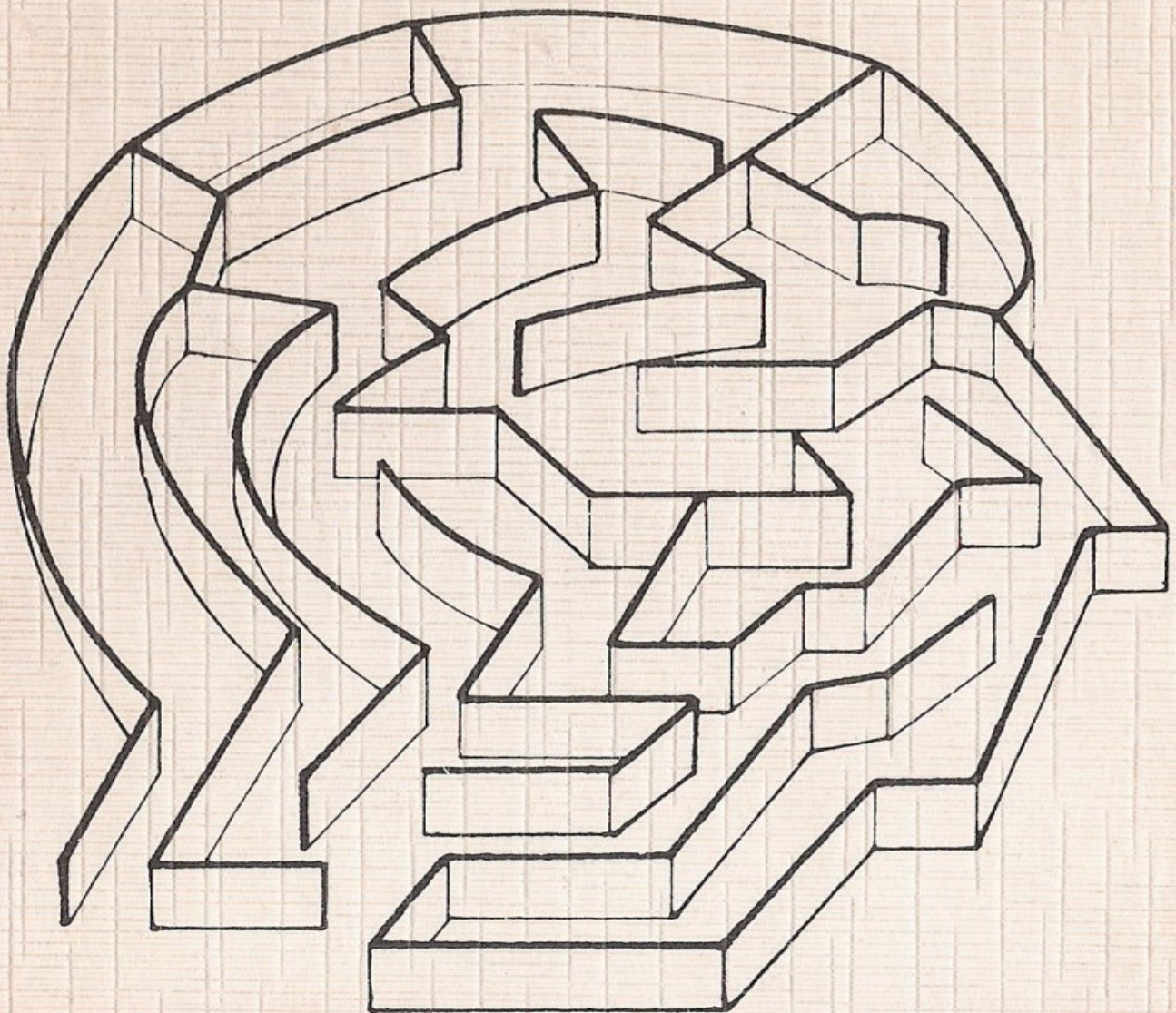
scienza · tecnica · società

JOHN R. SEARLE

MENTI, CERVELLI E PROGRAMMI

UN DIBATTITO SULL'INTELLIGENZA ARTIFICIALE

a cura di Graziella Tonfoni



CLUP · CLUED

JOHN R. SEARLE

MENTI, CERVELLI E PROGRAMMI

UN DIBATTITO
SULL'INTELLIGENZA
ARTIFICIALE

a cura di Graziella Tonfoni

CLUP-CLUED

Stampato presso le grafiche GV, viale umbria 36, milano
per conto della clup, piazza leonardo da vinci 32, milano
e della clued, via celoria 20, milano

Titolo originale:
Minds, Brains and Programs,
in *The Behavioral and Brain Sciences*, 1980
Cambridge University Press

in copertina:
disegno da un’idea di Giuliana Mengozzi

prima edizione: giugno 1984
ristampa:
III II I 1985 1986 1987
ISBN 88-7005-614-7
ISBN 88-7059-082-8

SOMMARIO

INTRODUZIONE

Graziella Tonfoni
The Artificial Intelligence Laboratory
MIT Cambridge, Mass., USA

Daniel Schneider
Dept. of Political Science
MIT Cambridge, Mass., USA

To Marvin Minsky and Gloria Rudisch Minsky
for their advice and wonderful friendship

Perché questa introduzione?

Scrivere un'introduzione al *Great Debate* nato dall'articolo di Searle "Minds, Brains and Programs" richiede senz'altro una scelta preliminare: si potrebbe cioè pensare a un'ulteriore interpretazione o valutazione del famoso esempio adottato da Searle, oppure a una presentazione dei fondamentali concetti che risultano essere i più frequentemente usati e dibattuti nell'ambito dell'attuale ricerca sull'Intelligenza Artificiale.

Il lettore ha già a sua disposizione una serie più che abbondante di interpretazioni, valutazioni e critiche alle tesi di Searle; si è per questo pensato di scegliere la seconda strada, ovvero di presentare un panorama teorico sufficientemente informativo della ricerca in corso nel settore dell'Intelligenza Artificiale, al fine di favorire piuttosto la comprensione delle tematiche già così ampiamente discusse nel corso del "Grande Dibattito" stesso. La presente introduzione viene quindi a costituire un capitolo a sé, quasi una breve storia critica delle più recenti e rilevanti acquisizioni teoriche.

Si vuole peraltro ricordare che quello dell'Intelligenza Artificiale è un campo assai giovane e dinamico, basato su assunzioni teoriche e problemi che hanno sempre affascinano, interessato e stimolato la mente umana fin dai tempi più antichi. Questa è la ragione che ci ha spinto a considerare quelle teorie di carattere filosofico, psicologico e matematico che costituiscono il background diretto e la "struttura portante" della ricerca attuale in Intelligenza Artificiale.

In breve, questa introduzione presenta una selezione specifica operata su tutta la vasta serie di informazioni che avremmo potuto dare. Tale selezione è avvenuta non arbitrariamente, ma col preciso intento di mettere in luce la pluralità delle interpretazioni relative a quei concetti specifici che sono i più usati nel corso del dibattito. In particolare abbiamo utilizzato la distinzione fra *ipotesi forte* dell'Intelligenza Artificiale (*strong Artificial Intelligence*) e *ipotesi debole* dell'Intelligenza Artificiale (*weak Artificial Intelligence*) così come già è stata adottata da Searle nel corso del suo intervento.

La nascita dell'Intelligenza Artificiale

Che cosa è l'Intelligenza Artificiale? Diamo un'occhiata alla seguente affermazione di Michie:

> Se possiamo formulare una teoria sufficientemente completa e precisa di ogni aspetto dell'intelligenza, allora possiamo anche convertirla in un programma di computer. Il programma stesso costituisce un'espressione della teoria, ma dovrebbe anche, se la teoria è valida, avere il potere

di far manifestare al computer un comportamento interamente simile a quello che la teoria pretende di saper descrivere.[1]

D'altro lato possiamo anche dire, seguendo Newell, che:

Le Scienze non sono definite, sono riconosciute.[2]

Invece di tentare di stabilire una definizione precisa di quello che l'Intelligenza Artificiale veramente significa, preferiamo presentare un riassunto delle tendenze di ricerca più rilevanti e dei maggiori eventi che hanno determinato in seguito sviluppi differenti.

Cominciamo con un evento storico: la Conferenza di Dartmouth. Possiamo considerarla come l'inizio ufficiale dell'Intelligenza Artificiale. Tenuta nell'estate del 1956, vi convennero scienziati provenienti da diversi ambienti: alcuni di essi erano matematici, altri psicologi, altri ingegneri elettronici. Idea comune era che i processi del pensiero possano aver luogo al di fuori della mente umana e che il miglior modo per riprodurre quei processi sia il computer. I nomi più rilevanti che devono essere menzionati a questo punto sono quelli di J. McCarthy, M. Minsky, A. Newell, C. Shannon, H. Simon. In effetti il termine di *Intelligenza Artificiale* non era in quel tempo particolarmente significativo, ma ha finito poi con l'essere quello con cui questa ricerca è stata denominata. Obiettivo iniziale era dunque quello di sviluppare un progetto di ricerca su tutto quanto riguarda i processi dell'intelligenza, dell'apprendimento e della simulazione.

Settori di ricerca dell'Intelligenza Artificiale

Diremo ora due parole sul significato del termine *intelligenza*. Invece di definirla, diremo piuttosto che l'intelligenza è basata su varie attività mentali, come il ragionare sulla base del cosiddetto "senso comune", il comprendere, il dedurre, l'indurre e così via. Durante gli ultimi anni sono stati costruiti molti sistemi per simulare tali processi.

Possiamo definire questo settore di studio come di Intelligenza Artificiale volta alla ricerca applicata. Esso presenta infatti un orientamento empirico e fondamentalmente ingegneristico. Il punto di avvio consiste in un corpo di tecniche computazionali su cui i sistemi creati dall'Intelligenza Artificiale sono basati: tali sistemi saranno poi sottoposti a una sperimentazione e, da ultimo, perfezionati. Per dare un'altra definizione generale dello stato di questa disciplina, possiamo dichiarare che l'Intelligenza Artificiale (IA) ha un fine scientifico più ampio, che consiste nel costruire una teoria dell'elaborazione dell'informazione e una teoria dell'intelligenza. La ricerca

in questo campo è ancora in atto: è certamente vero che una scienza dell'intelligenza può essere sviluppata, e che tale scienza potrà essere in grado di guidare la progettazione di macchine intelligenti e di presentare un'adeguata teoria del comportamento intelligente negli animali e negli esseri umani. Dobbiamo dire però che una tale idea è molto più un obiettivo ipotetico che una realizzazione effettiva.

L'IA può essere distinta in più aree di ricerca aventi differenti specializzazioni. Prima di tutto si deve evidenziare il *problem solving* ovvero la soluzione dei problemi basata sul ragionamento logico. Un altro importante tentativo viene fatto nel *theorem proving* ovvero nella ricerca di una prova o una smentita riguardante uno specifico teorema in matematica, ove si richieda la capacità di compiere processi di deduzione data l'esistenza di un'ipotesi specifica di partenza; tale processo richiede anche abilità come il *guessing*, ovvero l'immaginare ipotetico che deve poi essere soggetto a verifica. Si tratta infatti di "immaginare" quali lemmi devono essere utilizzati per provare il teorema principale e verificarne poi l'adeguatezza. Un matematico naturalmente usa un tipo di giudizio che è basato in larga misura su conoscenze specialistiche al fine di dividere un problema in una serie di sottoproblemi parziali e di risolverli poi progressivamente. I programmi che verificano un dato teorema e che sono già stati sviluppati presentano alcune di queste abilità, ma ancora in misura limitata. Lo studio della prova del teorema (*theorem proving*) è molto importante in IA. Una parte dei ricercatori in IA sostiene che formalizzare i processi deduttivi usando il linguaggio della Logica dei Predicati aiuta a capire più chiaramente alcuni dei componenti stessi del ragionamento. Un terzo campo di ricerca è quello che riguarda il linguaggio naturale e la sua comprensione. I problemi specifici riferiti a questo genere di ricerca sono relativi alla conoscenza derivante dal buon senso comune e dal ruolo giocato dalle aspettative create dal testo ogni volta che vogliamo realmente capire, tradurre o riprodurne una frase. I programmi che sono stati scritti finora possono essere usati per una serie semplice di frasi e riguardano una sfera abbastanza limitata di testi: la maggior parte dei problemi essenziali che sono connessi alla comprensione del testo devono infatti ancora essere risolti.

Un quarto campo di ricerca in IA è quello della programmazione. Il lavoro in tale settore è stato chiamato di "programmazione automatica" e tratta di sistemi che possono scrivere programmi per computer partendo da differenti descrizioni degli stessi esempi utilizzati. Molto lavoro è stato fatto ai fini di modificare e migliorare pure il processo di codificare (*coding*). Un quinto settore di ricerca in IA è quello connesso ai problemi dell'apprendimento. L'apprendere può essere considerato come uno degli aspetti più interessanti dell'intelligenza umana. Strategie che sono state usate al fine di migliorare l'apprendimento sono le procedure analogiche (*analogy procedures*), che

consistono nel fare acquistare nuove abilità guardando a precedenti esperienze oppure a conoscenze già acquisite per ricavare, da queste, nuove soluzioni.

Un sesto campo che deve essere ricordato è l'area dei sistemi esperti (*expert systems*): la ricerca relativa è conosciuta anche come "ingegneria della conoscenza" (*knowledge engineering*) ed è direttamente collegata a tecniche di applicazione. Un paio di parole su quello che può fare un sistema esperto: l'utente interagisce con un sistema esperto in ima conversazione che ha l'aspetto di una consultazione su problemi precisi nello stesso modo in cui interagirebbe con un essere umano, facendo cioè domande per cercare di dare soluzioni a problemi concernenti un settore in cui ha scarsa esperienza. I sistemi correnti hanno conseguito un alto livello di qualità di esecuzione in differenti campi.

Per dare qualche esempio ricorderemo qui i sistemi esperti bancari, i sistemi esperti nella diagnosi medica, e nell'analisi di dati chimici. L'attuale livello di ricerca presenta i sistemi esperti come intermediari tra esperti umani che interagiscono col sistema per acquisire informazioni, ed esperti umani che interagiscono col computer al fine di avere più informazioni o una conferma sulle loro ipotesi. L'obiettivo finale di questa ricerca è quello di sistemi che siano in grado di spiegare i loro ragionamenti e di fornire una consulenza che possa essere più accettabile e, infine, di aiutare l'utente a trovare, qualora esistano, i possibili errori nel sistema.

Molte tecniche vengono continuamente sviluppate per consentire una più efficace rappresentazione, immagazzinamento e recupero di una sempre crescente quantità di fatti. Il caso più interessante è naturalmente quello in cui si voglia fornire delle risposte che richiedano un ragionamento deduttivo a partire da fatti e informazioni presenti nella base-dati. Problema fondamentale nei sistemi esperti è rappresentare in modo utilizzabile la conoscenza che gli esperti umani in campo specifico hanno e usano correntemente. Un'ulteriore difficoltà sorge per il fatto che la conoscenza, in molti campi, risulta spesso essere frammentaria, incerta o episodica.

Un settimo campo di ricerca è quello connesso ai problemi di robotica e visione. Entrambi i settori di questa ricerca sono indirizzati alla costruzione di robot che devono diventare sempre più sofisticati e complessi.

Il background psicologico

La tradizione filosofica degli antichi greci si era già occupata di problemi riguardanti la mente e il corpo in vari modi, ma chi per primo raggiunse la conclusione che mente e corpo sono due cose del tutto diverse fu René Descartes, che divise gli atti umani in meccanici e razionali. I primi erano quelli che potevano essere imitati da automi, come il camminare e il

mangiare, a differenza dei secondi, come il volere e lo scegliere, che non potevano essere imitati. C'era comunque una relazione fra i due, che s'incontravano di fatto nella ghiandola pineale. Seguendo questa ipotesi si concludeva necessariamente che gli animali erano "macchine meravigliose". Gli esseri umani erano pure macchine, tranne il fatto che avevano anche una mente. Il dibattito sul dualismo cartesiano (la distinzione mente-corpo) divenne un fatto molto importante nella cultura occidentale. Cartesio distingueva pure due generi di idee, le une che venivano dall'esperienza sensoriale (idee derivate) e le più importanti, che si sviluppavano fuori dalla mente ed erano totalmente indipendenti dall'esperienza sensoria (idee innate). Leibnitz mantenne come base l'assunto che mente e corpo fossero separati, ma aggiunse che essi in effetti si combinavano esattamente, dando significato l'uno all'altro in un sistema di monadi. Il suo scopo era di ridurre il ragionamento a un'algebra di pensieri chiamata "calculus ratiocinator". Da questo genere di problema segue il parallelo dibattito sul comportamento.

Hobbes teorizzò che ogni aspetto del comportamento umano è semplicemente una testimonianza di un moto interno ed è specificamente ispirato dal timore e dall'interesse per sé stessi. Gli aspetti associativi della mente risultano collegati insieme non logicamente, ma solo contingentemente. Hobbes distingueva le associazioni puramente libere dal pensiero controllato. Hume teorizzò che le idee complesse fossero un risultato di quelle semplici e che la mente fosse appunto un flusso di sensazioni e ragionamenti. Sebbene le idee complesse siano come composte da quelle semplici, non assomigliano però a quelle semplici perché risultano da nuove combinazioni basate su differenti livelli di aggregazione.

De La Mettrie teorizzò "l'Uomo macchina". Essendo un fisico, eseguì un notevole numero di esperimenti sul corpo umano e alla fine concluse che le sostanze fisiche influenzano anzi "determinano" il modo in cui il pensare ha luogo. La prova era basata su esperimenti fatti sul modo di alimentarsi e sulla somministrazione di stupefacenti. Le idee di La Mettrie furono continuate e sviluppate da Diderot, che studiò gli esseri umani in quanto "macchine". Alla fine del diciottesimo secolo l'uomo come macchina era un'idea molto nota e ampiamente accettata, cosicché il cervello era effettivamente definito come un organo che "digerisce impressioni e secerne pensieri".[3]

Per concludere questa brevissima rassegna, ricordiamo l'idea di Kant sulla mente che possiede a priori princìpi che confermano e organizzano il mondo esterno e sé medesimi. In sostanza, il modo con cui il mondo è organizzato è determinato dalla nostra mente e non ha alcuna giustificazione da essa indipendente. Le più recenti ricerche storiche sembrano confermare questi ultimi assunti filosofici.

La psicologia fu una delle ultime scienze a staccarsi dalla filosofia e gli odierni collegamenti tra le due discipline sono ancora molto forti in numerose

aree di ricerca. Ne deriva che la filosofia della mente è strettamente connessa a un certo tipo di psicologia della mente, ed entrambe contribuiscono alla nuova Filosofia della Psicologia.

Per alcuni aspetti l'IA ha comunque applicato l'epistemologia, cercando di capire le condizioni del nostro pensare. La questione del funzionamento e dell'organizzazione della mente è naturalmente antichissima. I primi modelli, che risalgono ad Aristotele, trattavano dell'organizzazione delle idee, delle immagini e delle sensazioni. L'idea di Aristotele era di poter ricavare princìpi che ci dicano come combinare le entità mentali. Egli creò la mente "associazionista" e stabilì i tre famosi princìpi di associazione: similarità, opposizione e contiguità di tempo e di spazio. Queste leggi sono state completate da empiristi inglesi quali Hume e Locke mediante la postulazione delle "leggi secondarie di associazione". Solo più tardi, nel diciannovesimo secolo, psicologi come Ebbinghaus cercarono di affrontare in un modo non filosofico la questione se la composizione degli elementi potesse produrre un intero che fosse più di ciascuna delle sue parti singole. Tale linea di sviluppo mostra che è probabilmente più facile affrontare la questione di *come* usare le entità mentali che trattarne il contenuto. La psicologia strutturale, che oggi si può considerare estinta, ha sollevato l'importante questione di che cosa effettivamente fossero le entità mentali, e se potessero essere definite. B. Wundt, che è chiamato il padre della psicologia scientifica, ha creduto che ci debba essere una serie di "elementi mentali" comparabili alla tavola chimica. Egli ha considerato la mente stessa come un processo attivo che è ordinato e regolato in un modo non-arbitrario: funziona sulla base di leggi di associazione e sulla base di uno stimolo esterno. La teorizzata appercezione risulta così essere il prodotto congiunto della cognizione e della percezione. Questo brevissimo abbozzo della sua teoria mostra che già nel XIX secolo erano presenti due importantissimi concetti propri della moderna IA: primo, che la mente è considerata come un processo computazionale che funziona in un modo ordinato secondo elementi definiti; secondo, che la mente che lavora si basa sia su di un contesto "interno" che su uno "esterno", e conta sul potere di trasduttori.

Uno studioso spesso ricordato dai ricercatori dell'IA è il padre della "psicologia umanistica", cioè il biologo americano William James. Gli psicologi prima di James tendevano ad analizzare l'umano in differenti sottofenomeni: James si oppose invece a tale orientamento, introducendo di nuovo l'individuo nella psicologia. Per lui idee e reazioni sono il prodotto di un attore. Ciò non implica solo che la psicologia dovrebbe essere olistica, ma che la vita mentale è un processo che necessita di un supporto fisico. James dichiarò così che c'era una corrispondenza tra il mentale e il fisico, e la psicologia a sua volta divenne la scienza della vita mentale e del suo supporto fisico.

La questione mente-corpo, che abbiamo già menzionato, è ancora molto dibattuta in IA: la maggior parte dei ricercatori di IA crede in una certa indipendenza della mente, nel senso che si suppone che il cervello regga un sistema fisico di simboli che poi, a sua volta, può essere analizzato come sistema simbolico di simboli, perché i suoi processi sono organizzati simbolicamente, anche se dipendono fortemente da un certo tipo di "hardware". Ciò significa che l'organizzazione cerebrale, così come l'organizzazione di un computer, condiziona i processi simbolici. Questo non implica che la mente possa essere facilmente simulata da altri sistemi simbolici fisici quali un computer, ma piuttosto che le teorie su certi aspetti della mente possono essere esaminate attraverso il computer e che forse un giorno potranno essere costruiti computers tali da poter contenere sistemi simbolici fisici che si avvicinano molto di più ai nostri.

James ha presentato pure un'interessantissima teoria del "Sé". Il Sé ha tre componenti: il *Sé spirituale*, chiamato "Io", corrisponde alle capacità della mente. L'"Io" è quello che è cosciente, quello che ci definisce come parti di uno stato del flusso di coscienza. Il *Sé sociale*, il "Me", un aspetto spesso trascurato dall'IA, è creato dall'interazione con altre persone e riflette quello che sappiamo su noi stessi nel mondo. Infine il *Sé materiale*, il "Mio", ci definisce per mezzo delle cose che abbiamo, incluso il nostro corpo, cioè la nostra fisicità. I ricercatori dell'IA spesso si oppongono a tutte le forme di comportamentismo; nondimeno pensiamo che quella tradizione abbia avuto una grossa influenza sulla formulazione del paradigma dell'IA, ancora peraltro incompleto. Il comportamentismo radicale, come praticato da Watson e Skinner, è naturalmente contraddittorio rispetto all'idea dell'IA forte. Ma l'idea che la psicologia deve essere limitata a quello che può essere definito come il paradigma dello stimolo-risposta, sollevò l'importante questione della "dimostrabilità". Haugeland ha tentato di chiarire recentemente che i princìpi di reale dimostrabilità del behaviorismo possono essere in qualche modo evitati in un paradigma di IA. D'altro lato, una forte influenza di ricerca neobehavioristica, presente specialmente in Hull e Tolman, ebbe notevole peso nella formazione di alcuni concetti di IA. Hull molto presto riconobbe che anche un semplice modello stimolo-risposta deve contenere la nozione di "stimolo interno", e che il comportamento ha pure una guida interna. Tolman andò molto più in là: da un lato egli allargò la limitativa definizione di stimolo-risposta che fu sostituita dalla situazione in cui l'individuo reagisce allo stimolo con un'azione completa come risposta. D'altro lato, egli introdusse la "variabile infrapposta", un concetto che gli permetteva, di nuovo, di considerare un individuo come avente una mente. Concetti come quelli di mappe cognitive (*cognitive maps*), sistema dei bisogni (*need system*), matrice dei valori di credenza (*belief-value matrix*) permisero di disegnare la mente in un modo operazionalistico ben più dettagliato. Fu sempre Tolman

che discusse l'ormai famoso problema del ristorante, riguardante quello che una persona ha necessità di conoscere e di fare per ottenere cibo in un locale pubblico. Il tema è divenuto famoso in IA, dato il notevole contributo fornito da Schank a proposito dello *Script del ristorante*. Un'altra ancor viva tradizione, la psicologia della Gestalt, sollevò ulteriori, importanti questioni.

Tale tradizione, che già emergeva intorno al volger del secolo, accentuò l'importanza delle strutture dell'intero nei confronti della composizione dei singoli elementi. La più importante delle tre scuole della Gestalt fu fondata da Wertheimer, ma fu Ehrenfels che si distinse per la sua tesi che una melodia deve essere *di più* della semplice somma delle parti, per poter essere transposta: la melodia in tale caso sarà ancora la stessa, ma sarebbe composta di differenti note. I teorici della Gestalt avviarono quindi studi particolari sui problemi della percezione. Il modello S-O-R della percezione che sta per *Struttura* dello stimolo, *Organizzazione* della percezione, *Risposta* determinata dalla percezione fu più tardi allargato a un modello più generale per la comprensione. La tesi veramente fondamentale è che la realtà non è percepita direttamente, ma organizzata da un meccanismo di percezione strutturato gerarchicamente e piuttosto indipendente, definito come "campo fenomenologico". Quelli della Gestalt furono i primi ad accettare che il riconoscimento della struttura (*pattern recognition*) è fondamentale anche in riferimento al comprendere. Le cose ricevono il loro significato dall'essere associate a patterns di forme. Questo è vero anche per fenomeni complessi. La struttura di un pattern è a un certo punto indipendente dai suoi singoli elementi. Questo è facilmente dimostrato dal fatto che forme incomplete o strutture composte da differenti oggetti possono comunque essere riconosciute da un individuo. In conclusione, la mente umana necessita di meccanismi invarianti al fine di percepire e di comprendere la realtà. Ciò significa che ci devono essere strutture cognitive che permettono di immagazzinare forme generali nella memoria, e che ci sono processi che combinano i patterns in modo da mettere in relazione la conoscenza con uno specifico input.

Un ricercatore molto citato in IA è Bartlett, che studiò già nel 1932 l'area dell'analisi e del trattamento di un testo narrativo. La sua ricerca sulla memorizzazione mostrò che il richiamo è costruttivo: ci deve essere cioè un principio attivo (che egli chiamò *schema*) nella nostra memoria, che organizza e riorganizza elementi in complessi strutturati.

Il campo con il quale l'IA ha più interazioni oggi è proprio la psicologia cognitiva, lo studio del pensare. Questa teoria è piuttosto una escrescenza del neobehaviorismo e della teoria dell'apprendimento, se non si conta l'influenza che l'IA ha a sua volta avuto sulla psicologia per la creazione della nuova disciplina chiamata scienza cognitiva.

