regge perché una sintassi così costruita non ha poteri fisici e quindi non ha poteri causali fisici. Se, viceversa, si pensa in termini del mezzo fisico, allora l'analogia sussiste, ma non è un'analogia che interessi l'IA forte.

Poiché le osservazioni che ho fatto sono piuttosto ovvie (la sintassi non coincide con la semantica; i processi cerebrali causano fenomeni mentali) è naturale chiedersi come siamo finiti in questo pasticcio. Com'è possibile che qualcuno creda che una simulazione al calcolatore di un processo mentale possa coincidere con il processo mentale? In fin 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?

La risposta, in parte, è che abbiamo ereditato un residuo delle teorie psicologiche comportamentiste della generazione precedente. Nel test di Turing è racchiusa la tentazione di pensare che se qualcosa si comporta come se avesse certi processi mentali allora deve avere davvero questi processi mentali. E questo fa parte dell'erroneo assunto del comportamentismo che, per essere scientifica, la psicologia deve limitarsi a studiare il comportamento esterno osservabile. Paradossalmente, questo comportamentismo residuo è collegato a un residuo dualismo. Nessuno pensa che una simulazione al calcolatore della digestione possa realmente 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. 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.

Nel corso della storia, le dottrine scientifiche che in Occidente hanno trattato gli uomini semplicemente come parte dell'ordine fisico e biologico comune sono state spesso contrastate da manovre di retroguardia d'ogni sorta. Copernico e Galileo furono combattuti perché sostenevano che la Terra non è collocata al centro dell'universo; Darwin fu combattuto perché affermava che l'uomo discende da animali inferiori. Il modo migliore per inquadrare l'IA forte è di vederla come uno degli ultimi sussulti di questa tradizione antiscientifica, perché nega che esista nella mente umana qualcosa di sostanzialmente fisico e biologico. Secondo l'IA forte, la mente è indipendente dal cervello: è un programma di calcolatore e come tale non è legata ad alcun specifico substrato circuitale.

Molti di coloro che nutrono dubbi sulla portata psicologica dell'intelligenza artificiale pensano che i calcolatori potrebbero sì capire il cinese e ragionare sui numeri, ma non potrebbero fare certe cose squisitamente umane, cioè (e qui segue la specialità umana che preferiscono): innamorarsi, avere il senso dell'umorismo, percepire l'angoscia della società postindustriale nell'era del tardo capitalismo, o quant'altro. Ma i ricercatori di IA obiettano, giustamente, che così si sposta via via il traguardo. Appena all'intelligenza artificiale riesce una simulazione, essa cessa di avere importanza psicologica. In questo dibattito nessuno dei due contendenti percepisce la distinzione tra simulazione e riproduzione. Per quanto riguarda la simulazione, è facilissimo programmare un calcolatore in modo che scriva «Susi, ti amo»; «Ha ha»; oppure «Soffro l'angoscia della società postindustriale nell'era del tardo capitalismo». Ma è importante rendersi conto che la simulazione non coincide con la riproduzione e l'importanza di questo fatto è la stessa tanto per il pensare di aritmetica quanto per sentire l'angoscia. Non è che il calcolatore arrivi solo fino alla metà campo invece di arrivare fino all'area di rigore. Il calcolatore non parte neppure: non gioca a questo gioco.

BIBLIOGRAFIA

HAUGELAND JOHN (a cura), *Mind Design: Philosophy, Psychology, Artificial Intelligence*, The MIT Press, 1980.

SEARLE JOHN, *Minds, Brains, and Programs* in «Behavioral and Brain Sciences», 3, n. 3, pp. 417-458, 1980.

SEARLE JOHN R., *Minds, Brains, and Science*, Harvard University Press, 1984.

HARNAD STEVAN, *Minds, Machines and Searle* in «Journal of Experimental and Theoretical Artificial Intelligence», 1, n. 1, pp. 5-25, 1989.