La psicologia evolutiva propria della tradizione di Piaget è sempre risultata

attraente agli occhi di certi ricercatori di IA, anche se non è molto il lavoro svolto in quell'ambito. Lavorando sullo sviluppo cognitivo del bambino, Piaget creò una teoria generale dell'apprendimento. In sostanza, gli individui umani sono caratterizzati da un sostrato biologico che permette alla mente di svilupparsi in fasi diverse durante l'interazione con l'ambiente. L'adattamento segue i principi dell'assimilazione, cioè la capacità di cambiare l'"universo" in conformità alla mente, e dell'accomodamento, cioè della possibilità di adattare la mente all'ambiente. Gli schemi di Piaget sono in pratica le odierne strutture di conoscenza ordinata e le capacità di adattamento corrispondono all'uso creativo di queste durante la soluzione di problemi specifici. Di più, Piaget mostrò che una teoria dell'intelligenza deve anche essere una teoria dello sviluppo dell'intelligenza.

Dal *Logic Theorist* al *General Problem Solver*

I primi programmi di Intelligenza Artificiale furono limitati a compiti specifici quali giocare, provare teoremi e risolvere rompicapo. La teoria corrispondente era per lo più la teoria della programmazione euristica o la teoria del problem solving.

Il *Logic Theorist* fu un programma creato nel 1956 da A. Newell, J. Shaw e H. Simon, della Rand Corporation e del Carnegie Institute. Fu il primo programma in grado di risolvere problemi con metodi euristici, e fu il primo sistema di IA completamente programmato. Gli obiettivi primari della sua creazione furono di imparare come sia possibile risolvere difficili problemi in generale e, in particolare, capire i complessi processi, chiamati anche euristici, che risultano operare nel corso della risoluzione dei problemi. Contrariamente a quanto si era verificato per altri programmi precedenti, gli autori non erano interessati ad avere un programma che potesse garantire soluzioni richiedendo un'enorme quantità di calcolo, ma piuttosto a scoprire come un matematico procederebbe per trovare una soluzione a un problema per il quale non esiste alcun algoritmo noto. Questa brevissima descrizione delle finalità implicate in Logic Theorist mostra di nuovo il principio fondamentale dell'IA: che un programma è una teoria o almeno parte di una teoria sul comportamento intelligente e che quest'ultimo deve essere espresso almeno parzialmente in un programma.

Nonostante l'obiettivo principale consistesse nello sviluppare una teoria della soluzione dei problemi (*problem solving*), gli autori procedettero in maniera del tutto deduttiva. Certe operazioni di problem solving furono compiute sulla macchina e studiate sulla macchina. Solo più tardi nell'approccio proprio del *General Problem Solver* fu studiato anche il comportamento cognitivo umano. Più concretamente, gli autori decisero di

studiare come provare teoremi nella logica simbolica elementare di Whitehead e Russell. Questo ambito, come altri studiati dal problem solving meccanico, aveva regole definite. Ma il fatto di avere regole definite non significa che il modo in cui devono essere applicate o seguite è ugualmente ben definito. Come ogni matematico sa, prove che potrebbero apparire eleganti e facili sono raggiunte solo dopo un lavoro considerevole.

I *Principia Mathematica* sono uno dei lavori più difficili nonostante gli autori si siano necessariamente limitati a un settore specifico del complesso problema. Essi scelsero di lavorare sul calcolo proposizionale, cioè su un sistema matematicamente formalizzato consistente in espressioni costruite da combinazioni di simboli base. Una parte di queste espressioni è costituita da assiomi da cui tutte le altre espressioni, chiamate teoremi, vengono derivate. Al Logic Theorist fu data una serie di cinque assiomi e tre regole di inferenza che permettessero di fare deduzioni. Quando al programma fu dato un teorema da provare esso applicò le tre regole al teorema al fine di ridurlo ad assiomi, prese cioè una delle regole di inferenza e un assioma e li applicò al teorema per produrre una nuova espressione da esaminare finché essa risultò essere costituita da espressioni semplici. Tale modo molto inefficace di procedere è chiamato *Algoritmo del British Museum*: prese tale nome in seguito alla dichiarazione che tutti i libri del British Museum avrebbero potuto essere scritti da scimmie provviste di macchine da scrivere, concesso loro un margine di tempo sufficiente.

Il Logic Theorist non è mai stato un grosso successo: non ha infatti creato un programma intelligente che può agire come gli esseri umani in un certo ambito di competenza.

Peraltro ha contribuito a creare molte idee che furono usate più tardi in altri programmi. Suo diretto successore fu il *General Problem Solver*, un sistema creato da un potente programma di ricerca il cui scopo non era solo quello di risolvere problemi mediante la macchina, ma anche di sviluppare una teoria del modo in cui gli esseri umani risolvono tali problemi. Le più importanti affermazioni sui processi del problem solving umano sono le seguenti:

- gli esseri umani sono rappresentabili come sistemi di analisi di informazioni;
- è possibile simulare tali processi e possiamo descrivere un problema con una serie di dati;
- formalizzazione significa elaborare un programma in cui l’intero sistema di analisi dell’informazione può essere compreso. Un sistema consiste di programmi. Così la simulazione può essere complessa e automatizzata;
- differenze di compiti esistono fra i programmi. Questo significa che le persone affrontano i problemi in maniera diversa e che i problemi sono diversi l’uno dall’altro;
- un contesto può essere complesso, il risolutore umano del problema deve

ridurre tale complessità al fine di risolvere il problema.

Nonostante le numerose possibili critiche che possono esser mosse, certe nozioni della teoria del problem solving sono ancora in uso oggi.

Ora brevemente esamineremo i caratteri essenziali dell'importante principio di analisi dei mezzi e dei fini. Nel contesto in cui è stata scritta la parte principale del General Problem Solver (GPS), un problema è equivalente alla differenza tra uno stato corrente A del mondo e il desiderato stato B del mondo. Così un problema genera l'obiettivo di ridurre quella differenza, cioè di trasformare A in B, e cerca i mezzi per farlo. Un problema può essere delineato più precisamente in termini di oggetti e operatori. Gli oggetti sono descritti dai loro caratteri e dalle differenze che possono essere osservate tra coppie di oggetti. Un problema può così essere scomposto in sottoproblemi. Un operatore può essere applicato a certi oggetti ben definiti al fine di produrre oggetti diversi o nuovi. Ora ogni problem solver possiede le cosiddette "tabelle di differenza" che gli indicano come applicare un operatore per ridurre una differenza specifica. Il genere di conoscenza che ne deriva è dipendente dagli obiettivi previsti e contiene anche una descrizione del problema stesso. Per operare, un GPS utilizza diversi tipi di euristica che possono essere usati ricorsivamente. Quando al GPS si dà un problema, cioè la differenza fra due stati, esso si concentra sull'obiettivo di trasformare uno stato A in uno stato B. Tutto si fermerà se non si può scoprire alcuna differenza tra i due, altrimenti GPS cercherà di applicare un operatore ad A. Se ciò genera un A' che combacia con B, il processo si fermerà. Se ciò non avviene, e questo è generalmente il caso nella fase iniziale del processo, GPS cercherà di ridurre la differenza tra A e B in modo graduale, creando A' che sia più vicino a B di quanto non lo fosse A, e applicando di nuovo altri metodi. Un metodo viene scelto se sembra promettente; quando fallisce, viene fatto un altro tentativo.

Consideriamo ora un altro importante sistema di IA, il sistema di produzione (*Production System* - PS), che si occupa della definizione e specificazione di un algoritmo. Le principali caratteristiche sono le cosiddette regole di riscrittura (*rewrite rules*) che sono usate per definire grammatiche generative in linguistica. Una regola di riscrittura prende una catena di simboli e la riscrive in un'altra catena di simboli. Una grammatica generativa è una collezione di regole di scrittura che permettono di tradurre, in successive fasi, una catena di simboli in una parallela assiomatica catena di simboli. Questo procedimento è usato per eseguire *bottom-up parsing*. Lo stesso procedimento può essere usato nel modo opposto: partendo da un assioma può essere generata una stringa più lunga di simboli. Questo procedimento può essere usato in *top-down parsing*. Le tre componenti essenziali di un PS sono:

1. serie di regole di produzione (riscrittura);

2. una base-dati;
3. un interprete che usa 1) e 2) al fine di eseguire calcoli.

In un semplice sistema di produzione una regola ha la seguente forma generale: “Se <SITUAZIONE> allora <AZIONE>”. Ciò significa che se l’interprete scopre una certa situazione o condizione nella base-dati che combacia con la “situazione” di una regola data, la regola è pronta per essere usata. Così la funzione dell’interprete è di valutazione o di interpretazione. Esso esamina la base-dati al fine di trovare regole che possa eventualmente usare. Se trova solo una regola che combacia con una situazione, tutto è risolto; se trova parecchie regole che potrebbero essere usate, deve scegliere la migliore. Qui sta la sua funzione di controllo. Molto spesso deve basarsi su un’altra serie di regole al fine di potere proseguire. Si deve aggiungere che alcune regole potrebbero rivelare al sistema che un problema è già risolto o che non è risolvibile. Ora è naturalmente possibile che una regola di riscrittura possa essere riscritta, o che il sistema possa cambiare l’ordine in cui preferisce utilizzare una serie di regole.

Naturalmente, anche GPS aveva un’organizzazione più complessa di quella che abbiamo abbozzato sopra, e i sistemi di produzione più moderni hanno, talvolta, solo il nome in comune con gli originali, visto il veloce sviluppo della ricerca.

La Seconda Generazione

La cosiddetta “Seconda Generazione” nella ricerca di IA è caratterizzata da un allargamento di interessi. Molta attenzione viene data alla comprensione del linguaggio umano così come al calcolo simbolico e alla rappresentazione semantica. La Seconda Generazione ha visto anche la nascita dell’IA applicata: è da ricondurre infatti a tale periodo l’attuazione di expert systems oggi utilizzati. Appartengono a tale periodo programmi come SHRDLU, ELIZA e PARRY. SHRDLU contraddistingue un programma per la conversazione in linguaggio naturale sviluppato presso il MIT tra il 1968 e il 1970 da T. Winograd. SHRDLU “vive” in un mondo di semplici oggetti geometrici come blocchi, piramidi e cubi di cui conosce le misure, i colori, le posizioni. L’oggetto più completo è una scatola in cui gli altri oggetti possono essere posti; tutti questi erano originariamente collocati su di un tavolo e potevano essere mossi dall’immaginario braccio mobile di un robot. SHRDLU ha tre componenti: un analizzatore sintattico, un modulo semantico e un risolutore di problemi. Facciamo un esempio di come funziona; se si dice al programma: “Per favore, poni sui blocchi rossi un cubo verde o la piramide verde”, la prima operazione è il riconoscimento della “possibilità” sintattica della frase. Si passa poi all’interpretazione semantica. Infine si controlla

l’esistenza reale degli oggetti nominati e la reale possibilità di eseguire quei movimenti. Se esiste confusione o c’è ambiguità, SHRDLU chiederà precisazioni. SHRDLU dimostrò l’importanza dell’analisi del significato di una frase e non solo della sua sintassi. Il modello di comprensione del linguaggio che Winograd ne derivò, si basa sui seguenti princìpi:

1. le frasi in un linguaggio naturale corrispondono a realtà esistenti nel mondo;
2. è possibile creare un sistema di rappresentazioni formali tale che sulle strutture di rappresentazione si eseguano ragionamenti attraverso operazioni formali sistematiche.

L’importanza di SHRDLU sta nel fatto che si sperimenta la percezione del mondo esterno reale attraverso la costruzione di rappresentazioni di oggetti fisici.

La nascita del paradigma di Yale

Il principale obiettivo della ricerca in IA svoltasi a Yale è sempre stato la comprensione del linguaggio naturale. Cominciato come un lavoro individuale di Roger Schank negli ultimi anni Sessanta, è diventato un paradigma di ricerca importante in tutta la scienza cognitiva orientata verso l’IA. Partendo dall’idea che ci debba essere una rappresentazione di significato indipendente dalla varietà delle lingue, Schank proclamò che “c’è una predeterminata serie di possibili relazioni che creano una struttura di significato in ogni lingua”.[4] Queste relazioni furono modellate mediante regole concettuali, usate — fra le altre cose — anche per predire elementi concettuali mancanti o impliciti in una frase.

Ci sono parecchie fasi storiche di sviluppo nell’ambito del paradigma di Yale: gran parte degli elementi di una prima fase sono integrati nella fase successiva (questa è una situazione del tutto unica nel mondo dell’IA: in generale, infatti, la maggior parte delle ricerche in IA si basa sulla strutturazione di sistemi sempre nuovi, anche se talvolta influenzati dal lavoro precedente).

Consideriamo molto brevemente i principali stadi del paradigma di Yale, che in effetti ebbe il suo avvio a Stanford: partito come lavoro linguisticamente orientato, è diventato più specificamente di IA, quando Schank e i suoi studenti iniziarono a considerare il concetto di *evento* e più tardi si posero il problema di quali inferenze si devono fare per capire un semplice evento. La lista presenta brevemente le principali fasi storiche del paradigma di Yale:

1. traduzione automatica;
2. semantica computazionale;

3. inferenza concettuale;
4. strutture a più elevato livello di conoscenze;
5. comprensione integrata;
6. organizzazione di memorie.

Discuteremo ora alcuni aspetti del sistema MARGIE e della sua teoria. MARGIE, presentato in forma di libro nel 1975[5] fu uno dei primi programmi in IA che potevano trarre conclusioni o fare parafrasi partendo da un input di linguaggio naturale e presentando un output di nuovo in linguaggio naturale. Naturalmente MARGIE non fu inteso solo come un contributo alla linguistica computazionale, ma come l'espressione di una teoria generale sull'analisi dell'informazione umana. Il sistema ha tre moduli, un segmentatore del linguaggio, un modulo di memorizzazione e inferenza, e un generatore di linguaggio. Tutti questi tre moduli si basavano sulla teoria della *dipendenza concettuale* di Schank che è un modello di rappresentazione semantica non dipendente dalla specifica lingua di realizzazione ma di valore generale. Oggi si possono distinguere tre diversi tipi di uso della rappresentazione: l'uso in "micro" (afferrare il pieno significato di una frase), l'uso in "macro" (organizzare il testo), e l'uso "speciale" in certi programmi. Quando una data frase è rappresentata come una struttura di dipendenza concettuale essa è tradotta in una serie di concettualizzazioni collegate da relazioni logiche di causa-effetto. Schank descrive la teoria della dipendenza concettuale come una teoria della rappresentazione del significato delle frasi nei seguenti cinque punti:

1. per ogni coppia di frasi che sono identiche nel significato, indipendentemente dalla lingua, ci dovrebbe essere una sola rappresentazione;
2. ogni informazione che è implicita in una frase deve essere resa esplicita nella rappresentazione del significato di quella frase;
3. la proposizione del significato è chiamata concettualizzazione: quest'ultima può essere attiva o stativa;
4. una concettualizzazione attiva ha la forma: Attore-Azione-Oggetto-Direzione con (Strumento);
5. una concettualizzazione stativa ha la forma: Oggetto (è in) Stato con (Valore).

Un evento è un semplicissimo *frame* (Minsky, 1975); in generale un evento ha:

- un attore;
- un'azione rappresentata da quell'attore;
- un oggetto su cui l'azione è rappresentata;
- molto spesso una direzione in cui l'azione è orientata (da, a…);
- qualche volta un modo (vero, falso, forse).

Così un semplice evento fondamentalmente consiste in un nome e in una

lista di spazi che prendono diversi argomenti. In generale un sistema di IA come MARGIE cerca di riempire questi spazi. L'informazione che si deve inserire deriva o dalla frase analizzata, o dalla conoscenza del mondo che il sistema ha, o da più complicati meccanismi d'inferenza. Schank sviluppò anche un linguaggio di rappresentazione grafica che permette di presentare una serie di concetti in modo chiaro. Consideriamo un esempio: il concetto di "amare qualcuno" potrebbe essere rappresentato dalla seguente sequenza:

```
(AMA (ATTORE ?X)
    (OGGETTO ?Y))
```

"Amare" è rappresentato in un modo relazionale. Consiste in un predicato AMA e due argomenti, l'ATTORE che ama e l'OGGETTO che è amato. Un predicato AMA completato potrebbe essere il seguente:

```
(AMA (ATTORE SCHANK)
    (OGGETTO IA))
```

Schank sostiene che le rappresentazioni della dipendenza concettuale potevano consistere in un ristretto numero di primitivi, che includono atti primitivi e stati primitivi: egli ha presentato inizialmente una lista di circa 13 atti fondamentali, che sono in grado di rappresentare gran parte di tutti i possibili eventi. Questi atti basilari possono essere usati per descrivere e, più importante, per capire la maggior parte delle azioni e degli stati che capitano negli eventi di ogni giorno. Tali "primitivi" permettono di definire regole di inferenza e, quello che è più importante, permettono una combinazione di patterns relativamente semplice. Daremo solo un esempio di alcuni atti primitivi:

MBUILD è la costruzione di nuova informazione dalla vecchia, che include modi di pensiero come "inferire", "dedurre", "decidere", "pensare", ecc.;

PTRANS è il trasferimento dell'ubicazione fisica di un oggetto. Tale primitivo può anche essere usato per descrivere il camminare; si deve semplicemente dire che "X PTRANSED un oggetto da Y a Z tramite l'uso dei suoi piedi";

MTRANS è il trasferimento di informazione mentale tra persone o all'interno di una persona. Il verbo "dire" potrebbe essere un esempio di MTRANS.

Il lettore potrebbe chiedersi se una tale semplice serie di primitivi non perda troppa informazione. A controbattere tale punto, Schank dichiara che sarebbe piuttosto la rappresentazione completa di una frase in una forma più complicata a far perdere informazione. Spesso si usa una scala per rappresentare lo stato di un oggetto. Quella scala può essere numerica con un tratto da -10 a +10, dove "-" è negativo, cattivo, ecc. e "+" è positivo. Il "(SALUTE -10)" significa per esempio che una persona è morta. Ci sono anche scale verbali, come [positivo, negativo] o [rotto, depresso, non buono, soddisfacente, felice, perfetto], Schank usa anche una serie di primitivi per esprimere relazioni tra concettualizzazioni più semplici.

Varie azioni o stati possono influenzare solo un certo tipo di azioni o stati. Descrivere queste regolarità è la funzione della sintassi causale. L'uso di tale tipo di causalità è molto importante nel sistema MARGIE, ma meno notevole in sistemi più moderni, dove le inferenze causali sono spesso fatte utilizzando altre strutture di conoscenza.

Tale linguaggio di rappresentazione è stato pesantemente criticato da molti ricercatori nella comunità di IA. La maggior parte di essi ammette che i programmi di Schank sono capaci di eseguire un certo numero di compiti interessanti: nega però vigorosamente la plausibilità psicologica della rappresentazione della dipendenza concettuale. Alcuni dichiarano che i primitivi di Schank sono troppo semplici, altri negano la possibilità di ridurre il linguaggio naturale a primitivi. Questi ultimi ammettono la possibilità che ci sia qualcosa come un "mentalese" (o "linguaggio mentale"), ma sostengono che esso non è basato su primitivi. Schank stesso, del resto, era d'accordo con alcune delle critiche: ammetteva per esempio che la serie originale di primitivi poteva essere applicata solo a certi tipi di testo. Egli modificò anche l'idea che gli individui creino una struttura molto dettagliata e coerente del contenuto del testo; quando un testo è analizzato, è piuttosto collegato a strutture di più generale conoscenza. La teoria della dipendenza concettuale è sempre stata connessa alla segmentazione e analisi del linguaggio naturale (*parsing*). Il parsing di Yale si basa sempre in primo luogo sul significato e riduce enormemente la dipendenza del parser della sintassi.

Il primo parser frase-per-frase realmente riuscito è stato il parser di Riesbeck per il sistema MARGIE, più tardi nominato ELI. In ELI, una parola attiva un procedimento che seleziona e attende un successivo input. Questo avviene mediante processi di richiesta che sono una forma di produzione (regole di SE-ALLORA). In ELI ogni parola o, più precisamente, ogni gruppo di parole simili è così rappresentato come un tipo di programma. All'inizio, come nel sistema MARGIE, un programma di IA di Yale usò un tipo di parser che traduceva le frasi secondo criteri di dipendenza concettuale, prima che lo stesso programma di IA trattasse l'informazione. In questo caso una frase era tradotta prima in una struttura relativamente semplice e, successivamente, resa più complessa e collegata mediante nessi causali. Nel programma POLITICS, Carbonell operò dapprima un parsing integrato, dove il ragionare e il comprendere furono trattati allo stesso livello e allo stesso momento. Tale tipo di parsing semantico diretto aveva dei vantaggi perché un input veniva direttamente tradotto in qualcosa che un programma di IA può capire. Il completamento del sistema MARGIE fu un programma di inferenza che prendeva una proposizione e poteva dedurne un gran numero di altri fatti dalla memoria con l'aiuto di una serie di meccanismi di inferenza.

Il principio generale dell'inferenza concettuale è il seguente: ogni configurazione a dipendenza concettuale provoca una risposta che genera

molte nuove possibili predizioni sulla più ampia situazione di cui l'input potrebbe essere parte. Scopo primario è quello di scoprire le relazioni con altre configurazioni che già conosce, cioè di agganciarlo a quelle usando le relazioni causali. Nel fare ciò, il programma deve indagare perché certi eventi sono avvenuti in una "storia" (catena di eventi) e che cosa potrebbe accadere in futuro. Così la comprensione del linguaggio va molto al di là dello studio del linguaggio; interessa qualsiasi aspetto che possa essere rappresentato dal linguaggio e dall'organizzazione logica sottostante della mente che rende conto della nostra capacità di comprendere.

Consideriamo ora l'organizzazione della memoria: la memoria consiste di concetti interrelati, eventi, azioni, caratteri di azioni, stati, e strutture più complesse quali i processi dinamici. Facciamo un esempio: un concetto è un'entità che definisce qualcosa che esiste nel mondo. Consideriamo il concetto "filosofo". Un simbolo come "il filosofo Searle" darebbe un esempio concreto di quel concetto. Ma siccome ci sono molti Searle vogliamo definire il nostro Searle in un modo più preciso. In questo caso creiamo un *superatomo* "Searle" al quale noi aggiungiamo una serie di informazioni che sono chiamate *serie di occorrenza* e che consistono in un catalogo che presenta ogni caratteristica conosciuta su tale concetto. I caratteri sono relazioni come (PROFESSIONE Searle FILOSOFO). Ma nel programma argomenti come "Searle" o "filosofo" sono sostituiti da simboli che danno a ogni simbolo un numero. In questo modo la memoria è organizzata come una grossa rete in cui ciascun elemento è connesso almeno a un altro.

Ora possiamo brevemente volgerci al concetto di inferenza. Nel suo programma, Rieger sostiene che il cervello o la mente di un essere umano esegue un'incredibile quantità di computazione "nascosta" nel cosiddetto spazio di inferenza che non è soggetto al controllo conscio. Un input di linguaggio naturale attiva parte della struttura della memoria in certi punti: a causa della complessa natura della memoria e anche delle sue regole di inferenza tutti questi "punti" sono sistematicamente attivati durante la comprensione di un testo. Secondo Rieger chi comprende il linguaggio naturale ha bisogno di operare inferenze arbitrarie al fine di comprendere.

Un diverso approccio al problema è quello del modello di Charniak, dove certe strutture di conoscenza "restano in attesa" e attivano se stesse quando sono stimolate da agenti o esperienze esterne.

Un'inferenza può collegare qualcosa presente in memoria e confermarlo; una nuova inferenza può contraddire certa vecchia informazione: se né contraddice né collega qualche altra informazione, allora tale conoscenza è nuova e aumenta quella esistente. Un primo tipo di inferenza è l'Inferenza di Specificazione: se abbiamo, per esempio, due frasi come "Searle attacca l'IA" e "Searle scrisse un articolo" il programma dovrebbe essere in grado di ipotizzare che "Searle attaccò l'IA in un articolo". Inferenze Causative

devono dare il seguente contributo: data la nuova frase “Searle attaccò Schank” io posso ipotizzare che “Schank era probabilmente arrabbiato con Searle”. Le Inferenze di Motivazione devono invece essere applicate al fine di indagare perché qualcuno fece qualcosa. Per esempio “Schank attaccò Searle” potrebbe produrre l’inferenza che “Schank voleva restituire un attacco avuto da Searle”. Un’ultima classe d’inferenze che menzioneremo è la Predizione dell’Azione. Se qualcuno afferma una motivazione, farà probabilmente qualcosa per conseguire uno stato. Per esempio se un input dice che “Searle voleva attaccare l’IA” il programma dovrebbe scoprire che “Searle potrebbe (forse) scrivere un articolo”.

Questi sono naturalmente solo esempi che non possono spiegare i meccanismi di inferenza nel dettaglio. Essi mostrano non solo i princìpi molto generali, ma anche che tali meccanismi lavorano tutti insieme. Infatti certe predizioni fatte da una “molecola di inferenza” sono rinforzate o confermate da altre.

Questa breve descrizione ha comunque già mostrato implicitamente le debolezze del programma. È infatti possibile generare un numero a caso di inferenze se i testi che si devono capire sono composti solo da due o tre frasi. Ma non sembra naturale che ciò accada nei confronti di un testo ordinario. Poiché il significato del paragrafo di un testo è parzialmente contenuto in strutture di conoscenza più alta, non è necessario che il lettore operi tutte le possibili relazioni tra tutte le singole frasi; egli tende piuttosto a integrare il loro significato in un complesso più grande.

Sempre nell’ambito del programma di Yale, troviamo il generatore di testi di Goldman chiamato BABEL. Esso fu costituito per tre compiti: primo, per parafrasare una frase; secondo, per verbalizzare le inferenze fatte; terzo, per fare semplici traduzioni in tedesco. Il suo compito principale è di selezionare parole che rappresentano il significato degli elementi di una dipendenza concettuale. Per questo ha una rete di discriminazione che sa distinguere tra sensi diversi di una parola.

Prendiamo per esempio un primitivo semantico come PTRANS: BABEL sa che deve scegliere fra parole come “andare” e “mettere”. Se vede che un attore si muove da A a B usando le sue gambe, molto probabilmente sceglierà un verbo come “camminare” per rappresentare PTRANS.

Paranoidi e psichiatri

ELIZA è il primo programma a trattare il problema della comunicazione fra uomo e macchina. Scritto da J. Weizenbaum, ELIZA incorpora metodi generali per analizzare parti di una frase, frasi e serie di frasi; il contesto non era all’origine parte del programma.

ELIZA può essere definito come un meccanismo che riesce a dirigere operazioni diverse con tecniche diverse senza aver nulla di suo da dire. Il programma è basato sulle tecniche di interazione tra dottore e paziente durante una seduta psichiatrica. ELIZA riesce a creare l'illusione di avere realmente una conversazione con lo psichiatra. Per alcuni l'illusione fu così forte che effettivamente chiesero di poter parlare alla macchina in privato. ELIZA generava in effetti risposte plausibili alle domande poste, secondo le attese che i pazienti avevano.

È importante notare che questo programma non ha una reale comprensione di quello che l'utente vuole esprimere: non costruisce alcun modello di conoscenza sul campo specifico, ma opera solo identificando e usando parole chiave predeterminate.

PARRY, invece, fu creato da Colby nel 1973. Simula anch'esso una seduta psichiatrica e consiste di due parti; la prima è data da un "analizzatore" sintattico che riconosce le frasi, la seconda da un "interpretatore" delle frasi trasmessegli dall'analizzatore. In breve: la simulazione di Colby rappresenta una donna che si sottopone a sedute psichiatriche e che crede di essere sgradita al padre, ma in realtà rifiuta di ammettere che è piuttosto lei a odiarlo. Quel che ne risulta è ovviamente una conversazione "sfasata" e per capire il significato reale delle frasi della paziente, Colby non può riferirsi alla logica propria del comune buon senso, ma deve piuttosto prendere in considerazione tutti quei meccanismi di difesa che finiscono per distorcere ogni ragionamento. Su quella che potrebbe essere definita come l'interpretazione letterale della conversazione si applicano cioè meccanismi quali proiezioni, riflessioni e neutralizzazioni, che altro non sono che derivati dei concetti propri della tradizione freudiana.

Si ribadisce quindi ancora una volta che per avere una reale comprensione, si deve prendere in considerazione non solo l'aspetto della conoscenza logica, ma anche quello più complesso delle reazioni irrazionali e inconsce.

Sistemi Esperti

I Sistemi Esperti si collegano alla problematica relativa al *decision making* (prendere decisioni). Per valutare i criteri di scelta, la pratica dell'IA analizza l'informazione conosciuta e i costi per ottenerla. I Sistemi Esperti sono in grado di fornire a un utente non eccessivamente specializzato una conoscenza vasta e precisa in un determinato settore.

Consideriamo alcuni esempi concreti: il sistema MYCIN è stato realizzato a Stanford. Funziona come un consulente che fornisce la sua competenza per identificare, per esempio, la causa di un'infezione, selezionando il farmaco appropriato. Una situazione tipica può essere la seguente: un paziente

presenta segni di una malattia. MYCIN assiste il medico nell'interpretare i dati e nel formulare le possibili diagnosi e le adeguate terapie.

Il sistema DENDRAL, programmato a Stanford, procede alla ricerca della serie di possibili strutture molecolari di costituenti di atomi conosciuti che formano una molecola non conosciuta. DENDRAL, che deriva il nome dall'Algoritmo di Dendral, consiste in tre parti: una prima definisce il problema; una seconda crea le strutture molecolari possibili; la terza ordina in termini di probabilità la serie di strutture possibili simulandone il comportamento. Il risultato di tale programma è stato quello di raggiungere un livello di qualità pari a quello di un "esperto umano" nel settore.

La teoria di Minsky

M. Minsky ha elaborato una teoria parziale del pensiero combinando una serie di concetti propri del versante psicologico e linguistico oltre che, naturalmente, di IA. Fasi fondamentali della sua ricerca sono costituite dalla Teoria del Sistema del Frame (*Frame-System Theory*, 1975), dalla Teoria della Società delle Menti (*Society of Mind Theory*, 1979), e dalla Teoria della Memoria delle K-Lines (*K-Lines Theory of Memory*, 1980).

Il *frame* è definito da Minsky come una struttura di dati che rappresentano una situazione stereotipa con tipi diversi di informazione. Ci sono informazioni relative all'"uso del frame" così come informazioni relative alle aspettative che un frame può creare. I frames si collegano fra loro a creare sistemi complessi di frames. Ogni volta che una nuova situazione si presenta, le aggregazioni dei frames vengono riorganizzate in modi diversi: il modello risulta quindi avere natura dinamica; Minsky stesso usa, del resto, la nozione di "trasformazione" applicata al suo sistema. La processualità del pensiero viene descritta da Minsky nei seguenti termini:

> Il pensare comincia sempre con immagini suggestive ma imprecise che vengono progressivamente sostituite da idee migliori.

Un successivo sviluppo della teoria è costituito dall'idea della Società delle Menti che prende in considerazione l'aspetto biologico-evolutivo del cervello. Minsky parla a questo punto di "un'evoluzione organica" e immagina il cervello come una serie di "agenti" collegati mediante canali di comunicazione e gerarchicamente ordinati. Ogni agente è collegato ad alcuni altri canali, K-Lines; in altre parole la mente funziona attraverso entità decentrate solo parzialmente intercomunicanti. Ogni processo di comunicazione fra agenti è strettamente controllato e mai arbitrario: ciò significa che ogni aggregazione di agenti è determinata da regole. La mente crea quindi associazioni diverse a seconda delle diversità degli stimoli.

La terza fase evolutiva della ricerca di Minsky è costituita dalla “teoria della memoria”: si basa sul principio per cui, ogni volta che si ha un’idea e si vuole mantenerla, ricordarla e usarla nuovamente, si crea una K-Line ovvero uno “stato mentale parziale”, una serie di agenti mentali che agiscono tutti insieme, simultaneamente, che somiglia molto a quello originario. La teoria della memoria si basa quindi sull’assunto che, se si impara qualcosa, la rappresentazione di quel qualcosa viene costruita, immagazzinata ed, eventualmente, riutilizzata. La funzione della memoria consiste quindi nel ricreare uno stato mentale: in breve, la memoria aiuta la mente ad affrontare una situazione nuova facendo ricorso a processi e immagini “precedentemente incamerati”: idea evidentemente inammissibile all’interno di una teoria basata sull’ipotesi della mente come entità singola e fondata invece sull’ipotesi delle menti decentrate, parzialmente autonome, ma intercomunicanti, i cui agenti possono essere considerati come costituenti di una società altamente specializzata. La dinamicità insita nella Società delle Menti annulla ogni possibile rischio di contraddittorietà in quanto questa può venire eliminata attraverso continui riassestamenti e aggiustamenti; questo è del resto il modo stesso in cui il ragionamento ha luogo quotidianamente nell’individuo.

Per concludere, riassumiamo alcuni principi su cui la teoria di Minsky si fonda: 1) la comunicazione fra gli agenti è localizzata; 2) la memoria riattiva stati mentali già presenti; 3) conoscenza e memoria hanno carattere procedurale; 4) la scelta e la selezione nelle aggregazioni mentali è contingente, relativa agli scopi prefissi e soggetta a mutamenti.

L’apprendimento

Un settore fondamentale nell’attuale ricerca di IA è costituito dalla ricerca sui problemi dell’apprendimento (*learning*). Si distinguono tre diversi tipi di apprendimento: 1) mediante esempi; 2) mediante specifiche istruzioni; 3) mediante attività concreta.

Nell’area specifica dell’IA, l’aspetto considerato come il più interessante riguarda la creazione di un sistema che possa imparare, rappresentare la conoscenza acquisita e infine cambiare tale rappresentazione della conoscenza, qualora essa non risulti più soddisfacente e adeguata.

Il tipo di apprendimento privilegiato dalla ricerca di IA è il primo, ovvero quello che riguarda l’uso di esempi o “precedenti” ai fini di risolvere un nuovo problema. Tale apprendimento è basato sul riconoscimento degli elementi rilevanti. L’esempio più indicativo è costituito dal programma di P. Winston che impara a riconoscere strutture (archi, nello specifico) data una serie di esempi e controesempi, attraverso il ricorso a procedimenti “analogici”. Il programma inizia interpretando una serie di figure che

rappresentano lo stesso elemento e dando una descrizione dell'oggetto in riferimento ai suoi elementi e alle relazioni intercorrenti, riconoscendo quello che è pertinente e ricorrente. Il programma identifica anche le differenze fra il nuovo oggetto da riconoscere e il vecchio modello, in fasi successive. Tale ipotesi di lavoro è confermata da una vasta serie di ricerche nel settore psicologico che confermano appunto che chi deve imparare può risultare confuso qualora gli vengano presentate le differenze tutte in uno stesso tempo; così il programma di Winston considera una lunga serie di possibili differenze creando diversi "modelli" ognuno dei quali isola e rappresenta una sola specifica e rilevante differenza. Le conseguenze teoriche di tale programma possono essere trasferite dal campo dell'osservazione di oggetti del mondo fisico ad altri settori. Winston ha quindi creato una teoria più generale di come sia possibile imparare attraverso "precedenti" (siano essi oggetti o esercizi), basandosi sulle relazioni causali rilevanti.

Il ragionamento analogico è quindi basato sul principio che se due situazioni sono simili per alcuni aspetti, devono esserlo anche per altri e si basa sul meccanismo di regole "se-allora" (if-then rules); se cioè alcune relazioni in due situazioni sono per certi aspetti analoghe, saranno analoghe anche per la parte restante. Come stabilire l'importanza di certe relazioni rispetto ad altre pur esistenti? Tale importanza può essere determinata "a priori" dall'insegnante ovvero essere addebitabile ai nessi di causalità esistenti fra le relazioni presentate. La conclusione che ne deriva è la seguente: l'analisi di situazioni complesse mediante processi analogici implica che, date cause simili in situazioni simili, si possono predire anche analoghi effetti, esiti e risultati.

Il programma AM invece ha come fine la scoperta di nuovi concetti matematici e delle rispettive intercorrenti relazioni; non impara però attraverso esempi, come faceva il programma di Winston, ma parte piuttosto da un corpus selezionato di conoscenze e, attraverso un procedimento di apprendimento graduale (passo per passo), AM seleziona certi concetti matematici che vuole valutare e ne produce i relativi esempi. AM possiede più di trenta procedimenti euristici: un possibile procedimento di generazione è di produrre semplicemente un esempio a partire dalla stessa definizione del concetto; un altro possibile procedimento usato da AM consiste nell'assumere esempi simili, già esistenti in campi diversi, e verificarne i risultati. Tale programma è basato quindi sul "procedere operazionale" piuttosto che sulla verifica o la deduzione derivata dal già esistente.

I programmi di Yale della Terza Generazione

Si è già affermata l'evidenza psicologica del fatto che le persone usano un

ampio corpo di conoscenze organizzate tratte da esperienze già avute nel passato al fine di interpretare nuove situazioni. Si è già presentato, a tale proposito, il frame di Minsky.

Il concetto di *script*, formulato dal gruppo di Yale, è una versione più semplice, volta a rappresentare una possibile sequenza di eventi che definisce una situazione stereotipa.

La conoscenza dei differenti scripts viene usata per interpretare e prender parte attiva a eventi che le persone affrontano quotidianamente. Gli scripts hanno meccanismi semplici, che funzionano in modo speciale in base alla conoscenza relativa a istituzioni sociali e norme comportamentali convenzionali. Gli scripts accorpano quindi tutta una specifica serie di "aspettative": il loro pratico utilizzo si ha nella generazione dei testi linguistici poiché essi permettono di evitare di creare messaggi ridondanti in quanto facenti già parte dello script e quindi già conosciuti da parte del ricevente. Gli scripts sono inoltre relativi alle diverse prospettive adottate: lo stesso script del ristorante presenta infatti diverse interpretazioni a seconda che si adotti il punto di vista del cameriere o del cliente.

Oltre ai tipi "situazionali" e "comportamentali" di scripts, esistono anche quelli "personali" ovvero riguardanti una possibile teoria di tratti di personalità, come teorizzato da Carbonell. Esistono inoltre ulteriori scripts definibili come "strumentali" in quanto collegati alla progettazione e al conseguimento di scopi specifici.

Il primo sistema per la comprensione testuale elaborato a Yale sulla base di scripts è stato SAM (*Script Applier Mechanism*); SAM comprende solo una limitata serie di situazioni stereotipe quali il viaggiare o il mangiare. Una più elaborata versione di SAM è in grado di leggere articoli di giornale riguardanti incidenti stradali e visite ufficiali di capi di stato.

Per comprendere un testo linguistico però non bastano semplici scripts, ma occorrono inferenze e deduzioni operate in base all'informazione linguisticamente esplicitata ai fini di recuperare quella implicita, tuttavia presente. Per comprendere certe narrazioni il lettore deve così comprendere lo scopo finale di tali storie e confrontare le varie azioni presentate come stadi intermedi per l'ottenimento dello scopo finale. Ogni frase di un testo viene pertanto interpretata in relazione allo scopo o scopi presentati. Per esempio, la frase "Searle ha ricevuto la macchina da scrivere" può contenere come scopo implicito strumentale il fatto che Searle voglia scrivere un articolo contro l'Intelligenza Artificiale. Secondo la terminologia usata da Schank, ogni scopo è corredato da una serie di *plans* ovvero strategie, per il conseguimento del medesimo.

Indicativo di questa ultima prospettiva di ricerca è il programma PAM (*Plan Applier Mechanism*) scritto da R. Wilensky. Il suo scopo principale è quello di trovare spiegazioni per dati eventi. A tal fine dispone di due diverse

strategie: in primo luogo controlla se un evento è compreso entro un plan, predetto inizialmente da PAM; in seguito, se ciò non riesce, cerca di operare il numero più alto di processi di inferenziazione dall'input e controlla se queste inferenze contengano plans nell'ambito dei quali gli eventi considerati possano essere giustificati. In tal modo, PAM risulta capace di ricostruire e completare quell'informazione che non è presente esplicitamente nel testo, così come questo viene linguisticamente formulato, ma che è pur sempre indispensabile alla comprensione del medesimo. I più moderni sistemi possono integrare diversi modelli di conoscenza organizzata sia in scripts che in plans.

Si conclude con questa rassegna delle ipotesi teoriche e dei programmi di IA che risultano essere i più indicativi e rilevanti ai fini di una miglior comprensione della terminologia e delle tematiche discusse nel corso del "Grande Dibattito". Avremmo, come si è già precisato, potuto estendere tale rassegna a ulteriori settori altrettanto importanti di IA; dato però il carattere "funzionale" di tale premessa, l'accento è caduto su quegli argomenti che ci sono apparsi come i più collegati alle tematiche specifiche che seguono nella discussione "intorno a Searle". A tal proposito non abbiamo voluto presentare la nostra personale posizione teorica per permettere al lettore di confrontare e scegliere liberamente fra le diverse posizioni.

Ricordiamo che lo stile della traduzione vuole rispecchiare fedelmente quello variato e spesso disomogeneo dei testi originari. Si sottolinea anche che alcune delle "risposte" sono nate da registrazioni di nastri (hanno quindi caratteristiche orali) e sono poi state trascritte: si deve a questo fatto lo stile colloquiale e proprio di un procedimento discorsivo (interruzioni, riprese, periodare complesso). Vista comunque la ricchezza di sfumature, toni e registri, si è preferito talvolta riprodurre tale apparente contraddittorietà senza tentarne riaggiustamenti.

Bibliografia

Boden, M.A. (1977) *Artificial Intelligence and natural Man*, Hassocks: Harverster Press.

Carbonell, J.G. (1981) *Subjective Understanding. Computer Models of Belief Systems,* Ann arbor: UMI Research Press.

Charniak, E., Riesbeck, C.K., McDermott, D. (1980) *Artificial Intelligence Programming*, Hillsdale: Erlbaum.

De Michelis, G. (1983) Informatica e trasformazione del sapere, in *Skill*, I, Milano, pp. 17-29.

Dennett, D. (1969) *Content and Consciousness*, New York: Humanities Press.

Dreyfus, H. (1972) *What Computers Can't Do*, New York: Harper and Row.

Feigenbaum, E.A., Feldman, J. (1963) *Computers and Thought*, New York: McGraw-Hill.

Haugeland, J. (a cura di) (1981) *Mind Design, Philosophy, Psychology, Artificial Intelligence*, Cambridge (Ma): The MIT Press.

Hofstadter, D.R. (1980) *Gödel, Escher, Bach. An Eternal Golden Braid*, New York: Vintage Books.

Lenat, D. (1977) Automated Theory Formation in Mathematics, in *Proc. IJCAI*, 5.

Longuet-Higgins, H.C. (1981) Artificial Intelligence. A New Theoretical Psychology, in *Cognition*, 10, pp. 197-200.

McCarthy, J., Hayes, P. (1969) Some Philosophical Problems from the Standpoint of Artificial Intelligence, in *Machine Intelligence*, Vol. 4, New York: Elsevier.

McCorduck, P. (1979) *Machines Who Think. A Personal Inquiry into the History and Prospects of Artificial Intelligence*, San Francisco: Freeman.

Minsky, M. (1963) Steps Toward Artificial Intelligence, in Feigenbaum, E.A., Feldman, J. (a cura di), *Computers and thought*, New York: McGraw-Hill, pp. 406-450.

Minsky, M. (1975) A Framework for Representing Knowledge, in Winston, P.H., *The Psychology of Computer Vision*, New York: McGraw-Hill, pp. 211-277.

Minsky, M. (fall 1982) Why People Think Computers Can't, in *The AI Magazine*, pp. 3-15.

Newell, A. (1973) Artificial Intelligence and the Concept of Mind, in Schank, R.C., Colby, M.K., *Computer Models of Thought and Language*, San Francisco: Freeman.

Nilsson, N. (1981) *Principles of Artificial Intelligence*, Palo Alto, Tioga.

Rich, E. (1983) *Artificial Intelligence*, New York: McGraw-Hill.

Schank, R.C. (1980) Language and Memory, in *Cognitive Science*, 4, pp. 243-284.

Schank, R.C., Abelson, R.P. (1977) *Scripts, Plans, Goals and Understanding*, Hillsdale: Erlbaum.

Simon, H.A. (1969) *The Sciences of the Artificial*, Cambridge (Ma): The MIT Press.

Tonfoni, G. (1982) *A Frame-System Theory of Text*, Working Paper, AI Lab., MIT, Cambridge, Mass.

Tonfoni, G. (marzo, 1983) L'Intelligenza Artificiale: a colloquio con M. Minsky, in *Alfabeta*.

Wilensky, R. (1983) *Planning and Understanding. A Computational Approach to Human Reasoning*, Reading: Addison-Wesley.

Winston, P.H. (1977) *Artificial Intelligence*, Reading: Addison-Wesley.

MENTI, CERVELLI E PROGRAMMI

John R. Searle
Dipartimento di Filosofia, Univ. di California, Berkeley

Questo articolo può essere considerato come un tentativo d'esplorazione delle conseguenze derivanti da due affermazioni: (1) l'intenzionalità negli esseri umani (e animali) è un prodotto di caratteri causali inerenti il cervello. Io penso che questo sia un fatto empirico riguardante le relazioni causali effettive tra processi mentali e cervello. Significa semplicemente che certi processi del cervello sono sufficienti per l'intenzionalità; (2) instanziare un programma per il computer non è mai di per sé una condizione sufficiente d'intenzionalità. L'assunto principale di questo articolo è diretto a stabilire questa asserzione. Esso viene realizzato nel mostrare come un agente umano potrebbe instanziare un programma e non avere tuttavia l'intenzionalità relativa.

Queste due affermazioni hanno le seguenti conseguenze: (3) la spiegazione di come il cervello produce intenzionalità non può essere che lo fa semplicemente con l'instanziare un programma per un computer: questa è una conseguenza strettamente logica di 1 e 2; (4) ogni meccanismo capace di produrre intenzionalità deve avere poteri causali uguali a quelli del cervello: si considera che questa sia una semplice conseguenza di 1; (5) ogni tentativo di creare intenzionalità artificialmente proprio dell'ipotesi forte dell'Intelligenza Artificiale non potrebbe risultare semplicemente dall'organizzare programmi, ma dovrebbe riprodurre piuttosto i nessi causali propri del cervello umano: questo segue da 2 e 4.

"Potrebbe una macchina pensare?" In base all'assunto qui presentato solo le macchine potrebbero pensare, e solo tipi di macchine molto speciali, precisamente cervelli e macchine con nessi causali interni che sono equivalenti a quelli dei cervelli. E questo è il motivo per cui l'ipotesi "forte" dell'Intelligenza Artificiale ha poco da dirci intorno al pensare, poiché non riguarda le macchine, ma piuttosto i programmi, e nessun programma è di per sé capace di pensare.

Parole chiave: intelligenza artificiale; cervello; intenzionalità; mente.

Quale rilevanza psicologica e filosofica dovremmo dare agli sforzi recenti operati nella simulazione da parte del computer delle capacità cognitive umane? Nel rispondere a questo quesito trovo utile fare una distinzione tra ipotesi di IA "forte" e IA "debole" o "cauta". Secondo la IA debole il principale valore del computer nello studio della mente è di darci uno strumento molto potente. Per esempio, ci dà la possibilità di formulare ed esaminare ipotesi in un modo più rigoroso e preciso. Invece, secondo la IA forte, il computer non è semplicemente uno strumento nello studio della mente; piuttosto, il computer appropriatamente programmato è *realmente* una mente, nel senso che i computer, cui sono stati dati i programmi giusti, *capiscono* e hanno altri stati cognitivi. Nella IA forte, per il fatto che il

computer programmato ha stati cognitivi, i programmi non sono semplici strumenti che ci rendono possibile considerare spiegazioni psicologiche: piuttosto i programmi costituiscono di per sé le spiegazioni.

Non ho obiezioni da porre alle dichiarazioni della IA debole, almeno in questo articolo. Le mie obiezioni qui saranno dirette alle dichiarazioni che ho definito di IA forte, e specificamente alla dichiarazione che il computer, appropriatamente programmato, possiede letteralmente stati cognitivi, e che i programmi con ciò spiegano le capacità umane di conoscere. Quando mi riferisco alla IA, ho in mente la versione forte, come espressa da queste due dichiarazioni.

Intendo prendere in considerazione il lavoro di Roger Schank e dei suoi colleghi di Yale (Schank e Abelson, 1977), perché ho più consuetudine con esso che con qualunque altra tesi simile, e perché esso fornisce un esempio molto chiaro del genere di lavoro che desidero esaminare. Ma nulla di ciò che segue dipende nei dettagli dai programmi di Schank. Gli stessi argomenti si applicherebbero a SHRDLU di Winograd (Winograd, 1973), ELIZA di Weizenbaum (Weizenbaum, 1965) e, in pratica, a qualunque simulazione da parte di una macchina di Turing dei fenomeni mentali umani.

Molto brevemente e lasciando da parte i vari dettagli, si può descrivere il programma di Schank come segue: lo scopo del programma è di simulare l'abilità umana nel comprendere i racconti. È caratteristico della capacità di comprendere i racconti, propria degli esseri umani, il fatto che essi possano rispondere a domande sul racconto, anche se le informazioni che danno non sono mai state esplicitamente formulate nel racconto. Così, per esempio, supponiamo ci venga presentata la seguente storia: "Un uomo entrò in un ristorante e ordinò un hamburger. Quando l'hamburger arrivò, era tutto bruciato e l'uomo si precipitò fuori dal ristorante, furioso, senza pagare o lasciare la mancia". Ora, se vi si chiede: "L'uomo ha mangiato l'hamburger?" probabilmente risponderete: "No". Similmente, se vi si presenta la seguente storia: "Un uomo andò in un ristorante e ordinò un hamburger; quando l'hamburger gli fu portato, ne fu molto soddisfatto, e lasciando il ristorante diede alla cameriera una bella mancia prima di pagare il conto". E se vi si chiede: "L'uomo ha mangiato l'hamburger?" probabilmente risponderete: "Sì, l'ha mangiato". Ora le macchine di Schank possono ugualmente rispondere alle domande sui ristoranti in questo modo. Per fare questo, esse hanno una "rappresentazione" del genere di informazione che gli esseri umani hanno sui ristoranti, che rende loro possibile rispondere a domande come quelle sopra, una volta dato questo tipo di storie. Quando alla macchina si è data la storia e poi fatta la domanda, la macchina emetterà risposte del tipo che ci aspetteremmo da esseri umani qualora si raccontassero loro storie simili. I fautori di IA forte dichiarano che in questa sequenza di domande e risposte, la

macchina non solo simula un'abilità umana, ma anche che:

1. la macchina capisce letteralmente la storia e fornisce le risposte alle domande;
2. ciò che la macchina e il suo programma fanno, spiega l'abilità umana a comprendere la storia e a rispondere alle domande su essa.[6]

Entrambe le tesi mi sembrano totalmente non comprovate dal lavoro di Schank, come tenterò di mostrare in ciò che segue.

Un modo per esaminare qualunque teoria della mente è quello di chiedersi che cosa avverrebbe se la mia mente funzionasse in base a quei principi che la teoria stabilisce come comuni a tutte le menti. Applichiamo questa indagine al programma Schank con il seguente *Gedankenexperiment.* Supponiamo che io sia chiuso dentro una stanza e che mi si dia una serie di fogli scritti in cinese. Supponiamo inoltre (come infatti è il caso mio) che non conosca il cinese, né scritto né parlato, e che non sia nemmeno fiducioso di poter riconoscere uno scritto cinese in quanto tale, distinguendolo magari dal giapponese o da scarabocchi senza senso. Per me la scrittura cinese è proprio come tanti scarabocchi senza senso. Ora supponiamo ancora che, dopo questo primo esperimento, mi si dia un secondo pacco di fogli, sempre scritto in cinese, insieme con una serie di regole per mettere in relazione il secondo plico con il primo. Le regole sono in inglese e io capisco queste regole come qualunque altro inglese di madrelingua. Esse mi rendono possibile mettere in relazione una serie di simboli formali con un'altra serie di simboli formali (e tutto quello che formale significa qui, è che posso identificare i simboli interamente attraverso le loro forme). Ora supponiamo anche che mi si dia una terza serie di simboli cinesi con le relative istruzioni, sempre in inglese, che mi rendano possibile correlare elementi di questo terzo pacco con i primi due, e che queste regole mi istruiscano su come riprodurre certi simboli cinesi con certi tipi di forme datemi nel terzo plico. A mia insaputa, le persone che mi danno tutti questi simboli chiamano il primo pacco di fogli "uno scritto", chiamano il secondo "una storia", e il terzo "quesiti". Inoltre chiamano i simboli che rendo loro in risposta al terzo plico "risposte alle domande", e la serie di regole in inglese che mi hanno dato la chiamano "il programma". Ora, proprio per complicare un po' la storia, immaginiamo che queste persone mi diano pure delle storie in inglese, che mi facciano domande in inglese su queste storie, e io renda loro le risposte in inglese. Supponiamo anche che io diventi così bravo nel seguire le istruzioni per manipolare i simboli cinesi e che i programmatori diventino così bravi nello scrivere i programmi che dal punto di vista esterno — cioè dal punto di vista di qualcuno al di fuori della stanza nella quale sono chiuso — le mie risposte alle domande assolutamente non si distinguono da quelle di cinesi madrelingua. Nessuno che guardi bene alle mie risposte può dire che io non parli una parola di cinese. Supponiamo pure che le mie risposte alle domande in inglese siano, come senza dubbio

sarebbero, non distinguibili da quelle di altri inglesi nativi, per la semplice ragione che io sono di madrelingua inglese. Dal punto di vista esterno — dal punto di vista di qualcuno che legge le mie risposte — le risposte alle domande in cinese e a quelle in inglese sono egualmente buone. Ma nel caso del cinese, diversamente da quello dell'inglese, produco le risposte col manipolare simboli formali non interpretati. Per quanto riguarda il cinese, mi comporto semplicemente come un computer: eseguo operazioni calcolabili su elementi formalmente specificati. Per il caso del cinese, io sono semplicemente una istanziazione di un programma del computer.

Ora le dichiarazioni fatte dalla IA forte sono che il computer programmato capisce le storie e che il programma, in qualche modo, spiega il capire umano. Siamo in grado di esaminare queste tesi alla luce del nostro esperimento.

1) Per quanto riguarda la prima asserzione, mi sembra del tutto ovvio, nell'esempio, che non capisco neppure una parola delle storie cinesi. Ho immissioni ed emissioni di dati che non si distinguono da quelle di un cinese nativo e posso avere qualunque programma formale, ma continuo a non capire nulla. Per le stesse ragioni, il computer di Schank non capisce nulla di storie, sia in cinese che in inglese che in qualsiasi altra lingua, poiché nel caso cinese il computer sono io, e nei casi in cui il computer non sia io, il computer non ha niente di più di quello che ho io nel caso in cui non capisco nulla.

2) Per quanto riguarda la seconda dichiarazione, secondo cui il programma *spiega* il capire umano, possiamo costatare che il computer e il suo programma non forniscono sufficienti condizioni di comprensione poiché computer e programma funzionano eppure non c'è comprensione. Forniscono almeno una condizione necessaria o un contributo significativo al comprendere? Uno degli assunti sostenuti dai fautori della IA forte è che, quando capisco una storia in inglese, quel che faccio è esattamente la stessa cosa, sviluppata forse in misura maggiore, che facevo nel manipolare i simboli cinesi. È semplicemente il grado di manipolazione formale di simboli che distingue il caso del testo inglese, in cui io capisco, dal caso del cinese, dove non capisco. Pur non avendo dimostrato che questo assunto è falso, certo esso appare incredibile nell'esempio. La plausibilità che può avere l'assunto deriva dalla supposizione che possiamo costruire un programma che avrà le stesse immissioni ed emissioni di dati (*input* e *output*) dei nativi di madrelingua e, in aggiunta, assumiamo che i parlanti nativi hanno un certo livello di descrizione dove essi stessi sono instanze di un programma. Sulla base di queste due tesi sosteniamo che, anche se il programma di Schank non è una completa giustificazione del processo di comprensione, può nondimeno rappresentarne una parte. Suppongo che questa sia una possibilità empirica: finora comunque non è stata data la minima dimostrazione per credere che sia vera, dal momento che ciò che è suggerito — ma certamente non dimostrato — dall'esempio, è che un programma di computer è semplicemente

irrilevante per quanto riguarda la mia comprensione della storia. Nel caso della storia in cinese, ho tutto ciò che l'Intelligenza Artificiale può mettere in me per mezzo di un programma, ma io non capisco nulla; nel caso della storia in inglese capisco tutto e non c'è alcuna ragione al mondo per supporre che il mio grado di comprensione ha qualcosa a che vedere con programmi per il computer, cioè con operazioni di calcolo su elementi specificati in modo puramente formale. Finché il programma è definito in termini di operazioni computazionali basati su elementi definiti solo formalmente, quello che l'esempio suggerisce è che questi, di per sé, non hanno alcuna connessione interessante con il comprendere in sé e per sé. Sono certamente condizioni non sufficienti e non c'è la minima ragione per supporre che siano condizioni necessarie o perfino che diano un minimo contributo significativo al comprendere. Si noti che la forza di tale assunto non è semplicemente che macchine diverse possano avere lo stesso input o output mentre operano con principi formali diversi: non è assolutamente questo il punto. Piuttosto, qualunque principio puramente formale si metta nel computer, non sarà sufficiente per la comprensione, poiché un essere umano potrà seguire i principi formali senza capire nulla. Non si è presentata alcuna ragione per supporre che tali principi siano necessari o perfino minimamente utili, poiché non si è presentata alcuna ragione per supporre che, quando capisco l'inglese, debba per questo operare con un programma formale. Allora, che cosa ho nel caso delle frasi in inglese che non ho nel caso delle frasi in cinese? La risposta ovvia è che conosco che cosa significano le prime, mentre non ho la più pallida idea di quello che significano le seconde. Ma in che cosa consiste questa proprietà e perché non potremmo attribuirla a una macchina, qualunque cosa essa sia? Ritornerò su questa questione più avanti, ma prima voglio continuare con l'esempio. Ho avuto occasione di presentare questo esempio a diversi esperti di Intelligenza Artificiale e, cosa interessante, sembra che questi non siano d'accordo fra loro su quale possa essere l'appropriata risposta. Ricevo una sorprendente varietà di risposte, e in seguito considererò le più comuni di queste (specificate via via con le rispettive origini geografiche).

Ma prima voglio chiarire alcuni equivoci comuni riguardanti il *comprendere*:[7] in molte di queste discussioni si trovano numerose considerazioni assai semplici, ma elegantemente formulate intorno alla parola *comprendere.* I miei critici mostrano che ci sono molti gradi diversi di comprensione; che il *comprendere* non è un semplice predicato a due argomenti; che ci sono perfino diversi generi e livelli di comprensione, e spesso la legge del *terzo escluso* non si applica nemmeno in maniera diretta ad affermazioni del tipo "x comprende y"; che in molti casi è una questione di decisione e non un semplice dato di fatto che x capisce y, e così via. A questo proposito, voglio dire: certo, certo. Ma ciò non ha nulla a che fare con i punti

in discussione. Ci sono chiari casi in cui la comprensione, il *capire*, si applica letteralmente e altri in cui non si applica e questi due tipi di casi sono tutto quello di cui ho bisogno per questo assunto. Capisco le storie in inglese; a un grado minore posso capire le storie in francese; a un grado ancora minore le storie in tedesco; ma in cinese, niente del tutto. La mia auto e la mia macchina calcolatrice, al contrario, non capiscono nulla: non è di quello che si occupano.

Spesso attribuiamo il *comprendere* e altri predicati cognitivi, per metafora e analogia, alle auto, alle macchine calcolatrici, e così via, ma con tali attribuzioni non proviamo nulla. Diciamo: "La porta sa quando deve aprirsi grazie alle sue cellule fotoelettriche", "La macchina calcolatrice sa come fare addizioni e sottrazioni, ma non divisioni" e "Il termostato percepisce ciò che accade nella temperatura". La ragione per cui facciamo queste attribuzioni è molto interessante e ha a che vedere col fatto che noi estendiamo la nostra intenzionalità[8] ai mezzi meccanici; i nostri strumenti sono estensioni dei nostri scopi, e così troviamo naturale fare loro attribuzioni metaforiche di intenzionalità; ma penso che tali esempi non sfondino alcuna parete reale. Il senso in cui una porta automatica "capisce le istruzioni" dalla sua cellula fotoelettrica, non è affatto il senso in cui io comprendo l'inglese. Se si suppone che il senso in cui i computer programmati di Schank capiscono le storie sia il senso metaforico col quale la porta capisce, e non il senso in cui io comprendo l'inglese, la questione non meriterebbe una discussione. Ma Newell e Simon (1963) dichiarano che il genere di cognizione che ha il computer è esattamente lo stesso di quello degli esseri umani. Mi piace la franca immediatezza di questa affermazione ed è quella che prenderò in considerazione. Intendo mostrare che nel senso letterale il computer programmato comprende ciò che l'auto e la calcolatrice comprendono: cioè, esattamente, nulla. La comprensione del computer non è affatto parziale o incompleta (come la mia comprensione del tedesco): è zero.

E ora veniamo alle risposte.

La replica del sistema (Berkeley)

> Mentre è vero che l'individuo chiuso nella stanza non capisce la storia, sta di fatto che egli è semplicemente parte di un intero sistema, e il sistema effettivamente *comprende* la storia. La persona ha di fronte a sé un ampio registro in cui sono scritte le regole, ha fogli di carta per appunti e matite per fare calcoli, ha una quantità di dati riguardanti la serie di simboli cinesi. Ora, il comprendere non si attribuisce al solo individuo: si attribuisce piuttosto a questo intero sistema di cui tale individuo è parte.

La mia obiezione alla risposta della teoria dei sistemi è del tutto semplice: l'individuo interiorizza tutti questi elementi del sistema. Egli memorizza le

regole nel registro e i dati relativi ai simboli cinesi, e fa a mente tutti i calcoli. L'individuo incorpora l'intero sistema: non c'è proprio nulla del sistema che egli non comprenda. Possiamo perfino sbarazzarci della stanza e supporre che egli lavori fuori. Ciononostante egli non capisce nulla del cinese, e, *a fortiori*, neppure il sistema, perché non c'è nulla nel sistema che non sia in lui. Se lui non capisce, non c'è alcun modo per cui il sistema possa capire, poiché esso è proprio una sua parte.

Effettivamente, mi sento in qualche modo imbarazzato persino a dare questa risposta alla teoria dei sistemi, poiché la teoria mi sembra poco plausibile fin dall'inizio. L'idea è che, mentre una persona non comprende il cinese, in qualche modo la *combinazione* di quella persona e di pezzi di carta potrebbero, insieme, capire il cinese: non è facile per me immaginare che qualcuno (che non fosse nella stretta di un'ideologia) potrebbe trovare l'idea in qualche modo plausibile. E penso che molta gente che si è impegnata nell'ideologia di IA forte, alla fine sarà propensa a dire qualcosa di molto simile: perciò andiamo oltre.

Secondo una visione di questo tipo, mentre l'uomo nell'esempio non capisce il cinese come un cinese madrelingua (perché, per esempio, non sa che la storia si riferisce a ristoranti e a hamburgers, ecc.) tuttavia "l'uomo come sistema di manipolazione di simboli" *comprende* realmente il cinese. Il sottosistema dell'uomo che è il sistema di manipolazione di simboli per il cinese, non dovrebbe essere confuso col sottosistema per l'inglese.

Così ci sono realmente due sottosistemi nell'uomo: uno comprende l'inglese, l'altro il cinese, ed "è vero che i due sistemi hanno poco a che fare l'uno con l'altro". Ma, voglio rispondere, non solo essi hanno poco in comune, essi non sono nemmeno lontanamente simili. Il sottosistema che capisce l'inglese (supponendo per un momento che ci possiamo permettere di parlare in questo gergo di "sottosistemi") sa che le storie vertono su ristoranti e hamburgers, sa che gli si fanno domande su ristoranti e risponde alle domande meglio che può operando varie deduzioni dal contenuto della storia, e così via. Ma il sistema cinese non conosce nulla di questo. Mentre il sottosistema inglese sa che il termine *hamburgers* si riferisce ad hamburgers reali, il sottosistema cinese sa soltanto che *squiggle squiggle* è seguito da *squoggle squoggle*. Tutto quello che sa è che vari simboli formali vengono introdotti da una parte, e manipolati secondo regole scritte in inglese, e che, dall'altra parte, altri simboli salteranno fuori. Tutto il succo dell'esempio originale era di sostenere che tale manipolazione di simboli di per sé non potrebbe essere sufficiente per capire il cinese in alcun senso letterale, perché l'uomo potrebbe scrivere *squoggle squoggle* dopo *squiggle squiggle* senza capire nulla di cinese. E non ci si aiuta postulando sottosistemi nell'uomo, perché i sottosistemi non sono in condizioni migliori di quanto lo sia l'uomo: essi ancora non hanno nulla di nemmeno lontanamente simile a quello che ha

l’uomo (o il sottosistema) di madrelingua inglese. Infatti, nel caso descritto, il sottosistema cinese è semplicemente una parte del sottosistema inglese, una parte che è occupata in una manipolazione di simboli senza senso secondo regole in inglese. Chiediamoci che cosa motivi il sistema come risposta: cioè quali ragioni *indipendenti* si suppone ci siano per dire che l’agente deve avere un sottosistema in sé che letteralmente capisce le storie in cinese? Per quanto posso dire io, le uniche ragioni sono che nell’esempio ho lo stesso input e output dei nativi cinesi e un programma che va dall’uno all’altro. Ma il vero scopo degli esempi è stato quello di mostrare che ciò non potrebbe essere sufficiente per capire, nel senso in cui capisco le storie in inglese, perché una persona, e quindi il complesso di sistemi che concorrono a costituire una persona, potrebbe avere la giusta combinazione di input, output e programma, e tuttavia non capire nulla nel preciso senso in cui io capisco l’inglese. La sola motivazione per dire che ci deve essere un sottosistema in me che comprende il cinese è che io ho un programma e posso superare la prova di Turing; che cioè io posso ingannare i madrelingua cinesi. Ma precisamente uno dei punti in questione è l’adeguatezza della prova di Turing. L’esempio mostra che ci potrebbero essere due “sistemi” che superano entrambi la prova di Turing, dei quali però uno solo *comprende*; e contro questo punto non è una prova dire che, dal momento che entrambi superano la prova di Turing, devono entrambi capire, poiché questa asserzione non mette in discussione la tesi che il sistema in me che *capisce* l’inglese ha molto più del sistema che semplicemente *agisce* nel cinese. In breve: la risposta del sistema evade la questione ripetendo, senza prova valida, che il sistema *deve* capire il cinese.

Inoltre, il sistema come risposta sembrerebbe portare a conseguenze che sono assurde in sé e per sé. Se dobbiamo concludere che in me ci deve essere *cognizione* sulla base che ho un certo tipo di input e output e un programma, allora appare probabile che ogni genere di sottosistema non cognitivo è destinato a diventare cognitivo. Per esempio, c’è un livello di descrizione nel quale il mio stomaco esegue trattamenti di informazione e instanzia un numero qualunque di programmi per computer, ma io credo che noi non vogliamo per questo dire che esso abbia alcun tipo di capacità di comprensione (cfr. Pylyshyn, *Computation and Cognition*, BBS 3(I), 1980). Ma, se accettiamo il sistema come risposta, allora è difficile evitare di dire che stomaco, cuore, fegato e così via sono tutti sottosistemi dotati di comprensione, poiché non c’è alcun criterio fondato per distinguere l’argomento che il sottosistema cinese capisce da quello che lo stomaco capisce. Non costituisce, a proposito, una risposta a questo punto dire che il sistema cinese ha l’informazione come input e output e lo stomaco ha il cibo e i residui del cibo come input e output, poiché dal punto di vista dell’agente, dal mio punto di vista, non c’è informazione né nel cibo né nel cinese; il sistema cinese è costituito proprio da tanti piccoli elementi senza significato.

L’informazione, nel caso cinese, è solo negli occhi del programmatore e degli interpreti, e non c’è nulla che impedisca loro di trattare l’input e l’output dei miei organi digestivi come informazione, se lo desiderano.

Quest’ultimo punto si riferisce ad alcuni problemi indipendenti nell’ipotesi della IA *forte* e vale la pena per un momento di soffermarvisi. Se l’IA forte deve essere una branca della psicologia, deve poter distinguere quei sistemi che sono genuinamente mentali da quelli che non lo sono. Si devono poter distinguere i principi in base ai quali la mente opera da quelli in base ai quali operano i sistemi non mentali: altrimenti non si offrirà alcuna spiegazione di ciò che è specificamente mentale intorno al mentale. E la distinzione mentale/non mentale non può essere evidente solo all’occhio dell’osservatore, ma deve essere intrinseca ai sistemi: altrimenti il risultato sarebbe che ogni osservatore potrebbe trattare gli individui umani come non mentali e, per esempio, gli uragani come mentali, in modo puramente arbitrario. Ma molto spesso, nella letteratura di IA la distinzione viene a essere confusa in modi che a lungo andare si dimostrerebbero disastrosi rispetto alle tesi che l’IA è una ricerca cognitiva. McCarthy, per esempio, scrive: “Le macchine semplici come i termostati, si può dire abbiano delle opinioni, e avere delle opinioni sembra essere una caratteristica di moltissime macchine capaci di eseguire risoluzioni di problemi” (McCarthy, 1979).

Chiunque pensi che l’IA forte abbia la possibilità di essere una teoria della mente, dovrebbe riflettere sulle implicazioni di questa osservazione. Ci si chiede di accettare come una scoperta di IA forte che la barra di metallo usata per regolare la temperatura ha idee esattamente nello stesso senso in cui noi e i nostri figli abbiamo idee, e inoltre che “la maggior parte” delle altre macchine nella stanza — il telefono, il registratore, la macchina calcolatrice, l’interruttore elettrico — hanno pure idee in questo senso letterale. Non è scopo di questo articolo discutere il punto di McCarthy, per cui sosterrò semplicemente ciò che segue senza argomentarlo. Lo studio della mente comincia con fatti quale quello che gli esseri umani hanno idee, mentre termostati, telefoni e calcolatori non le hanno. Se si costruisce una teoria che nega questo punto, si è prodotto un controesempio alla teoria e la teoria è falsa. Si ha così l’impressione che le persone di IA che scrivono questo genere di cose, pensino che possano farlo perché non lo prendono veramente sul serio, e pensano che nessun altro lo farà. Io propongo, per un momento almeno, di prenderlo sul serio.

Pensa bene per un minuto a ciò che sarebbe necessario per stabilire che quella barra di metallo sulla parete ha reali convinzioni, convinzioni con orientamento e contenuto proposizionale, e condizioni di soddisfazione: convinzioni che abbiano la possibilità di essere o forti o deboli; convinzioni nervose, ansiose o sicure; convinzioni dogmatiche, razionali o superstiziose; fedi cieche o riflessioni esitanti: ogni genere di convinzioni. Il termostato non

è un candidato, né lo sono lo stomaco, il fegato, la calcolatrice o il telefono. Comunque, poiché stiamo prendendo l'idea sul serio, si noti che la sua verità sarebbe fatale alla proclamazione dell'IA forte come una scienza della mente. Perché ora la mente è ovunque. Quello che volevamo conoscere è che cosa distingue la mente dai termostati e dai fegati. E se McCarthy avesse ragione, l'IA forte non avrebbe alcuna speranza di dircelo.

La replica del robot (Yale)

> Si supponga che abbiamo scritto un genere diverso di programma da quello di Schank. Si supponga che mettiamo un computer dentro un robot e che questo computer non solo riceva simboli formali come input ed emetta simboli formali come output, ma faccia effettivamente funzionare il robot in modo tale che esso si comporti in modo molto simile al percepire, camminare, muoversi intorno, piantare chiodi, mangiare, bere e qualunque altra cosa gli piaccia. Il robot avrebbe, per esempio, una telecamera incorporata che gli permetterebbe di "vedere". Avrebbe braccia e gambe che lo metterebbero in grado di "agire", e tutto questo sarebbe controllato dal suo cervello computerizzato. Un tale robot, diversamente dal computer di Schank, avrebbe una genuina capacità di comprensione e altri stati mentali.

La prima cosa da notare sulla replica del robot è che esso concede tacitamente che la cognizione non è solamente una questione di manipolazione di simboli, poiché essa aggiunge un insieme di relazioni causali inerenti al mondo esterno (cfr. Fodor, *Methodological Solipsism*, BBS 3 (I), 1980). Ma la risposta alla replica del robot è che l'idea di tali capacità "percettive" e "motorie" non aggiunge nulla al programma originale di Schank in sostituzione della comprensione, in particolare, o dell'intenzionalità, in generale. Per convincersene, si noti che lo stesso esperimento di pensiero si applica al caso del robot.

Si supponga che invece del computer dentro il robot, si metta me dentro la stanza e, come nel caso originale cinese, mi si diano simboli cinesi con istruzioni in inglese per unire simboli cinesi a simboli cinesi, emettendo in risposta simboli cinesi. Si supponga che, a mia insaputa, alcuni dei simboli cinesi che mi sono dati provengano da un apparecchio televisivo inserito nel robot e che altri simboli cinesi che sto distribuendo servano a far sì che i motori all'interno del robot muovano le gambe o le braccia del robot. È importante sottolineare che tutto quello che faccio è manipolare simboli formali:, non conosco nessuno di questi altri fatti. Ricevo informazioni dall'apparato "percettivo" del robot e distribuisco istruzioni al suo apparato motorio senza conoscere l'uno o l'altro di questi fatti. Io sono l'*homunculus* del robot, ma diversamente dall'homunculus tradizionale, non so che cosa succede. Non capisco nulla tranne le regole per la manipolazione dei simboli. Ora, in questo caso, il robot non ha affatto stati intenzionali: semplicemente si muove intorno come risultato del suo sistema di fili elettrici e del suo

programma. E inoltre, instanziando il programma, io non ho stati intenzionali di tipo rilevante. Tutto quello che faccio è seguire istruzioni formali per manipolare simboli formali.

La replica del simulatore del cervello (Berkeley e MIT)

> Si supponga di disegnare un programma che non rappresenti l'informazione che abbiamo sul mondo, come l'informazione negli scritti di Schank, ma che piuttosto simuli l'effettiva sequenza delle esplosioni di neuroni e la sinapsi del cervello di un cinese nativo quando comprende e dà risposte su storie in cinese. La macchina inserisce storie in cinese e domande su di esse come input. Essa simula la struttura formale del cervello cinese nel comprendere queste storie ed emette risposte cinesi come output. Possiamo perfino immaginare che la macchina operi non con un singolo programma seriale, ma con un'intera serie di programmi operanti in parallelo, nella maniera in cui il cervello umano presumibilmente opera quando tratta il linguaggio naturale. Ora, in tale caso, dovremmo certamente dire che la macchina *capisce* le storie: e se rifiutiamo di dire ciò, non dovremmo anche negare che i cinesi nativi capiscono le storie? A livello della sinapsi, che cosa sarebbe o potrebbe essere diverso tra il programma del computer e il *programma* del cervello cinese?

Prima di ribattere, voglio soffermarmi a notare che per ogni partigiano dell'Intelligenza Artificiale (o del funzionalismo, ecc.) è ovvio rispondere che l'ipotesi di IA forte è che non abbiamo bisogno di conoscere come opera il cervello per sapere come opera la mente. L'ipotesi di base, o così avevo supposto, è che c'è un livello di operazioni mentali consistenti in processi computazionali basati su elementi formali che costituiscono l'essenza del mentale e possono essere realizzati in ogni tipo di diverso processo del cervello, allo stesso modo che qualunque programma di computer può essere realizzato nei diversi sistemi hardware: in base all'ipotesi di IA forte la mente sta al cervello come il software sta all'hardware, e così possiamo capire la mente senza fare neurofisiologia. Se dovessimo sapere come lavora il cervello per fare Intelligenza Artificiale, non ci occuperemmo affatto di Intelligenza Artificiale.

Comunque, anche se ci avviciniamo con questa ipotesi all'effettivo funzionamento del cervello, non è ancora sufficiente per giustificare la comprensione. Si immagini che, invece di un uomo che parla una sola lingua in una stanza e confonde simboli, abbiamo lo stesso uomo che opera un elaborato complesso di condutture per l'acqua con valvole che le congiungono. Quando l'uomo riceve i simboli cinesi, va a guardare nel programma, scritto in inglese, quali valvole deve aprire e chiudere. Ogni giuntura dei tubi per l'acqua corrisponde a una sinapsi nel cervello cinese, e l'intero sistema è organizzato in modo tale che, dopo avere azionato tutti i rubinetti giusti, le risposte in cinese saltano fuori dalla parte terminale della serie di tubi.

Ora, dov'è la capacità di comprensione in questo sistema? Esso prende il

cinese come input, simula la struttura formale della sinapsi del cervello cinese, e dà il cinese come output. Ma l'uomo certamente non comprende il cinese, e nemmeno le tubature dell'acqua, e se si è tentati di prendere in considerazione l'ipotesi, per me assurda, che in qualche modo l'unione dell'uomo e dei tubi dell'acqua capisca, si ricordi che, come regola generale, l'uomo può internalizzare la struttura formale delle tubature dell'acqua e comporre tutte le relative aggregazioni di neuroni nella sua immaginazione. Il problema, col simulatore del cervello, è che simula le cose sbagliate del cervello. Finché simula solo la struttura formale della sequenza delle aggregazioni di neuroni e della sinapsi, non avrà simulato ciò che importa nel cervello: precisamente le sue proprietà causali, la sua abilità a produrre stati intenzionali. E che le proprietà formali non siano sufficienti per le proprietà causali, è indicato dall'esempio della tubatura: possiamo avere tutte le proprietà formali disgiunte dalle relative proprietà causali neurobiologiche.

La replica combinata (Berkeley e Stanford)

> Mentre ciascuna delle tre precedenti repliche potrebbe essere, di per sé, non completamente convincente, se le si prende tutte e tre insieme, esse sono collettivamente molto più convincenti e addirittura decisive. Si immagini un robot con un cervello a forma di computer sistemato nella cavità del suo cranio; si immagini il computer programmato con tutte le sinapsi di un cervello umano. Si immagini che il comportamento del robot non si distingua dal comportamento umano, e si pensi a tutto l'insieme come a un sistema unificato e non solo come a un computer con input e output. In tale caso dovremmo certamente attribuire intenzionalità al sistema.

Sono completamente d'accordo sul fatto che in tal caso troveremmo razionale e davvero irresistibile l'ipotesi che il robot ha intenzionalità, finché non sappiamo nulla di più su di esso. In effetti però oltre all'apparenza e al comportamento, gli altri elementi della combinazione sono realmente irrilevanti. Se potessimo costruire un robot il cui comportamento non si distingue dal comportamento umano, attribuiremmo intenzionalità a esso, fino a prova contraria. Non avremmo bisogno di conoscere in anticipo che il suo cervello di computer è un analogo formale del cervello umano. Ma questo non è certo di alcun aiuto nei confronti degli assunti dell'IA forte; e il motivo è questo: secondo l'IA forte, instanziare un programma formale con il giusto input e output è una condizione sufficiente, anzi costitutiva, di intenzionalità. Come Newell (1979) spiega, l'essenza del mentale è l'operazione di un sistema di simboli fisici. Ma le attribuzioni di intenzionalità che diamo al robot in questo esempio non hanno nulla a che fare con i programmi formali. Sono semplicemente basati sull'assunto che, se il robot appare e si comporta approssimativamente come noi, allora noi supporremo, finché non è provato il contrario, che debba avere stati mentali come i nostri, che provocano (e sono espressi da) il suo comportamento, e che debba avere un meccanismo interno

capace di produrre tali stati mentali.

Se però fossimo in grado di giustificare il suo comportamento senza tali assunti, non gli attribuiremmo intenzionalità, specialmente se sapessimo che ha un programma formale. E questo è precisamente la base della mia precedente risposta alla seconda replica. Supponiamo di sapere che il comportamento del robot è interamente giustificato dal fatto che un uomo dentro di esso riceva simboli formali non interpretati dai sensori del robot e mandi ai suoi meccanismi motori simboli formali non interpretati, e che l'uomo faccia questa manipolazione di simboli in conformità a una serie di regole. Supponiamo inoltre che l'uomo non conosca nessuno di questi fatti sul robot: tutto quello che sa è quali operazioni eseguire e quali simboli senza significato utilizzare. In tal caso considereremmo il robot come un ingegnoso manichino meccanico. L'ipotesi che il manichino abbia una mente sarebbe ora non giustificata e non necessaria, perché non c'è più alcuna ragione per ascrivere intenzionalità al robot o al sistema del quale è parte (eccetto naturalmente per la intenzionalità dell'uomo nel manipolare i simboli). La manipolazione di simboli formali continua, l'input e l'output sono combinati correttamente, ma il solo *locus* reale dell'intenzionalità è l'uomo, ed egli non conosce alcuno degli stati intenzionali relativi; per esempio, non vede quel che il robot vede, non sente muovere il braccio del robot e non comprende alcuna delle osservazioni fatte al robot o da parte del robot. Né lo può, per le ragioni affermate prima, il sistema del quale uomo e robot sono una parte.

A riprova, si metta a confronto questo caso con i casi in cui troviamo completamente naturale ascrivere l'intenzionalità a membri di certe altre specie di primati come le scimmie o ad animali domestici come i cani. Le ragioni per cui lo troviamo naturale sono, grosso modo, due: non possiamo trovare un senso nel comportamento dell'animale senza l'attribuzione dell'intenzionalità, e possiamo notare che le bestie sono fatte di materiale simile al nostro — hanno cioè occhi, naso, pelle, e così via. Data la coerenza del comportamento dell'animale e l'assunzione dello stesso materiale causale soggiacente a esso, assumiamo sia che l'animale deve avere stati mentali sottostanti al suo comportamento, sia che gli stati mentali devono essere prodotti da meccanismi ricavati da una materia che è come la nostra. Avanzeremmo certamente ipotesi simili sul robot a meno di non avere qualche ragione per non farlo. Ma non appena fossimo a conoscenza che il comportamento è il risultato di un programma formale, e che le effettive proprietà causali della sostanza fisica sono irrilevanti, abbandoneremmo l'assunto dell'intenzionalità (cfr. *Cognition and Consciousness in Nonhuman Species,* BBS 1(4), 1978).

Ci sono altre due risposte al mio esempio, che ricorrono frequentemente (e sono quindi degne di discussione) ma realmente svisano il vero punto di dibattito.

La replica delle altre menti (Yale)

> Come si viene a conoscere che altre persone capiscono il cinese o qualsiasi altra cosa? Solo dal loro comportamento. Ora il computer può anch'esso superare i test di comportamento. Così, se si ha intenzione di attribuire alle persone la capacità conoscitiva, si deve attribuirla come regola anche al computer.

Questa obiezione in realtà è degna solo di una breve risposta. Il problema, in questa discussione, non verte sul come io so che le persone hanno stati cognitivi, ma piuttosto su che cosa io attribuisco loro quando li accredito di stati cognitivi. La forza dell'argomentazione è che non ci possono essere soltanto procedimenti computazionali e il loro output, perché i procedimenti computazionali e l'output possono esistere senza lo stato cognitivo. Non è una risposta a questo proposito ipotizzare la mancanza delle percezioni. Nelle scienze cognitive si presuppone la realtà e la conoscibilità del mentale nella stessa maniera che nelle scienze fisiche si deve presupporre la realtà e conoscibilità degli oggetti fisici.

La replica delle molte sedi (Berkeley)

> Tutta la tua argomentazione presuppone che l'Intelligenza Artificiale riguardi solo computer analogici e digitali. Ma ciò risulta vero solo allo stato attuale della tecnologia. In ogni caso, qualunque siano i procedimenti causali che tu dici essenziali per la intenzionalità (supponendo che tu sia nel giusto) prima o poi saremo in grado di costruire dispositivi che abbiano appunto tali procedimenti causali: ciò si chiamerà Intelligenza Artificiale. Così tali argomenti non sono in alcun modo diretti a obiettare contro la possibilità dell'Intelligenza Artificiale di produrre e spiegare la cognizione.

Veramente non ho obiezioni a questa risposta salvo a dire che banalizza il progetto di IA forte col ridefinirlo come qualunque cosa che artificialmente produca e spieghi la capacità cognitiva. L'interesse della tesi originale fatta a nome dell'Intelligenza Artificiale è di essere una tesi precisa, ben definita: i procedimenti mentali sono procedimenti computazionali che operano su elementi formalmente definiti.

Mia preoccupazione è stata quella di mettere alla prova quella tesi. Se la tesi è ridefinita in modo tale che non risulta più essere la stessa, le mie obiezioni non risultano più valide perché non c'è alcuna ipotesi attendibile da applicare ad esse.

Ritorniamo ora alla domanda cui ho promesso di rispondere: una volta garantito che comprendo l'inglese e non il cinese, e garantito che la macchina non comprende né l'inglese né il cinese, ci deve pure essere qualcosa in me che crea la situazione per cui capisco l'inglese e, corrispondentemente,

qualcosa che mi manca, per cui non capisco il cinese. Ora, perché non potremmo dare quel qualcosa, qualunque esso sia, a una macchina?

In effetti, non vedo alcuna ragione perché non potremmo dare a una macchina la capacità di capire l'inglese o il cinese, dal momento che fondamentalmente i nostri corpi con i nostri cervelli sono precisamente macchine di questo tipo. Vedo però delle ragioni molto valide per dire che non potremmo dare una tale cosa a una macchina se l'operazione della macchina è definita unicamente in termini di processi computazionali operanti su elementi formalmente definiti; se, cioè, l'operazione della macchina è definita come instanziazione di un programma di computer. Non è perché io sono l'instanziazione di un programma di computer che sono in grado di capire l'inglese e ho altre forme di intenzionalità (io sono, suppongo, l'instanziazione di un numero qualunque di programmi di computer), ma, per quanto sappiamo, è perché io sono un certo tipo di organismo con una certa struttura biologica (cioè chimica e fisica), e questa struttura, sotto certe condizioni, è causalmente capace di produrre percezione, azione, capacità di comprendere, di imparare, e altri fenomeni intenzionali. E nucleo di questo argomento è che solo qualcosa che ha quei poteri causali può avere quella intenzionalità. Forse altri processi fisici e chimici potrebbero produrre esattamente questi effetti; forse, per esempio, anche i marziani hanno l'intenzionalità, anche se i loro cervelli sono fatti di materiale diverso. Questa è una questione empirica, quasi come la questione se la fotosintesi può essere operata da qualcosa come una struttura chimica diversa da quella della clorofilla.

Ma il punto principale dell'argomento è che nessun modello puramente formale sarà mai sufficiente in sé per l'intenzionalità, perché le proprietà formali non sono di per sé costitutive di intenzionalità, e non hanno di per sé poteri causali tranne il potere, una volta instanziate, di produrre il livello successivo del formalismo quando la macchina funziona. E qualunque altra proprietà causale che particolari realizzazioni del modello formale hanno, non è importante rispetto al modello formale, perché possiamo sempre ipotizzare lo stesso modello formale in una diversa realizzazione, in cui quelle proprietà causali sono ovviamente assenti.

Anche se, per qualche miracolo, i nativi cinesi realizzano esattamente il programma di Schank, possiamo mettere lo stesso programma in nativi inglesi, in tubature per l'acqua, o in computers, nessuno dei quali comprende il cinese, a dispetto del programma. Quello che importa nelle operazioni del cervello sono le effettive proprietà delle sequenze della sinapsi. Tutti gli argomenti per la versione forte dell'Intelligenza Artificiale insistono col tracciare un confine intorno alle ombre prodotte dalla capacità cognitiva, dichiarando poi che le ombre sono la cosa reale.

Per concludere, voglio cercare di sottolineare alcuni dei punti filosofici generali impliciti nell'argomento. Per chiarezza tenterò di farlo in forma di domanda e risposta, e comincerò con la famosa domanda:

Una macchina può pensare?

La risposta è, ovviamente, sì. Noi siamo precisamente tali macchine.

Sì, ma può pensare un manufatto, una macchina fatta dall'uomo?

Supponendo che sia possibile produrre artificialmente una macchina con un sistema nervoso, neuroni e dendriti e tutto il resto, che siano sufficientemente simili ai nostri, di nuovo la risposta alla domanda sembra essere, ovviamente, sì. Se si possono raddoppiare esattamente le cause, si potrebbero raddoppiare anche gli effetti. Ed effettivamente potrebbe essere possibile produrre consapevolezza, intenzionalità e tutto il resto usando altri tipi di princìpi chimici, diversi da quelli che usano gli esseri umani. È, come ho detto, una questione empirica.

Bene, ma un computer digitale potrebbe pensare?

Se per computer digitale intendiamo qualsiasi struttura che ha un livello di descrizione che può essere correttamente definito con l'instanziazione di un programma di computer, allora di nuovo la risposta è, naturalmente, sì, dal momento che noi siamo le instanziazioni di un numero elevato di programmi di computer, e possiamo pensare.

Ma potrebbe una cosa pensare, comprendere, e così via, unicamente per il fatto di essere un computer con il giusto tipo di programma? Potrebbe l'instanziazione di un programma, del giusto programma naturalmente, essere di per sé una condizione sufficiente per comprendere?

Questa penso sia la giusta domanda da fare, sebbene sia generalmente confusa con una o più delle precedenti domande, e la risposta a essa non può che essere negativa.

Perché no?

Proprio perché le manipolazioni di simboli formali di per sé non hanno alcuna intenzionalità, esse sono del tutto prive di significato; non sono nemmeno manipolazioni di simboli, dal momento che i simboli non rappresentano nulla. Usando una terminologia linguistica si può dire che

hanno solo una sintassi, ma non una semantica. Tale intenzionalità, quale sembra abbiano i computer, è solamente nelle menti di quelli che li programmano e di quelli che li usano, di quelli che immettono l'input e di quelli che interpretano l'output.

Scopo dell'esempio della stanza cinese era esattamente questo: mostrare che non appena mettiamo qualcosa nel sistema che realmente ha intenzionalità (un uomo), e poi lo programmiamo col programma formale, si può constatare che il programma formale non porta alcuna intenzionalità addizionale. Non aggiunge nulla, per esempio, alla capacità di un uomo di intendere il cinese. Precisamente quella caratteristica di Intelligenza Artificiale che sembrava così seducente — la distinzione tra il programma e la realizzazione — risulta fatale alla tesi che la simulazione potrebbe essere un duplicato. La distinzione tra il programma e la sua realizzazione sembra essere parallela alla distinzione tra il livello delle operazioni mentali e il livello delle operazioni del cervello. E se potessimo descrivere il livello delle operazioni mentali come un programma formale, allora potremmo descrivere che cosa è essenziale alla mente senza fare né psicologia introspettiva né neurofisiologia del cervello.

Ma l'equazione "la mente sta al cervello come il programma sta alla struttura del computer" fallisce in diversi punti, fra i quali i tre seguenti:

1. La distinzione tra programma e realizzazione ha la conseguenza che lo stesso programma può avere ogni genere di folli realizzazioni che non hanno alcun genere di intenzionalità. Weizenbaum (1976), per esempio, mostra in dettaglio come costruire un computer usando un rotolo di carta igienica e un mucchio di piccole pietre. In modo simile, il programma che capisce la storia cinese può essere programmato in una sequenza di tubature d'acqua o in un inglese che parla solo la sua lingua; nessuno dei due acquista da esso peraltro la capacità di comprendere il cinese. Le pietre, la carta igienica, le tubature dell'acqua sono il genere di materiale sbagliato per avere intenzionalità — solo qualcosa che ha gli stessi poteri causali del cervello può avere intenzionalità — e sebbene l'inglese madrelingua possieda il giusto genere di materiale per l'intenzionalità, si può facilmente vedere che egli non riceve alcuna extra-intenzionalità col memorizzare il programma, poiché memorizzarlo non gli insegnerà il cinese.

2. Il programma è puramente formale, ma gli stati intenzionali non sono formali in quel modo. Essi sono definiti nei termini del loro contenuto, non della loro forma. La convinzione che stia piovendo, per esempio, non è definita come una certa struttura formale, ma come un certo contenuto mentale con condizioni di soddisfazione (cfr. Searle, 1979). Infatti la convinzione in quanto tale non ha una struttura formale in senso sintattico, dal momento che a una sola convinzione si può dare un numero indefinito di

espressioni sintattiche diverse in sistemi linguistici diversi.

3. Come ho ricordato prima, gli stati e gli eventi mentali sono letteralmente un prodotto dell'operazione del cervello, mentre il programma non è nello stesso modo un prodotto del computer.

Ma se i programmi non sono in alcun modo costitutivi dei processi mentali, perché tanti hanno creduto il contrario? Questo punto necessita almeno di qualche spiegazione.

Non so davvero la risposta a questo punto. L'idea che le simulazioni del computer potessero essere la cosa reale avrebbe dovuto sembrare sospetta fin dall'inizio, perché il computer non può in alcun modo simulare le operazioni mentali. Nessuno suppone che la simulazione — da parte di un computer — di un incendio distruggerà un quartiere o che la simulazione di un temporale ci lascerà tutti fradici. Per quale motivo uno dovrebbe supporre che la simulazione da parte di un computer della comprensione effettivamente produca comprensione? Si dice, qualche volta, che sarebbe terribilmente difficile far sentire dolore ai computer o farli innamorare, ma l'amore o il dolore non sono né più difficili né più facili della capacità cognitiva o di qualsiasi altra cosa. Per la simulazione, tutto quello di cui si ha bisogno è il giusto input o output e un programma che trasforma il precedente input nel seguente output. Questo è tutto ciò che il computer ha per qualunque cosa faccia. Confondere la simulazione con la duplicazione è il risultato dello stesso sbaglio, sia esso dolore, amore, capacità di conoscere, incendio o temporale.

Ancora, ci sono parecchie ragioni perché l'Intelligenza Artificiale abbia potuto sembrare — e a molti forse ancora sembri — in qualche modo riprodurre e con ciò spiegare i fenomeni mentali, e credo che non riusciremo a rimuovere queste illusioni finché non abbiamo completamente esposto le ragioni che danno loro l'avvio.

La prima, e forse la più importante, è una confusione intorno alla nozione di *trattamento dell'informazione*: molti nelle scienze cognitive credono che il cervello umano, con la sua mente, faccia qualcosa definibile come trattamento dell'informazione e che analogamente il computer con il suo programma faccia trattamento dell'informazione; ma fuochi e temporali, d'altro lato, non fanno alcun trattamento dell'informazione. Così, sebbene il computer possa simulare i caratteri formali di qualunque processo, sta in una speciale relazione con la mente e il cervello, perché quando il computer è appropriatamente programmato, idealmente con lo stesso programma del cervello, il trattamento dell'informazione è identico nei due casi, e questo trattamento è realmente l'essenza del mentale. Ma il guaio di questa tesi è che si fonda su una ambiguità esistente nella nozione di "informazione". Nel

senso in cui gli individui trattano informazione quando riflettono, diciamo, su problemi di aritmetica o quando leggono e rispondono alle domande sulla storia, il computer programmato *non fa* trattamento di informazione. Piuttosto, ciò che esso fa è manipolare simboli formali. Il fatto che il programmatore e l'interprete dell'output usano i simboli come sostituti di oggetti nel mondo è totalmente al di fuori degli obiettivi del computer. Il computer, per ripeterci, ha una sintassi ma non una semantica. Così, se si batte sulla tastiera "2 + 2 è uguale a ?" il computer batterà "4". Ma non ha alcuna idea che "4" significa 4 o qualcos'altro. Il punto non è che esso manca di qualche informazione di secondo ordine per l'interpretazione dei suoi simboli di primo ordine, ma piuttosto che i suoi simboli di primo ordine non hanno interpretazioni almeno per quanto riguarda il computer. Tutto ciò che il computer ha, sono simboli.

L'introduzione della nozione di trattamento dell'informazione perciò provoca un dilemma: o costruiamo la nozione di trattamento dell'informazione in modo tale che essa implichi l'intenzionalità come parte del processo o non lo facciamo. Nel primo caso, il computer programmato non tratta informazione, ma manipola soltanto simboli formali. Nel secondo caso, sebbene il computer esegua un trattamento d'informazione, lo fa solo nel senso in cui le macchine calcolatrici, le macchine da scrivere, lo stomaco, i termostati, i temporali e gli uragani trattano informazioni; precisamente, essi hanno un livello di descrizione per cui possiamo accettarli come capaci di assumere informazione da una parte, trasformarla e produrre informazioni come output. In questo caso, spetta a osservatori esterni interpretare l'input e l'output come informazioni nel senso ordinario del termine. E fra il computer e il cervello non si stabilisce alcuna similarità in termini di similarità nel trattamento dell'informazione.

Secondo, in gran parte dell'IA c'è un residuo di comportamentismo o di operazionalismo. Poiché i computer, programmati appropriatamente, possono avere modelli input-output simili a quelli degli esseri umani, siamo tentati di postulare stati mentali nel computer simili agli stati mentali umani. Ma una volta che vediamo che è possibile, sia concettualmente che empiricamente, che un sistema abbia capacità umane in qualche campo senza avere alcuna intenzionalità, dovremmo poter superare questo impulso.

La mia calcolatrice da tavolo ha capacità di calcolare, ma nessuna intenzionalità, e qui ho cercato di mostrare che un sistema potrebbe avere capacità di input e output che raddoppiano quelle di un cinese madrelingua e tuttavia non comprendere il cinese, indipendentemente da come era stato programmato. Il test di Turing è tipico di questa tradizione in quanto è apertamente comportamentista e operazionalista, e io credo che se gli esperti di IA ripudiassero totalmente il comportamentismo e l'operazionalismo, molta della confusione tra simulazione e duplicazione sarebbe eliminata.

Terzo, questo operazionalismo residuo è congiunto a una forma residua di dualismo: infatti l'IA forte è significativa solo insieme all'assunto dualistico per cui, dove si ha a che fare con la mente, il cervello non c'entra. Nella IA forte (e nel funzionalismo, pure), ciò che importa sono i programmi, e i programmi sono indipendenti dalla loro realizzazione nelle macchine; infatti, finché si tratta di Intelligenza Artificiale, lo stesso programma potrebbe essere realizzato da una macchina elettronica, una sostanza mentale cartesiana, o uno spirito del mondo hegeliano. La scoperta più sorprendente che ho fatto nel discutere questi temi è che molti esperti in IA sono assai turbati dalla mia idea che i reali fenomeni mentali umani possano dipendere da proprietà chimico-fisiche reali di cervelli umani reali. Ma se ci si pensa un attimo, non si può restare sorpresi, poiché, a meno che non si accetti qualche forma di dualismo, il progetto di IA forte non ha possibilità di successo.

Il progetto è di riprodurre e spiegare il mentale col delineare programmi: ma a meno che la mente non sia — non solo concettualmente, ma anche empiricamente — indipendente dal cervello, il progetto non può essere realizzato, poiché il programma è completamente indipendente da ogni realizzazione. A meno che non si creda che la mente è separabile dal cervello sia concettualmente che empiricamente — dualismo in una forma forte — non si può sperare di riprodurre il mentale scrivendo e mettendo in esecuzione programmi, dal momento che i programmi devono essere indipendenti dai cervelli o da qualunque altra particolare forma di instanziazione. Se le operazioni mentali consistono di operazioni computazionali su simboli formali, ne consegue che non hanno alcuna connessione interessante con il cervello; la sola connessione possibile sarebbe che il cervello è uno dei moltissimi tipi di macchine capaci di instanziare il programma. Questa forma di dualismo non è la tradizionale verità cartesiana che dichiara che ci sono due generi di sostanze, ma è cartesiana nel senso che conferma che ciò che è specificamente mentale intorno alla mente non ha alcuna connessione intrinseca con le reali proprietà del cervello.

Questo dualismo soggiacente è mascherato dal fatto che la letteratura dell'Intelligenza Artificiale contiene frequenti denunce contro il "dualismo"; quello di cui gli autori sembrano non essere consapevoli è che la loro posizione presuppone una versione forte di dualismo.

Potrebbe una macchina pensare?

La mia opinione è che *solo* le macchine possono pensare, e solo tipi di macchine molto speciali: precisamente i cervelli e le macchine che hanno gli stessi poteri causali del cervello. Questa è la ragione principale per cui la IA forte ha avuto poco da dirci intorno al pensare, poiché non ha nulla da dirci sulle macchine. Per sua propria definizione verte intorno ai programmi, e i

programmi non sono macchine. Qualunque cosa sia l'intenzionalità, è un fenomeno biologico, e quindi è verosimile che sia causalmente dipendente dalla biochimica specifica delle sue origini come la lattazione, la fotosintesi, o qualunque altro fenomeno biologico. Nessuno dovrebbe supporre che potremmo produrre latte e zucchero eseguendo una simulazione su computer delle sequenze formali nella lattazione e nella fotosintesi, ma quando si tratta della mente molti sono disponibili a credere in tale miracolo a causa del dualismo profondo e perenne: la mente che essi suppongono è una questione di procedimenti formali ed è indipendente da cause materiali specifiche, mentre latte e zucchero non lo sono.

A difesa di questo dualismo si esprime spesso la speranza che il cervello sia un computer digitale (i primi computers, a proposito, erano spesso chiamati "cervelli elettronici"). Ma ciò non serve. Naturalmente il cervello è un computer digitale. Poiché tutto è un computer digitale, il cervello lo è pure. Il fatto è che la capacità causale del cervello di produrre intenzionalità non può consistere nell'instanziazione di un programma di computer, poiché per qualunque programma si voglia, è possibile che qualcosa instanzi tale programma e tuttavia non abbia, per questo, alcuno stato mentale. Qualunque cosa faccia il cervello per produrre intenzionalità, questa non può consistere nell'instanziare un programma, poiché nessun programma, di per sé, è sufficiente per l'intenzionalità.

Riconoscimenti

Sono obbligato a un numero abbastanza cospicuo di persone per la discussione di questi argomenti e per i loro pazienti tentativi di vincere la mia ignoranza sull'Intelligenza Artificiale. Vorrei particolarmente ringraziare Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky e Terry Winograd.

IL DIBATTITO

La tesi di Searle è solo un insieme di simboli cinesi

Robert P. Abelson
Dipartimento di Psicologia, Yale University, New Haven, Conn. 06520

Searle sostiene che i programmi che sembrano dotati di senso comune del progetto di Intelligenza Artificiale di Yale, in realtà non offrono una comprensione significativa del testo. Per lui il computer che analizza una storia sulla visita al ristorante è solo un manipolatore di simboli cinesi, che ciecamente applica regole non comprese a un testo non compreso. Ciò che manca, dice Searle, è la presenza di stati intenzionali.

Searle con questa sua critica incorre in almeno due fraintendimenti. Prima di tutto, non è cosa comune scrivere regole per trasformare i "simboli cinesi" del testo di una storia nei "simboli cinesi" di risposte appropriate a domande intorno alla storia. Liquidare questo fatto come una semplice questione di regole è come degradare un'opera di letteratura a qualcosa che le scimmie del British Museum possono produrre. Il programmatore necessita di una comprensione assai acuta del lavoro reale per scriverne le regole appropriate. Regole mediocri producono un output non intelligente e devono essere riscritte. Rese nitide e chiare le regole, l'output diventa sempre più convincente, sicché il processo di sviluppo delle regole è convergente. Questa è una caratteristica della comprensione di un'area di contenuto, non del cieco esercizio al suo interno.

Certo, Searle direbbe che tale capacità di comprensione è nel programmatore e non nel computer. Ebbene sì: ma che cosa è il risultato? Più precisamente, la capacità di comprensione è nella serie delle regole di programmazione che il computer esercita. Nessuno, che io sappia (a Yale, almeno), ha preteso autonomia per il computer. Il computer non è nemmeno necessario alla teoria rappresentazionale; è solo molto, molto conveniente e molto, molto convincente.

Ma supponiamo di sostenere che il computer stesso capisca il contenuto della storia. Come si potrebbe difendere tale tesi, dato che il computer semplicemente mastica dichiarazioni in codice di programmazione e produce altre dichiarazioni in codice di programmazione che (traduzione successiva) sono considerate da osservatori esterni come corrette e forse perfino intelligenti? Che genere di comprensione è quella? È — affermerei — il genere di comprensione che le persone mostrano per esporre un nuovo contenuto tramite il linguaggio o altri sistemi di simboli. Quando un bambino impara a sommare, che cosa fa se non applicare delle regole? A che punto entra in gioco il "comprendere"? È comprendere il fatto che i risultati dell'addizione si applicano indipendentemente dal contenuto, cosicché m + n = p significa che, se hai m cose e le unisci con n cose, avrai p cose? Ma quella

pure è una regola. È comprendere il fatto che le unità possano essere tradotte in pennies, decine in decimi, centinaia in dollari, cosicché le addizioni dei numeri sono isomorfiche alle addizioni di denaro? Ma quella è una regola che connette sistemi di regole.

In generale, se più regole su un dato contenuto sono incorporate, specialmente se vengono collegate con altre aree di contenuto, sentiamo che il comprendere è in aumento. A che punto una persona passa dal *semplice* manipolare regole al *reale* comprendere?

Gli educazionisti vorrebbero saperlo, e anch'io, ma sarei disposto a scommettere che, in base al test dei simboli cinesi, la maggior parte delle persone che leggono questa pagina, realmente non comprendono il numero trascendentale o l'inflazione o come le barche a vela possono andare controvento (sii onesto con te stesso!). L'argomento stesso di Searle, veleggiando controvento in un ambito carico di simboli che è intrinsecamente difficile da capire, potrebbe ben esser visto come pura manipolazione di simboli. La sua regola fondamentale è che se si considerano i simboli cinesi come "operazioni computazionali formali", allora si possono emettere i simboli cinesi senza affatto capire.

Dato l'esercizio assai comune nelle pratiche umane di interscambio linguistico in aree dove non è dimostrabile che sappiamo di che cosa stiamo parlando, potremmo dare al computer il beneficio del dubbio quando e se opera come noi. Se riteniamo che le persone capiscano in virtù dell'apparente capacità di realizzazione verbale da esse mostrata, potremmo estendere la stessa cortesia alla macchina. È frutto di presunzione, non di capacità di introspezione, dare più credito a se stessi, a parità di realizzazione. Ma Searle tranquillamente liquida questa tesi delle "altre menti" e insiste sul fatto che il computer manca di qualcosa di essenziale. Le regole dei simboli cinesi raggiungono solo questo punto, e per lui, se non si ha tutto, non si ha nulla. Dovrei piuttosto pensare che se non si ha tutto, non si ha tutto.

In ogni caso, l'ingrediente mancante per Searle è il suo concetto di intenzionalità. Nel suo intervento egli non giustifica perché questo è il fattore chiave. Sembra più ovvio che quello di cui manca il manipolatore di simboli cinesi sia la validità estensionale. Il non conoscere che il simbolo *menu* si riferisce a quella cosa che in tutto il mondo puoi tenere, piegare e guardare da vicino, significa perdere ogni possibile reale comprensione di ciò che si intende per menu. Senza difficoltà convengo sull'importanza di tale conoscenza sensomotoria. Comprendere come una barca risalga il vento attraverso il tocco della vela e del timone, è certamente valido e non è lo stesso che una spiegazione verbale.

Quei programmi di computer verbali-concettuali che mancano della connessione sensomotoria col mondo possono certamente fallire. Immaginiamo il seguente brano: "Giovanni disse a Enrico che non riusciva a

trovare il libro. Enrico volse gli occhi al soffitto". I modelli di inferenza del senso comune possono fare varie predizioni sui rapporti di Enrico con il libro e la sua introvabilità. Forse l'ha prestato a Giovanni e perciò è contrariato che sia andato perduto. Ma il significato unico e non scomponibile del volger gli occhi è difficile da afferrare per un modello tranne che attraverso una precisa e concreta voce lessicale di dizionario. Un individuo che comprende, d'altro lato, può imitare il volger dell'occhio di Enrico apertamente o nell'immaginazione e sperimentare globalmente la rassegnata frustrazione che Enrico deve provare: è importante esplorare un ambito di esempi come questo.

Ma perché invece l'*intenzionalità* è così importante per Searle? Se recitiamo la sua litania di parole — speranze, timori e desideri — non arriviamo al punto. Un computer o un uomo non hanno certo bisogno di avere speranze o timori sul cliente per poter comprendere la storia di una visita al ristorante. E l'uso inferenziale di questi concetti è ben all'interno dell'ambito delle capacità dei modelli di comprensione del computer. Le inferenze basate sugli scopi, per esempio, sono un meccanismo standard nei programmi del progetto di IA di Yale. Anzi, lo stato decisivo della intenzionalità per la conoscenza è la considerazione delle condizioni per la sua falsificazione. In che senso il computer comprende che l'asserzione "Giovanni legge il menu" potrebbe essere vera o no e che ci sono altri modi utilizzabili nel mondo reale per esprimere tale concetto?

Bene, Searle presenta a questo punto un argomento importante, sebbene io non lo veda come la carta vincente nella partita che egli pensa di stare giocando. Il computer opera in un modo che è suscettibile di inganno: considera come vera ogni asserzione. Ci sono così certi problemi della conoscenza che non sono stati considerati nei programmi di IA per la comprensione del linguaggio; per esempio, la questione di che cosa fare quando un'opinione sul mondo è contraddetta dai dati: si deve, a questo punto, modificare l'opinione o mettere in questione i dati? Questi problemi sono stati discussi da psicologi nella considerazione e trattazione della conoscenza umana, ma i risultati vanno oltre le capacità reali di Intelligenza Artificiale. Dovremo vedere che cosa succede in quest'area. L'ingenuità del computer sulla validità di quello che gli diciamo è forse toccante, ma sembra difficile che essa giustifichi il totale disappunto mostrato da Searle. Ci sono molte aree di conoscenza entro le quali le questioni di falsificabilità sono del tutto secondarie: la comprensione di un'opera letteraria come un romanzo, per esempio. Searle non ha reso convincente il suo caso a favore dell'essenzialità fondamentale dell'intenzionalità nella comprensione. A ogni modo, il mio manipolatore di simboli cinese non è certo in procinto di estrarre il simbolo corrispondente ad "arrendersi".

Ciò che le intuizioni sull'homunculus non mostrano

Ned Block
Dipartimento di Linguistica e Filosofia, Massachusetts Institute of Technology, Cambridge, Mass. 02139

L'argomento di Searle dipende, per la sua forza, dall'intuizione che certe entità non pensano. Ci sono due semplici obiezioni a questa tesi, basate su considerazioni generali circa quello che può essere mostrato dall'intuizione che può esistere qualcosa che non è in grado di pensare.

Primo, siamo disposti, e giustamente, ad accettare conseguenze controintuitive di asserzioni per le quali abbiamo evidenza sostanziale. Una volta sembrava intuitivamente assurdo asserire che la Terra ruotava nello spazio a velocità vertiginosa, ma di fronte dell'evidenza della tesi copernicana, una tale intuizione doveva essere (e alla fine lo fu) rigettata come non rispondente a verità. Più precisamente, una goccia di protoplasma grigio, grande quanto un pompelmo, sembra, almeno a prima vista, una sede per l'intuizione per niente plausibile. Ma se le tue intuizioni ancora esitano ad accettare il cervello come sede della mentalità dovresti ignorare le tue intuizioni come irrilevanti rispetto alla verità della questione, data la notevole evidenza del ruolo del cervello nella nostra vita mentale. Searle presenta alcune conseguenze dichiaratamente controintuitive della considerazione della conoscenza come manipolazione di simboli formali. Ma la sua tesi non ha nemmeno la forma giusta, perché per sapere se dovremo rigettare la dottrina a causa delle sue conseguenze dichiaratamente controintuitive, dobbiamo sapere che tipo di evidenza c'è in favore della dottrina. Se l'evidenza per la dottrina è schiacciante, allora le intuizioni incompatibili dovrebbero essere ignorate, proprio come lo dovrebbe essere l'affermazione che il cervello non può essere la sede di formazione delle opinioni. Quindi la tesi di Searle implica una premessa mancante perché l'evidenza non è sufficiente a fare accantonare le intuizioni.

È vera, allora, tale premessa mancante? Io penso che chiunque seguisse un corso di primo livello di psicologia cognitiva dovrebbe vedere evidenza sufficiente per trascurare le intuizioni a cui si appella Searle. Molte teorie nella tradizione del pensare inteso come manipolazione di simboli formali hanno un grado moderato (anche se certamente non schiacciante) di supporto empirico.

Un secondo punto contro Searle ha a che fare con un altro aspetto della logica del fare appello all'intuizione. Al meglio, le intuizioni rivelano fatti dipendenti dai nostri concetti (al peggio, fatti intorno a una svariata quantità di fattori come i nostri pregiudizi, l'ignoranza, e, ancor peggio, la nostra mancanza di immaginazione come quando la gente accettò che due linee

diritte non possano incrociarsi due volte). Così, anche se noi dovessimo accettare l'appello di Searle alle intuizioni per mostrare che le teste degli homunculi che formalmente manipolano simboli non pensano, questo dimostrerebbe che le nostre teorie formali di manipolazione di simboli non forniscono una condizione sufficiente per l'applicazione dei nostri concetti ordinariamente intenzionali. Il quesito più interessante comunque è se la manipolazione di simboli formali della testa dell'homunculus cade nello stesso genere naturale scientifico (cfr. Putnam, 1975a) dei nostri processi intenzionali. Se è così, la testa dell'homunculus pensa nel ragionevole senso scientifico del termine — e tanto peggio per il concetto di "ordinario". Inoltre, se a noi interessano molto i concetti intenzionali ordinari, possiamo dare condizioni sufficienti per la loro applicazione costruendoci condizioni *ad hoc* intese a escludere i presunti controesempi. Un primo colpo — inadeguato, ma perfezionabile (cfr. Putnam, 1975b, p. 435 e Block, 1978, p. 292) — consisterebbe nell'aggiungere la condizione che, al fine di pensare, le realizzazioni del sistema manipolatore dei simboli non devono presentare operazioni mediate da entità che presentino, a loro volta, la manipolazione di simboli tipica dei sistemi intenzionali. Il fatto che tale condizione sia *ad hoc* non è un'obiezione, dato che quello che tentiamo di fare è "ricostruire" un concetto semplice ordinario partendo da uno scientifico: possiamo aspettarci che il concetto ordinario sia caratterizzabile scientificamente solo in un modo innaturale (cfr. l'opinione di Fodor su Searle, in questa stessa pubblicazione). Infine c'è una buona ragione per pensare che la spiegazione data da Putnam e Kripke riguardo alla semantica del "pensiero" e ad altri termini intenzionali è corretta. Se è così, e se la manipolazione dei simboli della testa dell'homunculus cade nello stesso genere naturale dei nostri processi cognitivi, allora la testa dell'homunculus pensa realmente, nel senso ordinario come nel senso scientifico del termine.

Il risultato di questi due punti è che il nodo reale del dibattito si fonda su un fatto che Searle non menziona più di tanto: l'evidenza cioè della manipolazione di simboli. Ricordiamo che l'obiettivo di Searle è la teorizzazione della cognizione come manipolazione di simboli, cioè come manipolazione di rappresentazioni attraverso meccanismi che tengono conto solo delle forme delle rappresentazioni. Le teorie della cognizione orientate alla manipolazione di simboli postulano una varietà di meccanismi che generano, trasformano e confrontano rappresentazioni. Una volta che uno vede questa teoria come il reale obiettivo di Searle, può semplicemente ignorare le obiezioni rivolte a Schank. L'idea che una macchina programmata alla Schank abbia qualcosa di simile alla "razionalità" non è degna di essere presa sul serio, e getta tanto dubbio sulla teoria del pensiero come manipolazione di simboli quanto Hitler ne getta sulle teorie che favoriscono un esecutivo autoritario nel governo. Ogni plausibilità dell'idea che una

macchina di Schank pensa, sembrerebbe derivare da una rozza versione del comportamentismo che è un anatema per la maggior parte di coloro che considerano la cognizione come una manipolazione di simboli.[9]

Si consideri un robot simile a quello disegnato nella risposta numero 2 di Searle (omettendo i punti che hanno a che vedere con la sua critica a Schank). Esso simula il vostro comportamento input-output usando una teoria della manipolazione di simboli del genere or ora delineato dei vostri processi cognitivi (insieme con una teoria dei vostri processi mentali non cognitivi, una qualificazione omessa d'ora in avanti). Il suo "corpo" è come il vostro tranne che, al posto di un cervello, ha un computer attrezzato e organizzato con una vera teoria cognitiva su di voi. Voi ricevete un input: "Chi è il vostro filosofo favorito?" Ci pensate un po' su e rispondete: "Eraclito". Se il robot, vostro sosia, riceve lo stesso input, un meccanismo converte l'input in una descrizione dell'input. Il computer usa la sua descrizione dei vostri meccanismi cognitivi per dedurre una descrizione del prodotto della vostra riflessione. Questa descrizione è poi trasmessa a un congegno che trasforma le descrizioni nella risposta concreta: "Eraclito".

Mentre il robot appena descritto si comporta esattamente come fareste voi, dato un qualunque input, non è ovvio che esso abbia degli stati mentali. Voi riflettete per rispondere alla domanda, ma quello che ha luogo nel robot è la manipolazione delle descrizioni della vostra riflessione al fine di produrre la stessa risposta. Non è ovvio che la manipolazione delle descrizioni della riflessione in questo modo sia essa stessa riflessione.

In questo caso, le mie intuizioni si accordano con Searle, ma sul suo argomento ho trovato poco accordo. In assenza di una intuizione ampiamente condivisa, chiedo al lettore di fingere di avere l'intuizione di Searle e la mia su questa faccenda. Ora chiedo un altro tipo di favore, che dovrebbe essere fermamente distinto dal primo: fare il salto dall'intuizione al fatto (un salto che, come ho sostenuto nei primi quattro paragrafi di questo intervento, Searle non ci dà alcuna ragione di fare). Supponga, in grazia della tesi, che il robot descritto sopra non abbia infatti stati intenzionali.

Quello che voglio porre in evidenza è che, se anche concediamo a Searle tutto ciò, la teoria che la cognizione è manipolazione di simboli formali rimane dichiaratamente indenne. Poiché non è parte dell'interpretazione della cognizione come manipolazione di simboli il fatto che il genere di manipolazione attribuita alle descrizioni dei nostri processi cognitivi manipolanti simboli è esso stesso un processo cognitivo. Quelli che credono alle teorie dell'intenzionalità basate sulla manipolazione di simboli formali, devono assegnare intenzionalità a ogni cosa per cui le teorie si dimostrino vere ma non ci si può aspettare che le teorie siano applicabili a quei dispositivi che le usano per imitare gli esseri per i quali esse risultano vere.

Finora ho sottolineato che le intuizioni che il tipo di testa dell'homunculus

ipotizzato da Searle non pensa, non infirmano la dottrina che il pensare è manipolazione di simboli formali. Tuttavia, si può considerare anche una variante dell'esempio di Searle, simile nella forza intuitiva, ma che evita la critica che ho appunto abbozzata. Rammentiamo che scopo della psicologia cognitiva è decomporre processi mentali in successive combinazioni di processi nei quali i meccanismi generano rappresentazioni, altri meccanismi trasformano rappresentazioni, e altri meccanismi ancora confrontano rappresentazioni, producendo spiegazioni per altri meccanismi mentre l'intera rete è appropriatamente connessa a trasmettitori sensori di input e congegni motori di output. Il traguardo di tale teorizzazione è scomporre questi processi fino al punto in cui i meccanismi che eseguono le operazioni non hanno altri comportamenti che possano essere di nuovo scomponibili attraverso una manipolazione simbolica presentando ulteriori meccanismi. Tali meccanismi definitivi sono descritti come "primitivi", e sono spesso raffigurati in diagrammi di flusso come "scatole nere" la cui realizzazione è una questione di struttura e la cui operazione deve essere spiegata dalle scienze fisiche, non dalla psicologia (cfr. Fodor 1968, 1980, e Dennett 1975).

Ora consideriamo una teoria idealmente completa lungo queste linee, una teoria dei nostri meccanismi cognitivi. Immaginiamo un robot il cui corpo è come il nostro, ma la cui testa contiene una serie di homunculi, uno per ogni scatola nera. Ogni homunculus fa il lavoro di manipolazione dei simboli proprio della scatola nera che occupa, trasmettendo il suo "output" ad altri homunculi per telefono, in conformità con la teoria cognitiva. Questa testa di homunculi è appunto una variante di quella che usa Searle, ed evita completamente la critica prima abbozzata, perché la teoria cognitiva che completa è effettivamente valida rispetto a quanto intende giustificare. Chiamiamo questo robot la mente cognitiva degli homunculi (la mente cognitiva degli homunculi è discussa in maggior dettaglio in Block 1978, pp. 305-10). Ribadirò che, se anche avessimo l'intuizione che la mente cognitiva dell'homunculus non ha intenzionalità, non dovremmo guardare a questa intuizione come a un dubbio sulla verità delle teorie del pensiero basate sulla manipolazione dei simboli.

Una linea argomentativa contro la testa cognitiva degli homunculi è che il suo potere di persuasione può essere dovuto all'illusione del "non vedere la foresta al posto degli alberi" (cfr. il testo di Lycan in questo volume). Un altro punto è che l'intuizione bruta e incolta tende a esitare nell'assegnare l'intenzionalità a qualsiasi sistema fisico, incluso il prediletto "cervello" di Searle. Pensa veramente Searle che è un'idea inizialmente congeniale che un piccolo ammasso di gelatina grigia sia la sede della sua intenzionalità? (si potrebbe immaginare un candidato meno probabile?). Quello che rende la materia grigia così soddisfacente per Searle è ovviamente la sua conoscenza che i cervelli sono la sede della nostra intenzionalità. Ma a questo punto

proviamo difficoltà nel fidarci delle intuizioni, precisamente perché esse dipendono dalle nostre opinioni, e fra le opinioni, le più probabili a giocare un ruolo nel caso in questione sono proprio le nostre dottrine sul problema se la teoria del pensiero basata sulla manipolazione di simboli formali sia vera o falsa.

Permettetemi di illustrare questo punto e un altro ancora per mezzo di un esempio (Block 1978, p. 291). Supponiamo che ci sia una parte dell'universo che contiene materia che sia divisibile all'infinito. In quella parte dell'universo ci sono creature intelligenti molto più piccole delle nostre particelle elementari che decidono di dedicare le prossime centinaia di anni a emettere dalla loro materia sostanze aventi le caratteristiche chimiche e fisiche (tranne al livello di particelle subelementari) dei nostri elementi. Essi costruiscono schiere di navi spaziali di diversa varietà sulle misure dei nostri elettroni, protoni e altre particelle elementari e lanciano in volo le navi in modo da imitare il comportamento di queste particelle elementari. Le navi contengono un'attrezzatura per scoprire e produrre il tipo di radiazioni che le particelle elementari emettono. Esse fanno questo per produrre ingenti (per i nostri standard) masse di sostanze con le caratteristiche chimiche e fisiche di ossigeno, carbonio e altri elementi. Voi partite per una spedizione verso quella parte dell'universo e scoprite "ossigeno" e "carbonio". Non conoscendo la loro realtà naturale, voi impiantate una colonia, usando questi "elementi" per coltivare piante per alimenti, fornire aria per respirare, e così via. Poiché le molecole vengono costantemente scambiate con l'ambiente, voi e gli altri colonizzatori venite a essere composti principalmente della "materia" di cui è composta la gente delle navi spaziali.

Se alcune intuizioni sulle menti degli homunculi sono chiare, è evidente che esser fatti della materia prodotta dagli homunculi non colpirebbe il vostro modo di pensare. Così vediamo che l'intuizione non ha bisogno di esitare nell'assegnare intenzionalità a un essere la cui intenzionalità è fondamentalmente debitrice alle azioni di homunculi interni. Perché è così ovvio che esser fatti di materia infestata da homunculi non dovrebbe ferire il nostro sapere o sentire? Io credo che ciò sia dovuto al fatto che abbiamo tutti assorbito abbastanza neurofisiologia per sapere che variazioni nelle particelle del cervello che non colpiscono i meccanismi cerebrali (elettrochimici) di base, non colpiscono la mentalità.

Le nostre intuizioni sulla mentalità delle teste degli homunculi sono ovviamente influenzate (se non determinate) da quello che noi crediamo. Se è così, è Searle che ha il dovere di provare che l'intuizione che la testa cognitiva degli homunculi non ha intenzionalità (un'intuizione che io e molti altri non condividiamo), non è dovuta a una teoria ostile alla spiegazione dell'intenzionalità come prodotto di una manipolazione di simboli. Insomma, una tesi come quella di Searle richiede un attento esame dell'origine

dell’intuizione da cui la tesi dipende, un esame che Searle non comincia.

Riconoscimenti

Sono grato a Jerry Fodor e a Georges Rey per i commenti fatti a una versione precedente.

Cervelli + programmi = menti

Bruce Bridgeman
Commissione degli studi psicologici, Università di California, Santa Cruz, Calif. 95064

Il mio intervento si divide in due parti: una prima afferma che le macchine possono incorporare qualche cosa di più di quanto Searle immagini, e una seconda che afferma che gli uomini incorporano qualche cosa di meno. La mia conclusione sarà che i due sistemi possono per principio conseguire livelli simili di funzionamento.

La mia risposta all’esempio di Searle è una variante della replica del robot: il robot semplicemente necessita di più informazioni, sia ambientali che a priori, di quante Searle sia disposto a dargli. Il robot può internare il significato solo se può ricevere informazioni relative a una definizione di significato, cioè informazioni con una relazione nota con il mondo esterno. Primo, esso necessita di alcune idee innate kantiane, come il fatto che alcune linee di input (per esempio, input dai due occhi o da posizioni nello stesso occhio) sono collegate topograficamente l’una all’altra. In cervelli biologici questo si fa con linee etichettate. Alcuni degli inputs, come quelli visivi, saranno connessi primariamente ai programmi di sviluppo spaziale mentre altri, come quelli acustici, saranno più strettamente riferiti al trattamento del tempo. Inoltre sarà costruito un sistema per evitare alcuni inputs (quelli che rappresentano il dolore, per esempio) e cercarne altri (l’acqua quando si ha sete). Queste proprietà e molte altre ancora sono costruite nella struttura del cervello umano geneticamente, ma possono essere costruite altrettanto facilmente in un programma come una base di dati. Può essere che l’homunculus rappresentato in questo programma non conosca che cosa accade, ma impari presto, perché ha tutte le informazioni necessarie a costruire una rappresentazione di eventi nel mondo esterno.

Il mio super-robot imparerebbe il numero cinque allo stesso modo di un bambino, per interazione con il mondo esterno dove il verificarsi della stringa di simboli rappresentanti “cinque” nei suoi inputs visivi o acustici corrisponde alla più diretta esperienza di cinque oggetti qualunque. Il fatto che dei numeri possono essere codificati nel computer in modi più economici, non è più

rilevante del fatto che il numero cinque è codificato nelle dita della mano di un bambino. Sia la conoscenza a priori che quella ambientale potrebbero essere rese simili in quantità e qualità a quelle disponibili a un essere umano. Ora cercherò di mostrare che l'intenzionalità umana non è così diversa qualitativamente dagli stati della macchina come potrebbe sembrare a un "introspezionista". Il cervello è simile a un programma di computer in quanto riceve esso pure solo input e produce solo output. Gli inputs sono piccoli segnali a 0.1 volt che entrano in gran quantità lungo i nervi afferenti, e gli output sono segnali fisicamente identici che abbandonano il sistema nervoso centrale sui nervi efferenti; il cervello è sordo, muto e cieco, cosicché i segnali elettrici (e alcuni messaggi ormonali che qui ora non devono interessarci direttamente) sono i soli modi che il cervello ha per avere conoscenza del suo mondo o per agire su di esso.

L'eccezione a questa regola è l'informazione esistente, immagazzinata nel cervello, sia quella data in sviluppo genetico che quella aggiunta per esperienza. Ma anche quella è venuta senza un'intenzionalità del tipo che Searle sembra richiedere, poiché l'informazione genetica è ricevuta da lunghe stringhe di sequenze a base di DNA (chiaramente non c'è intenzionalità in questo caso) e i precedenti inputs sono composti delle stesse correnti di segnali a 0.1 volt che costituiscono l'input attuale. Ora è chiaro che nessun neurone che riceve questi segnali o segnali simili generati dentro il cervello ha idea di quello che succede. Il neurone è soltanto una semplice macchina che riceve inputs e genera outputs come una funzione delle relazioni strutturali. Asserire ancora altre proprietà del cervello è il peggior genere di dualismo.

Searle assicura che gli umani hanno intenzionalità e verso la fine del suo articolo ammette che anche degli animali potrebbero avere intenzionalità. Ma quanto lontano giù per la scala filogenetica è disposto ad andare? (cfr. *Cognition and Consciousness in Nonhuman Species*, BBS 1(4), 1978). Un animale monocellulare ha intenzionalità? Certo che no, perché è una semplice macchina che riceve fisicamente inputs identificabili e "automaticamente" genera outputs di riflesso. L'idra con alcune dozzine di neuroni si potrebbe spiegarla nello stesso modo, una semplice rete di nervi con inputs e outputs che sono ristretti, relativamente facili da capire, e trattati secondo modelli fissati. Ora che diremo del mollusco con alcune centinaia di neuroni, dell'insetto con alcune migliaia, dell'anfibio con alcuni milioni, o dei mammiferi con bilioni? Per render convincente questa tesi, Searle necessita di un criterio per porre una linea divisoria nel suo implicito dualismo.

Siamo rimasti con un cervello umano che ha una struttura esente da intenzione, geneticamente determinata, sulla quale sono sovraimposti i risultati di tempeste di sottili segnali nervosi. Da questo in qualche modo deduciamo una intenzionalità che non può essere assegnata alle macchine.

Searle usa l'esempio delle manipolazioni aritmetiche per mostrare come gli umani "comprendano" qualcosa che le macchine non comprendono. Credo che né gli umani né le macchine capiscono i numeri nel senso che intende Searle.

La comprensione di numeri più grandi del cinque è sempre un'illusione, perché gli umani possono trattare con numeri più grandi solo usando espedienti di memorizzazione piuttosto che vera comprensione. Se voglio aggiungere 27 a 54, non uso una comprensione numerica diretta o un analogo di natura spaziale o elettrica nel mio cervello. Applico invece regole che ho memorizzato alle scuole elementari senza realmente sapere che cosa significavano, e combino queste regole con fatti memorizzati sull'addizione di numeri di una cifra per arrivare a una risposta senza comprendere i numeri stessi. Sebbene io abbia la sensazione che eseguo operazioni su numeri, nei termini degli algoritmi che uso non c'è nulla di numerico. Allo stesso modo posso sommare numeri al livello di miliardi, anche se né io né alcun altro abbiamo alcun concetto di che cosa significano questi numeri in termini di quantità percettivamente significativa. Ogni ulteriore comprensione del sistema numerico che possiedo è irrilevante, perché non è usata nell'eseguire semplici calcoli.

L'illusione di avere una consapevolezza dei numeri è simile all'illusione di avere un campo visivo a colore pieno, ben focalizzato: tale concetto esiste nella nostra coscienza, ma la realtà fisiologica è assai lontana dall'introspezione. L'informazione di colori ad alta qualità è disponibile solo nei trenta gradi centrali del campo visivo, e la migliore informazione spaziale in solo uno o due gradi. Suggerisco che la percezione dell'intenzionalità è un'illusione cognitiva simile alla percezione dell'immagine visiva ad alta qualità. La consapevolezza è un sistema neurologico come qualunque altro, con funzioni come la direzione a lungo termine del comportamento (intenzionalità), che ha accesso a ricordi a lungo termine, e parecchie altre caratteristiche che fanno di tale sistema un processore potente, sebbene di capacità limitata, di informazioni biologicamente utili.

Tutte le risposte di Searle al suo *Gedankenexperiment* sono variazioni sul tema che ho qui descritto, che una macchina disegnata adeguatamente potrebbe includere l'intenzionalità come una qualità emergente anche se le parti individuali (transistors, neuroni, o qualunque altra cosa) non ne hanno. Tutte le risposte hanno un elemento di verità: i loro difetti consistono più nell'incapacità di comunicare a Searle la similarità del cervello e delle macchine che in qualche debolezza interna. Forse la differenza più importante tra il cervello e le macchine sta non nella loro organizzazione ma nella loro storia, perché gli umani si sono sviluppati nel senso che sono in grado di eseguire una varietà di funzioni, che includono la riproduzione e la sopravvivenza in un complesso contesto sociale ed ecologico. I programmi,

essendo tracciati senza evoluzione estensiva, hanno scopi e motivazioni più limitati.

L'accusa di Searle al dualismo in Intelligenza Artificiale cade ben al di là del segno perché il meccanicista non insiste su un particolare meccanismo nell'organismo, ma dichiara solo che i processi "mentali" sono rappresentati in un sistema fisico quando il sistema è funzionante. Un programma su disco riposto in un angolo non è più consapevole di un cervello conservato in un vaso di vetro, e il fatto che un programma, se letto in un computer appropriato, funzionerebbe con intenzionalità, conferma soltanto che la macchina adeguata consiste in un'organizzazione imposta su un sostrato fisico. L'organizzazione non è più "mentalistica" del sostrato stesso. L'Intelligenza Artificiale verte sui programmi piuttosto che sulle macchine, solo perché il processo di organizzazione dell'informazione e gli inputs e outputs in un sistema di informazione è stato largamente risolto da computers digitali. Perciò, il programma è il solo passo nel processo di cui preoccuparsi.

Searle può ben avere ragione sul fatto che i programmi attuali (come in Schank e Abelson, 1977) non instanziano l'intenzionalità conformemente alla sua definizione: il problema non sta nel fatto se gli attuali programmi lo facciano, ma nel fatto se sia possibile per principio costruire macchine che fanno piani e conseguono mete. Searle non ci ha dato alcuna prova che ciò non sia possibile.

L'uso e la menzione di termini e la simulazione della comprensione linguistica

Arthur C. Danto
Dipartimento di Filosofia, Columbia University, New York, NY 10027

Nel balletto di Coppelia una ballerina imita una bambola che danza a ritmo d'orologio, la quale a sua volta simula una ballerina. I movimenti che imitano la danza sono prevedibilmente meccanici, date le discrepanze della somiglianza esterna tra ballerine del carillon e ballerine vere. Queste discrepanze possono diminuire fino a zero col progresso tecnologico, finché una ballerina che imita una ballerina di carillon simulante una ballerina può presentare uno spettacolo di tre ballerine indistinguibili in un passo a tre. In base a criteri comportamentistici nulla ci renderebbe possibile identificare la bambola, e la questione se la bambola del carillon stia veramente danzando o semplicemente sembri danzare, appare puramente verbale — a meno che non adottiamo un criterio di significato favorito dal comportamentismo che rende la questione stessa priva di senso.

La questione se le macchine instanziano predicati mentali è stata posta proprio negli stessi termini a partire da Turing, e attraverso il tacito appello

alla indistinguibilità esteriore la questione se le macchine capiscono è eliminata o banalizzata. È in parte un'obiezione contro l'assimilazione del significato dei predicati mentali a semplici criteri di comportamento — un'assimilazione della quale Abelson e Schank sono chiaramente colpevoli, poiché li rende comportamentisti a dispetto di sé stessi — che anima lo sforzo di Searle di imitare un pensatore a orologeria che simula la comprensione: nella misura in cui egli organizza lo stesso programma, esso agisce e fallisce nel comprendere ciò che è capito da quelli che la macchina simula — anche se l'output dei tre non può essere discriminato — e la macchina stessa fallisce nel comprendere. L'argomentazione è pittoresca e può non esser vincolante per quelle persone volte a definire termini come "comprendere" in base a criteri esterni. Così riesaminerò la tesi di Searle in termini logici che devono forzare gli oppositori o ad ammettere che le macchine non capiscono o altrimenti, al fine di sostenere che potrebbero capire, ad abbandonare la teoria essenzialmente comportamentista del significato per adottare quella dei predicati mentali.

Consideriamo, come fa Searle, una lingua che uno non capisce, ma che può dirsi, in senso limitato, che legga. Così posso non comprendere il greco, ma conosco le lettere greche e i valori fonetici associati a esse, e sono capace di pronunciare le parole greche. Le figlie di Milton erano capaci di leggere ad alta voce al loro padre cieco testi dal greco, latino ed ebraico, anche se non avevano la più pallida idea di ciò che leggevano. E sapevano, come me, rispondere a certe domande sulle parole greche, quante lettere ci sono, quali sono i loro nomi, qual è la loro pronuncia. In breve, nei termini della distinzione che i logici formulano tra l'uso e la menzione di un termine, esse conoscevano, come me, quelle proprietà delle parole greche che possono essere identificate da qualcuno che è incapace di usare parole greche in frasi greche. Designamo queste come M-proprietà, in contrasto con le U-proprietà, essendo queste ultime le proprietà che uno deve conoscere al fine di usare il greco. La questione allora è se una macchina programmabile per simulare il comprendere è limitata alle M-proprietà, cioè se il programma è tale che la macchina non può usare le parole che peraltro si può dire manipoli utilizzando le M-regole e le M-leggi. Se è così, la macchina esercita i suoi poteri su quello che possiamo riconoscere nelle parole di una lingua che non capiamo, senza peraltro pensare in quella lingua. È abbastanza evidente infatti che la macchina opera soprattutto attraverso il riconoscimento di modelli in modo molto simile alle infelici figlie di Milton.

Ora devo dare per scontato che non possiamo definire le U-proprietà delle parole in maniera esauriente attraverso le loro M-proprietà. Se questo è vero, le macchine di Schank, limitate alle M-proprietà, non possono pensare nella lingua in cui simulano di pensare. Ci si può chiedere se è possibile alle macchine mostrare l'output che effettivamente mostrano, se tutto quello che

hanno è M-competenza, se no devono avere qualche genere di U-competenza. Ma la difficoltà della questione posta in questi termini è che ci sono due modi nei quali l'output può essere valutato: come se mostrasse comprensione reale o soltanto sembrasse farlo; come tale la struttura del problema è tutt'uno con la struttura del problema mente-corpo, per quanto riguarda il seguente aspetto. Qualunque comportamento esterno, specialmente di un essere umano, noi volessimo descrivere con un predicato psicologico (o mentale) — per esempio, il braccio si è alzato — l'azione di sollevare un braccio ha una descrizione fisica che è vera, sia che la descrizione psicologica sia vera o no. La descrizione fisica allora sottodetermina la distinzione tra movimenti del corpo e azioni, o tra azioni e movimenti del corpo che somigliano esattamente a quelle. Così qualunque sia il predicato psicologico (Ψ) che prende un comportamento esterno, esso prende anche un predicato fisico (Φ) che sottodetermina se il primo è vero o falso rispetto a ciò rispetto a cui il secondo è vero. Perciò non possiamo inferire da una descrizione Φ se si può applicare o no una descrizione Ψ. Per essere sicuri possiamo tranquillamente definire termini come termini Φ, nel qual caso l'inferenza è facile ma banale, ma allora non possiamo più, come Schank e Abelson, spiegare il comportamento esteriore con concetti quali il comprendere. In ogni caso, la distinzione tra M-proprietà e U-proprietà è esattamente parallela: ogni cosa che al livello di output saremmo preparati a descrivere in U-termini ha una M-descrizione di esso che risulta vera e che sottodetermina se la U-descrizione è vera o no. Così nessun modello di output implica che sia usata la lingua, né è da lì che la fonte dell'output comprende, poiché può essere stato intelligentemente disegnato per emettere un modello esaurientemente descrivibile in M-termini.

Il problema è perfettamente cartesiano. Possiamo preoccuparci del fatto che qualcuno dei nostri esseri è un automa. La questione è se la macchina di Schank (SAM) è così programmata che solo le M-proprietà si applicano al suo output. Allora, per quanto esso simuli quello che qualcuno con capacità di comprendere mostrerebbe nel suo comportamento, non un passo è stato fatto verso la costruzione di una macchina che realmente comprenda. E Searle ha veramente ragione, poiché mentre la U-competenza non può essere definita in M-termini, una simulazione M-specificata può essere data in ogni U-realizzazione, per quanto protratta e intricata. Il simulatore mostrerà soltanto, senza averle, le proprietà della U-realizzazione. Le prestazioni possono essere indistinguibili ma si costituisce un uso del linguaggio solo se chi lo emette usa di fatto il linguaggio. Ma non si può dire che si usi il linguaggio se il suo programma, come esso è, è scritto solamente in M-termini.

I princìpi sulla base dei quali un utente della lingua struttura una storia o un testo sono così diversi dai princìpi sulla base dei quali uno potrebbe predire, da certe M-proprietà, quali ulteriori M-proprietà attendersi, che anche se gli outputs sono indistinguibili, i princìpi devono invece essere distinguibili. E

proprio al grado che essi deviano, un programma che impieghi questo secondo tipo di princìpi fallisce nel simulare i princìpi impiegati nel comprendere storie o testi. Il grado di deviazione determina il grado per il quale le tesi forti di IA sono false. E questo è ancor di più il caso se gli M-princìpi non devono essere potenziati mediante gli U-princìpi.

Ognuno di noi può predire quali suoni una persona può emettere quando risponde a certe domande, ma questo accade perché noi capiamo dove essa sta andando. Se dovessimo sviluppare la facoltà di predire suoni solo sulla base di altri suoni, potremmo conseguire una sorprendente congruenza con quello che sarebbe stato il nostro comportamento se avessimo saputo che cosa stava accadendo. Anche se nessuno avrebbe potuto dirlo, la comprensione sarebbe stata nulla.

D'altro lato, resta la questione se la macchina di Schank usa le parole. Se lo fa, Searle ha fallito come un simulatore di qualcosa che non simula, ma genuinamente possiede la capacità di comprendere. Se ha ragione, c'è un'interessante conseguenza. Le M-proprietà producono, per così dire, figure di parole: le macchine, se codificano diverse proposizioni, lo fanno in figure.

Il latte dell'intenzionalità umana

Daniel Dennett
Centro di studi avanzati nelle Scienze del Comportamento, Stanford, Calif. 94305

Voglio separare gli argomenti di Searle, che considero sofisticheria, dalla sua idea positiva, che lancia un'utile sfida all'Intelligenza Artificiale, perché dovrebbe indurre una formulazione più approfondita dei principi di IA. In primo luogo, devo sostenere il carico della sofisticheria col diagnosticare, brevemente, i trucchi con gli specchi che danno al suo caso una certa — spuria — plausibilità. Poi commenterò brevemente la sua idea positiva.

La forma d'argomentazione di Searle è familiare ai filosofi: egli ha costruito quello che si può chiamare "pompa di intuizione", un congegno che provoca una serie di intuizioni col produrre variazioni su un esperimento di pensiero basico. Una pompa di intuizione non è, tipicamente, uno strumento di scoperta, ma uno strumento di persuasione o pedagogico — un modo di far vedere alla gente le cose nel tuo modo, una volta che hai visto la verità, come pensa d'aver fatto Searle. Sarei l'ultimo a screditare l'uso delle pompe d'intuizione — mi piace usarle, infatti — ma il rischio è che si può farne abuso. In questo esempio penso che Searle si fondi quasi interamente su conclusioni errate: intuizioni favorevoli generate da esperimenti presentati in maniera fuorviarne.

Searle comincia con un compito di IA nello stile di Schank, dove sia l'input

che l'output sono oggetti linguistici, frasi in cinese. Per un aspetto, forse, questo è legittimo, poiché Schank e altri hanno certamente legittimato entusiastiche dichiarazioni di comprensione per programmi i cui limiti sono talvolta passati inosservati; ma da un altro punto di vista è un colpo facile poiché è da tempo un'idea ben familiare all'interno dell'IA che tali programmi — li chiamo programmi forzati in un letto poiché i loro unici modi di percezione e azione sono linguistici — al più realizzano un severo taglio nell'interessante compito di modellare una reale capacità di comprendere. Tali programmi non presentano transizione *Language-entry* e *Language-exit*, per usare i termini di Wilfrid Sellars, e non hanno alcuna capacità per una percezione non linguistica o azione corporea. I difetti di tali modelli sono ampiamente riconosciuti da anni in IA: per esempio, il riconoscimento era implicito nella decisione di Winograd di dare a SHRDLU qualcosa da fare al fine di avere qualcosa di cui parlare. "Un computer i cui input e output siano solo verbali sarà sempre cieco rispetto al significato di ciò che è stato scritto" (Dennett 1969, p. 182). L'idea è stata in circolazione per un lungo periodo. Così molti, se non tutti, i sostenitori di IA forte concorderebbero con Searle che nella sua versione iniziale della stanza cinese, nessuno e nulla può capire il cinese, se non in un qualche senso molto bizzarro, ellittico e impreciso. Quindi ciò che Searle chiama "la replica del robot (Yale)" non è una sorpresa, sebbene il fatto che venga da Yale suggerisca che anche Schank e la sua scuola sono ora favorevoli rispetto a questo punto.

La risposta di Searle alla replica del robot è di rivedere il suo esempio, dichiarando che non fa alcuna differenza. Il nostro eroe nella stanza cinese ora controllerà le azioni non linguistiche di un robot, e riceverà le sue informazioni percettive. Ancora (Searle vi chiede di consultare le vostre intuizioni a questo punto) nessuno e nulla capirà veramente il cinese. Ma Searle non si sofferma su quanto grande sia la differenza che questa modificazione produce su quello che ci si chiede di immaginare.

Né Searle si ferma a fornire vivaci dettagli quando di nuovo riconsidera il suo esperimento di pensiero per incontrare la "replica dei sistemi". La risposta dei sistemi suggerisce, a mio parere del tutto correttamente, che Searle ha confuso diversi livelli di spiegazione e attribuzione. Io capisco l'inglese; il mio cervello, no; né in particolare, la parte di esso (se può essere isolata) che opera per "processare" le frasi ed eseguire le mie intenzioni in atti di parola. La descrizione e la discussione della risposta dei sistemi di Searle non è convincente, ma egli è preparato a ritirarsi in ogni caso: la sua proposta è che possiamo di nuovo modificare il suo esempio della stanza cinese, se lo desideriamo, per aggiustare l'obiezione. Dobbiamo immaginare il nostro eroe nella stanza cinese pronto a "internare tutti questi elementi del sistema" cosicché egli "incorpora l'intero sistema". Il nostro eroe non è più ora una

parte subpersonale di un supersistema al quale la comprensione del cinese potrebbe essere appropriatamente attribuita, poiché non c'è alcuna parte del supersistema che sia esterna alla sua pelle. Searle insiste ancora (con un altro pretesto per il nostro sostegno di intuizioni) che nessuno — né il nostro eroe né alcuna altra persona di cui egli possa in qualche senso metafisico essere ora una parte — si può dire che capisca il cinese.

Ma le nostre intuizioni sosterranno Searle quando immaginiamo questo caso in dettaglio? Mettendo insieme entrambe le modificazioni, dobbiamo immaginare il nostro eroe che controlla sia il comportamento linguistico che quello non linguistico di un robot che è *lui stesso*! Quando le parole cinesi per "Mani in alto! Questa è una rapina!" sono modulate direttamente al suo orecchio, incomprensibilmente (e a velocità altissima) simula il programma che porta a por mano al suo portafoglio mentre chiede pietà, in cinese, con le sue labbra. Ora, è davvero del tutto ovvio che, immaginato in questo modo, nessuno nella situazione comprenda il cinese? In effetti, Searle semplicemente non ci ha detto come intende che noi immaginiamo questo caso, che ci è permesso di prendere in considerazione adottando le sue due modificazioni. Dobbiamo supporre che, se le parole fossero state in inglese, il nostro eroe avrebbe risposto appropriatamente nel suo inglese nativo? O è così assorbito nel suo massiccio compito di homunculus che risponde con la incomprensione (simulata) che sarebbe appunto la risposta indotta dal programma a questa serie di inputs incomprensibili (al robot)? Se è vero il secondo caso, il nostro eroe ha preso congedo per sempre dai suoi amici che parlano inglese, annegato all'interno dello spazio di una "persona" che parla cinese dentro il suo corpo. Se è vero il primo caso, la situazione ha drasticamente bisogno di ulteriore descrizione da parte di Searle, perché proprio quello che egli immagina è lungi dall'essere chiaro. Ci sono parecchie alternative radicalmente diverse, tutte così stranamente irrealizzabili da indurci a non affidare a esse le nostre reazioni viscerali. Quando immaginiamo il nostro eroe "incorporante l'intero sistema" dobbiamo pensare che schiaccia bottoni con le dita al fine di far muovere le sue braccia? Sicuramente no, poiché tutti i bottoni sono ora interni. Dobbiamo forse immaginare che quando egli risponde al cinese che gli chiede: "Passa il sale, prego" muovendo la mano per afferrare il sale e portarlo in una certa direzione, non nota che questo è ciò che sta facendo? In breve, uno che si è perfezionato in questo immaginario esercizio può mancare di diventare "scorrevole" in cinese? Forse, ma tutto questo dipende dai dettagli di questo esempio, il solo fondamentale nell'armamentario di Searle che Searle però finisce per non fornire.

Searle ci dice che quando per la prima volta presentò versioni di questo articolo al pubblico di IA, furono sollevate obiezioni che era preparato a incontrare, in parte, modificando il suo esempio. Perché allora non ha

presentato a noi, cioè al suo pubblico successivo, l'esperimento modificato, invece di dilungarsi in quisquilie? Potrebbe essere perché è impossibile raccontare la storia doppiamente modificata in modo tale che si avvicini a un grado di persuasività e dettaglio tali da non provocare intuizioni non volute? Detto in dettaglio, la storia doppiamente modificata suggerisce o che ci sono due persone, una delle quali comprende il cinese, abitante un corpo, o che una persona che parla inglese è, in effetti, stata inglobata entro un'altra persona, una persona che capisce il cinese (tra molte altre cose ancora).

Queste e altre simili considerazioni mi convincono che possiamo voltare le spalle all'esempio della stanza cinese almeno finché una versione migliore non venga esposta.

Nell'attuale stato di incompletezza posso fargli "pompare" le mie opposte intuizioni, almeno nella stessa misura di quelle di Searle.

Che cosa dire, invece, a proposito del contributo che ritengo positivo? A conclusione del suo articolo, Searle osserva: "Nessuno pensa di produrre latte e zucchero mediante una simulazione su computer delle sequenze formali della lattazione e della fotosintesi, ma là dove la mente è coinvolta, molte persone sono disposte a credere in tale miracolo". Non credo che questa sia solo una curiosa illustrazione della visione di Searle; penso che esprima vivacemente la caratteristica che più radicalmente distingue la sua prospettiva attuale dalle tesi prevalenti della dottrina. Per Searle l'intenzionalità è piuttosto come una meravigliosa sostanza secreta dal cervello nello stesso modo in cui il pancreas secerne l'insulina. Il cervello produce l'intenzionalità, dice, mentre altri oggetti, come i programmi del computer, non lo fanno, anche se accade che siano disegnati per imitare il comportamento input-output del cervello o di una parte di esso. C'è inoltre un notevole disaccordo su cosa sia il prodotto del cervello. La maggior parte degli studiosi in IA (e la maggior parte dei funzionalisti nella filosofia della mente) direbbe che il suo prodotto è qualcosa come il controllo: quello a cui un cervello è adibito, è il controllo delle giuste, appropriate, intelligenti relazioni di input-output, dove queste sono considerate, alla fine, relazioni tra inputs sensori e outputs di comportamento di qualche genere. Questo sembra a Searle essere un certo genere di comportamentismo, e lui non vuole avere niente a che fare con esso. Superare il test di Turing può essere, al primo aspetto, una prova che se qualcosa ha intenzionalità, realmente ha una mente, ma "non appena sappiamo che il comportamento era il risultato di un programma formale, e che le effettive proprietà causali della sostanza fisica erano irrilevanti, dovremmo abbandonare l'assunto dell'intenzionalità".

Così nell'opinione di Searle le "giuste" relazioni input-output sono sintomatiche, ma non forniscono una prova conclusiva di intenzionalità; la prova del pudding si ha alla presenza di qualche proprietà causale (interamente non specificata) che è interna alle operazioni del cervello.

Questa internalità necessita di alcuni chiarimenti. Quando Searle parla di proprietà causali si può pensare dapprima che quelle proprietà causali fondamentali per l'intenzionalità sono quelle che collegano le attività del sistema (cervello o computer) alle cose nel mondo con le quali il sistema interagisce, incluso in primo luogo l'attivo corpo sensibile il cui comportamento è controllato dal sistema. Ma Searle insiste che queste non sono le proprietà causali rilevanti. Egli ammette la possibilità in linea di principio di duplicare la competenza di input e output di un cervello umano mediante un "programma formale" che, convenientemente innescato, possa guidare un corpo attraverso il mondo, esattamente come vorrebbe il cervello di quel corpo, acquistando così tutte le relative proprietà causali extrasistemiche del cervello. Ma tale sostituto del cervello fallisce totalmente nel produrre intenzionalità, sostiene Searle, perché manca di alcune altre proprietà causali delle operazioni interne del cervello. Tuttavia, come possiamo sapere che manca di queste proprietà, se tutto quello che sappiamo è che è lo sviluppo di un programma formale? Dal momento che Searle ammette che qualsiasi operazione — anche quella di un cervello umano — può essere descritta in termini di esecuzione di un programma formale, la semplice esistenza di un tale livello di descrizione di un sistema non potrebbe precludere la possibilità che esso abbia intenzionalità. Sembra che sia solo quando possiamo vedere che il sistema in questione consiste nello sviluppo di un programma formale, che possiamo concludere che non produce intenzionalità. Ma nulla può consistere solo nello sviluppo di un programma formale: i computers emanano calore e rumore durante le operazioni: perché quindi non dovrebbero produrre anche intenzionalità?

Inoltre, qual è il prodotto più importante e quale il prodotto secondario? Searle può difficilmente negare che i cervelli in effetti producono delle quantità di controllo corporeo adeguato e appropriato. Essi fanno questo, pensa, producendo intenzionalità, ma ammette anche che un computer con le giuste regole di input e output potrebbe produrre il controllo senza produrre o usare alcuna intenzionalità. Ma allora il controllo è il prodotto principale e l'intenzionalità solo un mezzo, senza dubbio naturale, di ottenerlo. Se i nostri antenati fossero stati organismi mutanti privi di intenzionalità con semplici sistemi di controllo, la natura li avrebbe prontamente selezionati (devo questa considerazione a Bob Moore). O, per guardare l'altro lato della medaglia, cervelli con una quantità di intenzionalità, ma nessuna competenza di controllo, sarebbero un prodotto ecologicamente irrilevante, che l'evoluzione non proteggerebbe. Fortunatamente per noi, tuttavia, il nostro cervello produce intenzionalità: se non lo facesse, ci comporteremmo proprio come facciamo ora, ma naturalmente non ne avremmo consapevolezza!

Sicuramente Searle non sostiene la tesi che ho appena ridicolizzato, anche se sembra che lo faccia. Egli non può davvero considerare l'intenzionalità

come un meraviglioso fluido mentale: per cui, che cosa tenta di dimostrare? Io penso che le sue perplessità riguardo alle proprietà interne dei sistemi di controllo siano un maldestro tentativo di catturare il punto di vista interiore di un agente consapevole: egli non vede come un semplice computer possa giustificare tale punto di vista. Ma questo è dovuto al fatto che egli guarda troppo in profondità. È appunto altrettanto misterioso se scrutiamo nella giungla di sinapsi del cervello e ci chiediamo dove è nascosta la consapevolezza. Non è a quel livello di descrizione che si troverà un adatto soggetto di consapevolezza. Esso è la risposta dei sistemi che Searle non vede ancora come un passo nella giusta direzione lontano dalla sua versione aggiornata di *élan vital.*

Una prospettiva dualista-interazionista

John C. Eccles
Ca à la Gra, Contra (Locarno) CH-6611, Svizzera

Searle afferma chiaramente che la base della sua valutazione critica dell'IA poggia su due proposizioni. La prima è: "L'intenzionalità negli esseri umani (e negli animali) è il prodotto delle caratteristiche causali del cervello". Egli sostiene questa proposizione con l'affermazione che "è un fatto empirico che riguarda le effettive relazioni causali tra processi mentali e cervello. Ciò significa semplicemente che certi processi mentali sono *sufficienti* per l'intenzionalità" (il corsivo è mio).

Questo è un dogma della teoria dell'identità psiconervosa, che è una variante delle teorie materialiste della mente. Non c'è menzione dell'ipotesi alternativa dell'interazionismo dualista che Popper e io pubblicammo qualche tempo fa (1977) e che ho ulteriormente sviluppato più recentemente (Eccles 1978, 1979). In base a quell'ipotesi l'intenzionalità è una proprietà della mente autocosciente (*Mondo 2* di Popper), essendo il cervello usato come strumento nella realizzazione delle intenzioni. Mi riferisco alla figura E 7-2 di Popper ed Eccles (1977), dove le intenzioni appaiono nello spazio del Mondo 2, con frecce che indicano il flusso delle informazioni per mezzo del quale le intenzioni della mente causano mutamenti ed eventualmente movimenti volontari.

Non ho invece alcuna difficoltà con la seconda proposizione, ma suggerirei che i punti 3, 4 e 5 siano riscritti con "mente" sostituito da "cervello". Di nuovo l'affermazione "Solo una macchina potrebbe pensare, e solo tipi molto speciali di macchine… con poteri causali interni equivalenti a quelli del cervello" è il dogma della teoria dell'identità. Dico dogma perché è incontestato e senza sostegno empirico. La teoria dell'identità è molto debole empiricamente, essendo semplicemente la teoria di una promessa.

Finché Searle parla dell'attività umana senza guardare all'intenzionalità come a una proprietà del cervello, posso apprezzare il fatto che ha prodotto argomenti provanti contro la teoria dell'IA forte. La storia dell'hamburger con il *Gedankenexperiment* dei simboli cinesi è collegato ai tentativi di Premack di insegnare allo scimpanzé Sara un livello primitivo di linguaggio umano espresso in simboli (cfr. Premack, *Lo scimpanzé ha una teoria della mente?*, BBS 1(4) 1978). La critica di Lenneberg (1975) era che, per mezzo del condizionamento, Sara aveva imparato un gioco simbolico, usando strumentalmente simboli, ma non aveva alcuna idea che fossero collegati al linguaggio umano. Egli addestrò studenti di scuola superiore con i metodi descritti da Premack, applicando rigidamente lo studio di Premack. I soggetti umani furono presto in grado di ottenere risultati considerevolmente migliori dello scimpanzé: ma non furono capaci di tradurre correttamente una sola frase completa in inglese. Infatti essi non capivano che c'era corrispondenza tra i simboli e il linguaggio: erano convinti che il loro compito fosse di risolvere semplici indovinelli.

Penso che questo banale esperimento indichi un difetto fatale in tutto il lavoro di IA. Non importa quanto sia complessa l'esecuzione instanziata dal computer: non può infatti essere che un trionfo per il progettista di simulazione con il calcolatore. La macchina di Turing è il successo di un mago o un incubo! È sorprendente che, dopo le particolareggiate affermazioni riguardanti cervello e mente, non abbia trovato la parola "cervello" nel testo di Searle, nella sua intera argomentazione: dove egli usa mente, stati mentali, comprensione umana e stati cognitivi esattamente come si farebbe in un testo sull'interazionismo dualista. Il termine cervello non appare fino alla "replica del robot", dove si trova come "cervello del computer". Comunque, dalla "risposta simulatrice del cervello" alle affermazioni e alle critiche delle altre varie repliche, cervello, neuroni, sinapsi e simili sono usati a profusione in un modo abbastanza semplicistico. Per esempio, "immaginare il computer programmato con tutte le sinapsi di un cervello umano" è più di quello che posso fare e per diversi ordini di grandezza! Così "la replica della combinazione" è una fantasia e non ha alcun riscontro concreto.

Sono d'accordo che è un errore confondere la simulazione con la duplicazione. Ma non mi oppongo all'idea che la distinzione tra il programma e la sua realizzazione sull'hardware sembra essere parallela alla distinzione tra le operazioni mentali e il livello delle operazioni del cervello. In ogni caso, Searle crede che l'equazione "la mente sta al cervello come il programma sta all'hardware" fallisca in diversi punti. Preferirei sostituire "programmatore" a "programma", perché, come interazionista dualista, accetto l'analogia che come esseri coscienti noi funzioniamo come programmatori dei nostri cervelli. In particolare rifiuto il terzo argomento di Searle: "Stati ed eventi mentali sono letteralmente un prodotto delle operazioni del cervello, mentre il

programma non è allo stesso modo un prodotto del computer" e così più avanti ci si dice: "Qualunque cosa sia l'intenzionalità, essa è un fenomeno biologico, ed è probabile che sia altrettanto causalmente dipendente dalla biochimica specifica delle sue origini quanto la lattazione, la fotosintesi, o qualunque altro fenomeno biologico". Ho la sensazione di essere trasportato indietro nel diciannovesimo secolo, quando, come fu ironicamente registrato da Sherrington (1950), "l'oracolare professor Tyndall, presidente dell'Associazione Britannica a Belfast, disse al suo pubblico che, come la bile è una secrezione del fegato, così la mente è una secrezione del cervello".

Riepilogando, le mie critiche sorgono da fondamentali differenze di opinione rispetto al problema cervello-mente. Finché Searle si riferisce alle intenzioni e attività umane senza riferimento al problema cervello-mente, posso apprezzare le critiche che egli lancia contro le opinioni di IA che un computer appropriatamente programmato è una mente che letteralmente comprende e ha altri stati cognitivi.

La maggior parte delle critiche di Searle sono accettabili per un interazionismo dualista. È veramente ora che l'IA forte sia messa in dubbio.

Searle su quello che solo il cervello sa fare

J.A. Fodor
Dipartimento di Psicologia, Massachusetts Institute of Technology, Cambridge, Mass. 02139

1. Searle ha certamente ragione quando afferma che instanziare lo stesso programma che instanzia il cervello non è, in sé e per sé, una condizione sufficiente per avere quelle attitudini proposizionali caratteristiche di un organismo che ha il cervello. Se qualcuno in IA pensa che lo sia, sbaglia. Per quanto riguarda il test di Turing, esso ha tutte le difficoltà proprie delle predizioni di "nessuna differenza"; non si può distinguere la verità della predizione dalla insensibilità dello strumento usato per il test.[10]

2. Non convince affatto, invece, il modo in cui Searle tratta la "replica del robot". Dato che ci sono i giusti tipi di collegamenti causali tra i simboli che il robot manipola e le cose nel mondo compresi i trasduttori efferenti ed afferenti del dispositivo, non è per niente chiaro che l'intuizione si rifiuti di attribuire a esso attitudini proposizionali. Tutto quello che l'esempio di Searle mostra è che il genere di collegamento causale che immagina — quello che è, in effetti, mediato da un uomo "seduto nella testa di un robot" — non è, e la cosa non sorprende, del tipo giusto.

3. Non sappiamo dire quali sono i giusti tipi di collegamenti causali. Anche questo non sorprende, poiché non sappiamo come rispondere alla domanda strettamente collegata su quali siano i generi di connessione tra una formula e

il mondo che determinano l'interpretazione secondo la quale viene impiegata la formula. Non abbiamo una risposta a questa domanda per nessun sistema simbolico; *a fortiori*, non l'abbiamo per rappresentazioni mentali. Queste domande sono strettamente collegate perché, data l'ipotesi della rappresentazione mentale, è naturale ritenere che ciò che rende gli stati mentali intenzionali è anzitutto che essi implicano relazioni con oggetti mentali semanticamente interpretati: ovvero, relazioni del giusto tipo.

4. Mi sembra che Searle abbia frainteso il punto principale sul trattamento dell'intenzionalità con teorie rappresentazionali della mente: non è sorprendente poiché i proponenti della teoria — speciamente in IA — sono stati notevolmente oscuri nell'esporla. Il punto principale è questo: proprietà intenzionali di attitudini proposizionali sono considerate come eredità di proprietà semantiche di rappresentazioni mentali (e non come derivanti dal ruolo funzionale delle rappresentazioni mentali, a meno che "ruolo funzionale" non sia considerato abbastanza largamente da inglobare le relazioni simbolo-mondo). In effetti, ciò che viene proposto è una riduzione del problema di che cosa renda intenzionali gli stati mentali, al problema di che cosa fissi le proprietà semantiche di un simbolo. Questa riduzione appare promettente perché dovremo comunque rispondere alla seconda domanda (per esempio, nel costruire teorie di linguaggi naturali): e noi comunque abbiamo bisogno della nozione di rappresentazione mentale (per esempio, per fornire appropriati ambiti per i processi mentali). Può essere importante aggiungere che non c'è nulla di nuovo riguardo a questa strategia. Locke, per esempio, pensava: a) che le proprietà intenzionali degli stati mentali sono ereditate dalle proprietà semantiche (referenziali) di rappresentazioni mentali; b) che i processi mentali sono formali (associativi); c) che gli oggetti dai quali gli stati mentali ereditano la loro intenzionalità sono gli stessi in base ai quali i processi mentali sono definiti, cioè le idee. È mia opinione che nessuna seria alternativa a questo trattamento di attitudini proposizionali sia stata mai proposta.

5. Dire che un computer (o un cervello) esegua operazioni formali su simboli non è lo stesso che dire che esegue operazioni su simboli formali, cioè non interpretati. Questo equivoco si presenta ripetutamente nell'articolo di Searle, e causa considerevole confusione. Se ci sono rappresentazioni mentali esse devono, naturalmente, essere interpretate come oggetti: è perché sono oggetti interpretati che gli stati mentali sono intenzionali. Ma per tutto ciò il cervello potrebbe essere un computer.

6. Questa situazione — che necessita di una nozione di connessione causale, ma senza sapere quale nozione di connessione causale è quella giusta — è del tutto familiare in filosofia. È per esempio estremamente plausibile che *a percepisca b* possa essere vero solo dove c'è il giusto genere di connessione causale tra *a* e *b.* Ma noi non sappiamo qual è qui il giusto

genere di connessione causale. Dimostrare che certi tipi di connessione causale sono errati, non pregiudicherebbe, naturalmente, la tesi. Per esempio, supponiamo di avere interposto un omino tra *a* e *b*, la funzione del quale è di riferire da *a* a *b*. Avremo allora (*inter alia*) una specie di legame causale da *a* a *b*, ma non avremo la specie di legame causale che si richiede ad *a* per percepire *b*. Sarebbe naturalmente errato concludere, dal fatto che questo collegamento causale non riesce a ricostruire la percezione, che nessun collegamento causale ci riuscirebbe. La tesi di Searle contro la "replica del robot" è proprio un'idea errata di questo genere.

7. È del tutto ragionevole (infatti deve essere vero) che il giusto genere di relazione causale è quello che esiste tra il nostro cervello e i nostri meccanismi di inferenza (da un lato) e tra il nostro cervello e gli oggetti distanti (dall'altro). Non ne seguirebbe con questo che *solo* il nostro cervello può trasferire tali relazioni mediante inferenze; non ne seguirebbe neppure che essere lo stesso genere di cosa che è il nostro cervello (in ogni senso biochimico di stesso genere) sia una condizione necessaria per essere in quella relazione; e non ne seguirebbe neppure che le manipolazioni formali dei simboli non sono tra gli anelli in tali catene causali. E, anche se i nostri cervelli fossero le uniche entità che possono essere in quella relazione, il fatto che essi lo siano, potrebbe non essere proprio di alcun interesse particolare, in quanto dipenderebbe dal perché è vero.[11] Searle non dà alcuna prova del perché egli pensa che la biochimica è importante per l'intenzionalità, e, *prima facie*, l'idea che ciò che conta è *come* l'organismo è connesso al mondo, sembra di gran lunga più plausibile. Dopo tutto, è abbastanza facile immaginare, grosso modo, come il fatto che il mio pensiero è causalmente connesso a un albero potrebbe legarsi al suo essere un pensiero intorno a un albero. Ma è difficile immaginare che il fatto che il mio pensiero consiste di idrocarburi potrebbe essere importante, eccetto nell'improbabile ipotesi che solo gli idrocarburi possono essere causalmente connessi agli alberi nel modo in cui lo è il cervello.

8. La prova empirica per credere che la "manipolazione dei simboli" è implicata nei processi mentali deriva largamente dal considerevole successo del lavoro in linguistica, psicologia e Intelligenza Artificiale che è stato fondato su quell'assunto. Pochi dei dati relativi interessano la simulazione del comportamento o il superamento del test di Turing, sebbene Searle scriva come se tutto ciò lo fosse davvero. Searle non dà indicazione alcuna di come i fatti che questo lavoro giustifica devono essere spiegati se non in base alla tesi di *processi-mentali-che-sono-processi-formali*. Dichiarare che non c'è alcuna prova che la manipolazione dei simboli è necessaria per l'analisi mentale, mentre si ignorano sistematicamente tutte le prove che sono state allegate in favore della tesi, suscita l'impressione che tutto ciò sia una curiosissima strategia da parte di Searle.

9. Alcune condizioni necessarie sono più interessanti di altre. Mentre le connessioni col mondo e le manipolazioni di simboli sono entrambe presumibilmente necessarie per i processi intenzionali, non c'è ragione (fin qui) di credere che esse forniscano un dominio teoretico per una scienza: mentre invece, c'è una considerevole ragione a posteriori per supporre che i processi intenzionali lo facciano. Se questo è giusto, ciò fornisce qualche giustificazione per la pratica dell'Intelligenza Artificiale; se non lo è, per la retorica dell'Intelligenza Artificiale.

10. *Parlare* implica eseguire certe operazioni formali su simboli: collegare le parole insieme. Eppure, non tutto ciò che può legare le parole insieme può parlare. Da queste banali osservazioni non segue che ciò che pronunciamo sono suoni non interpretati, o che non capiamo ciò che diciamo, o che chiunque parla dice cose senza senso, o che solo gli idrocarburi possono fare affermazioni: succede la stessa cosa, *mutatis mutandis*, qualora si sostituisca "pensare" a "parlare".

Programmi, poteri causali e intenzionalità

John Haugeland
Centro per lo studio avanzato nelle Scienze del Comportamento, Stanford, Calif. 94305

Searle è in un laccio. Egli nega che ogni test di Turing per l'intelligenza sia adeguato: e cioè che il comportarsi intelligentemente sia una condizione sufficiente per essere intelligente. Ma non osa negare che creature fisiologicamente molto diverse dalle persone possano nondimeno essere intelligenti. Così egli necessita di un criterio intermedio: non così specifico per noi da escludere gli alieni, e neppure così lontano dalle specificità da ammettere un qualunque oggetto che presenti il giusto comportamento. Il suo suggerimento è che solo oggetti con "i giusti poteri causali" possono avere intenzionalità, e da qui che solo tali oggetti possono genuinamente comprendere alcunché o essere intelligenti. Questo suggerimento, comunque, è incompatibile con la principale tesi di questo articolo.

Palesemente, la tesi è contro la dichiarazione che lavorare secondo un certo programma possa mai essere sufficiente per capire alcunché — non importa quanto intelligentemente il programma sia tracciato in modo che l'oggetto relativo (computer, robot, o qualunque altra cosa) si comporti *come se* capisse. L'azione cruciale è riporre il processore centrale (CPU) in una persona superveloce (che potremmo ben chiamare "il demone di Searle"): Searle sostiene che un demone che parli inglese potrebbe benissimo seguire un programma per simulare un cinese madrelingua, senza capire egli stesso una parola di cinese. Il guaio è che la stessa strategia funziona altrettanto bene

contro ogni specificazione dei “giusti poteri causali”; invece di manipolare simboli formali in base alle specificazioni di qualche programma di computer, si manipoleranno piuttosto stati fisici o variabili in base alla specificazione delle “giuste” interazioni causali. Proprio per essere concreti, immaginiamo che quelli giusti siano i poteri che i nostri neuroni devono “stuzzicare” l’uno con l’altro mediante trasmettitori. Gli alieni verdi possono essere intelligenti, anche se sono costituiti di materiale siliceo, perché i loro neuroni (al silicio) avrebbero lo stesso potere di stuzzicamento reciproco. Ora immaginiamo di coprire ciascuno dei neuroni di un criminale cinese con un sottile strato, che non ha alcun effetto, se non di renderlo inaccessibile ai trasmettitori neurali. Immaginiamo inoltre che il demone di Searle possa vedere il problema e venga in soccorso; egli scruta attraverso lo strato superficiale in ciascuna estremità neurale, determina quale trasmettitore (se ce n’è) è stato emesso, e poi colpisce le estremità adiacenti in modo da ottenere lo stesso effetto che si sarebbe avuto da quel trasmettitore. Fondamentalmente, il demone sostituisce i neurotrasmettitori.

Per ipotesi, il comportamento della vittima è immutato: in particolare, essa agisce ancora come se capisse il cinese. Ora, comunque, nessuno dei suoi neuroni ha i giusti poteri causali, ma il demone li ha, e capisce solo l’inglese. Perciò, avere i giusti poteri causali (anche se racchiuso in un sistema tale che l’esercizio di questi poteri porta a un comportamento “intelligente”) non può essere sufficiente per la comprensione. Inutile dire che una corrispondente variazione funzionerà, qualunque siano i poteri causali rilevanti.

Nessuno di questi argomenti dovrebbe risultare una sorpresa. Un programma di computer è appunto una specificazione dell’esercizio di certi poteri causali: il potere di manipolare vari simboli formali (oggetti fisici o stati di qualche genere) in certi modi specificati, in dipendenza dalla presenza di certi altri simboli. Naturalmente è un modo particolare di specificare esercizi causali di un genere particolare, ed esso è quello che dà al “paradigma computazionale” il suo carattere distintivo. Ma Searle non fa uso di questa particolarità: la sua tesi dipende solo dal fatto che i poteri causali possono essere specificati indipendentemente da qualunque cosa sia ciò che possiede il potere. Questo è precisamente ciò che rende possibile interpretare il demone, sia nel programma dell’interazione dei simboli che nei casi di interazione dei neuroni.

Non si può evitare di sostenere che questa è una tesi “dualistica” dei poteri causali, non intrinsecamente connessa con le “attuali proprietà” degli oggetti fisici. Parlare di poteri causali in un modo che permetta la generalizzazione (ad alieni verdi, per esempio) implica, *ipso facto*, un’astrazione dai particolari di ogni data “realizzazione”. Il punto è indipendente dall’esempio; funziona infatti altrettanto bene per la fotosintesi. Organismi simili a piante, del colore della carne, su un pianeta alieno, potrebbero fotosintetizzare (in senso pieno e

letterale) se contenessero una sostanza chimica (non necessariamente clorofilla) che assorba luce e usi energia per fare zucchero e liberare ossigeno da diossido di carbonio e acqua. Questo significa specificare la fotosintesi come un *potere causale*, piuttosto che appunto una proprietà che è, per definizione, idiosincratica alla clorofilla. Ma ora, naturalmente, il demone può arrivare e rimpiazzare sia la clorofilla che il suo sostituto: esso divora i fotoni e, così energizzato, produce zucchero da CO_2 e H_2O. Mi sembra che il demone fotosintetizzi.

Ma mettiamo da parte l'argomento demone. Searle suggerisce che "non c'è alcuna ragione per supporre" che la capacità di comprendere (o l'intenzionalità) "abbia qualcosa a che fare" con i programmi di computer. Pure questo, penso, si basa sul suo mancato riconoscimento che specificare un programma è (in un modo distinto) specificare una serie di poteri causali e interazioni. Il risultato principale è quello che differenzia l'intenzionalità originale da quella derivata. La prima è l'intenzionalità che una cosa (sistema, stato, processo) possiede "per suo proprio diritto"; la seconda è l'intenzionalità che è "presa in prestito da" o "conferita da" qualcos'altro. Così (per ammissione generale, che non metterò qui in discussione) l'intenzionalità del pensiero conscio e della percezione è originale, mentre l'intenzionalità (significato) dei simboli linguistici è semplicemente conferita a essi dagli utenti del linguaggio: si tratta di parole che non hanno alcun significato in sé e per sé, ma solo in virtù del fatto che gliene diamo uno. Questi sono casi paradigmatici: molti altri casi falliranno chiaramente per un lato o per l'altro, o saranno discutibili, o forse perfino marginali. Nessuno nega che se i sistemi di Intelligenza Artificiale non hanno intenzionalità originale, hanno almeno intenzionalità derivata, in senso non comune (non banale) perché hanno interpretazioni non banali. L'obiezione di Searle, sostenuta da molti, è che sistemi di Intelligenza Artificiale anche abbastanza buoni non hanno (né avranno) intenzionalità originale.

Simboli del pensiero, come credenze articolate e desideri, e simboli linguistici, come le espressioni di credenze articolate e desideri, sembrano avere molto in comune: come evidenziato, per esempio, da Searle (1979c). In particolare, eccetto per la distinzione originale-derivata, essi hanno (o almeno sembrano avere) strutture e variazioni semantiche strettamente parallele. Ci deve essere allora qualche altra distinzione di principio tra esse, in virtù della quale le prime possono essere intenzionali in maniera originaria e le seconde solo in maniera derivata. Un buon candidato per questa distinzione è l'idea che i pensieri sono attivi semanticamente, mentre i simboli di frase trascritti, diciamo, su una pagina, sono semanticamente inerti. I pensieri sono costantemente interagenti l'uno con l'altro e il mondo, in modi che sono semanticamente appropriati al loro contenuto intenzionale. Le interazioni causali di frasi-simbolo scritte, d'altro lato, non riflettono in maniera

consistente il loro contenuto (eccetto quando interagiscano con la gente). I pensieri sono incorporati in un "sistema" che fornisce canali normali perché essi interagiscono col mondo, e tali che queste normali interazioni tendono a ottimizzare la corrispondenza tra loro e il mondo; tramite la percezione, le opinioni tendono alla verità: e, tramite l'azione, il mondo tende a ciò che è desiderato e prefissato. Ci sono canali di interazione tra pensieri (vari tipi di inferenza) tramite i quali la serie dei pensieri tende a diventare più coerente, e a contenere più conseguenze dei suoi membri. Naturalmente altri effetti introducono aberrazioni e "rumore" nel sistema; ma i canali normali tendono alla lunga a predominare. Non ci sono canali comparabili di interazione per simboli scritti: infatti, in base a questa stessa tesi standard, le sole interazioni semanticamente sensibili che i simboli scritti abbiano mai avuto sono con pensieri: per quanto essi tendano a esprimere delle verità, è perché esprimono credenze e, per quanto tendano a causare le loro proprie condizioni di accettabilità, è perché tendono a soddisfare i rispettivi desideri. Così, le sole interazioni semanticamente significative che i simboli scritti hanno con il mondo avvengono tramite i pensieri e, se l'idea funziona, ciò avviene perché l'intenzionalità è derivata.

Le interazioni che i pensieri hanno fra sé stessi (all'interno di un "sistema" singolo) sono particolarmente importanti, perché è in virtù di queste che il pensiero può essere sottile e indiretto, relativo alle sue interazioni con il mondo, cioè non facilmente ambiguo o incoerente. Così, nell'esprimere giudizi teniamo in considerazione qualcosa di più che la prova immediatamente attuale, e più delle opzioni immediatamente presenti nel fare progetti. Noi valutiamo quanto desideriamo, cerchiamo ulteriori informazioni, tentiamo realizzazioni per vedere se funzioneranno, formuliamo massime generali e leggi, soppesiamo risultati e costi, andiamo in biblioteca, cooperiamo, manipoliamo, schematizziamo, esaminiamo e riflettiamo su ciò che stiamo facendo. Tutte queste o sono o implicano un sacco di interazioni fra pensiero e pensiero, e tendiamo, a lungo andare, ad allargare e perfezionare la connessione e corrispondenza tra pensiero e mondo. Queste sono tipiche manifestazioni sia dell'intelligenza che dell'indipendenza.

Io dò per scontato che tutte le interazioni menzionate sono, in qualche senso, causali dal momento che è fra i "poteri causali" dei sistemi che si possono avere (organizzare, realizzare, produrre) pensieri che interagiscono col mondo e fra di loro. È difficile dire che questi sono i generi di poteri causali che Searle ha in mente, sia perché non lo dice, sia perché non sembrano così simili alla fotosintesi e alla lattazione. Ma, in ogni caso, essi mi sembrano essere forti candidati per il genere di potere che potrebbe distinguere i sistemi con l'intenzionalità — intenzionalità *originale* — da quelli senza. E la ragione è che questi sono i soli poteri che consistentemente riflettono il carattere distintamente intenzionale degli interattori:

precisamente, il loro “contenuto” o “significato” (anche se, per così dire, passivamente, come nel caso dei simboli scritti che vengono letti). Ciò significa che il potere di avere stati che sono semanticamente attivi è il “giusto” potere causale per l’intenzionalità. È questa tesi che sottostà all’affermazione (sufficientemente sviluppata) che i sistemi di IA possano veramente *essere* intelligenti e avere intenzionalità *originale.* Si può sicuramente affermare che le loro “rappresentazioni” sono semanticamente attive (o, almeno, che lo sarebbero se un sistema fosse costruito all’interno di un robot). Ricordiamo che noi concediamo loro almeno intenzionalità derivata, così gli stati in questione hanno veramente un contenuto, relativamente al quale possiamo giudicare della “appropriatezza semantica” delle loro interazioni causali. E la scoperta centrale di tutta la tecnologia del computer è che i dispositivi elettronici possono essere progettati tali che, relativamente a una certa interpretazione, certuni dei loro stati agiranno sempre causalmente in modi semanticamente appropriati, finché i congegni hanno le prestazioni loro attribuite dal disegno, cioè questi stati possono avere “canali normali” di interazione (l’uno con l’altro e con il mondo) più o meno comparabili a quelli che sottostanno all’attività semantica dei pensieri. Questo punto difficilmente può esser negato, finché è formulato nei termini dell’intenzionalità derivata dai sistemi di computazione: quello che sembra si debba aggiungere all’intenzionalità derivata, archetipica (e “inerte”) di un, diciamo, testo scritto, è precisamente l’attività semantica. Così, se l’attività semantica sufficientemente ricca è ciò che distingue l’intenzionalità originale da quella derivata (in altre parole, il “giusto” potere causale), allora sembra che i sistemi di computazione (sufficientemente ricchi) possano avere intenzionalità originale.

Ora, come Searle, sono incline a discutere questa conclusione: ma per ragioni completamente diverse. Non posso credere che ci sia confusione *concettuale* nel supporre che i giusti poteri causali per l’intenzionalità originale sono quelli che sarebbero catturati specificando un programma (cioè, una macchina virtuale). Quindi non penso che l’argomento della plausibilità precedentemente esposto possa essere messo da parte: né posso immaginare di essere convinto che, non importa a quale livello di qualità, la ricerca di IA sia giusta, ma che essa sia comunque ancora “debole” — cioè, che non abbia creato una intelligenza “reale” — perché ancora procede alla specificazione di programmi. Mi sembra che la questione interessante sia molto più empirica: supposto che i programmi *possano* essere il giusto modo per esprimere la fondamentale struttura causale, sono essi tali nei fatti? È a questo proposito che mi attendo un no. In altre parole, non mi curo molto del demone di Searle che lavora attraverso un programma per la perfetta simulazione di un cinese madrelingua — non perché tale demone non esiste, ma perché tale programma non esiste. O piuttosto, dal mio punto di vista le

questioni importanti sono se esiste un tale programma e, se non esiste, perché non esista.

Riduzionismo e religione

Douglas R. Hofstadter
Dipartimento di Computer Science, Indiana Univ., Bloomington, Ind. 47406

Questa diatriba religiosa contro l'IA, mascherata da serio argomento scientifico, è uno degli articoli più errati e indisponenti che io abbia mai letto nella mia vita. Nel suo potere di seccare è pareggiato solo dal famoso articolo *Menti, macchine e Gödel* di J.R. Lucas (1961).

Mi immedesimo facilmente nelle preoccupazioni di Searle. Come me, Searle ha profonde difficoltà nel percepire come mente, anima e "Io" possano uscire dal cervello, dalle cellule e dagli atomi. Per mostrare il suo imbarazzo, egli dà alcune belle parafrasi di questo mistero. Una delle mie favorite è la simulazione di un cervello mediante la tubatura dell'acqua. Va diritto al nocciolo del problema mente-corpo. La cosa strana è che Searle semplicemente neghi ogni possibilità di esser conscio a un tale sistema, con una manciata di "assurdo" (veramente penso che egli mal rappresenti la complessità di tale sistema sia ai lettori sia nella propria mente, ma questo è un problema in qualche modo a parte). Il fatto è che noi dobbiamo trattare con una realtà della natura: e talvolta le realtà della natura sono assurde. Chi avrebbe creduto che la luce consistesse di particelle prive di massa che ubbidiscono a un principio di incertezza mentre viaggiano attraverso un universo curvo a quattro dimensioni? Il fatto che intelligenza, capacità di comprensione, mente, consapevolezza, anima scaturiscano da una sorgente improbabile — un tessuto enormemente confuso di corpi cellulari, sinapsi e dendriti: è assurdo, eppure innegabile. Come questo possa creare un "Io" è difficile da capire, ma una volta che accettiamo quel fatto fondamentale, strano, disorientante, allora non dovrebbe sembrare più strano dare un "Io" a una tubatura d'acqua.

Il modo di Searle di trattare con questa realtà della natura è dichiarare che l'accetta: per poi non accettarne le conseguenze. La conseguenza principale è che la "intenzionalità" — il suo nome per anima — è un risultato di processi formali. Ammetto che ho evitato una ulteriore premessa a questo punto: cioè che i processi fisici sono formali, ovvero governati da regole. In altre parole, l'ulteriore premessa è che non c'è alcuna intenzionalità a livello di particelle (forse ho inteso male Searle. Può darsi che sia un mistico e che sostenga che c'è intenzionalità a quel livello. Ma poi come fa uno a spiegare perché si sente consapevole solo quando le particelle sono disposte in certe configurazioni speciali — di cervello — ma non, diciamo, in configurazioni a tubature di

ogni genere e misura?). L'unione di questi due pareri mi sembra ci costringa ad ammettere la possibilità di tutte le speranze dell'IA, nonostante il fatto che ci forzerà a pensare a noi stessi, in fondo, come a sistemi formali.

Per gente che non ha mai programmato, la distinzione tra i livelli di sistemi di elaborazione — programmi che girano su altri programmi o su hardware — è elusiva. Io credo che Searle non capisca realmente questa sottile idea, e così confonda molte distinzioni mentre ne crea altre artificiali per approfittare delle risposte emotive evocate nel processo di formulazione di idee non familiari.

Egli comincia con una situazione relativamente innocente: un uomo in una stanza con una serie di istruzioni in inglese per manipolare simboli cinesi. Dapprima si pensa che l'uomo risponda a delle domande (anche se non note a lui) riguardanti ristoranti, usando gli *scripts* di Schank. Poi Searle scivola casualmente nell'idea che questo programma possa superare il test di Turing! È un incredibile salto di complessità — forse con un aumento di milioni di livelli, se non di più. Searle non sembra essere consapevole di come l'intero quadro muti radicalmente solo nell'insinuare quella minima ipotesi. Ma perfino la situazione iniziale, che risulta abbastanza plausibile, è, in effetti, altamente irrealistica. Immaginiamo un essere umano che simula un complesso programma di IA, quale può essere un programma di "comprensione" basato su scripts. Digerire un'intera storia, esaminare le scritture e produrre una risposta, prenderebbe probabilmente una faticosa giornata di otto ore a un essere umano. In realtà, si suppone che questo programma simulato supererà il test di Turing; non soltanto che risponderà ad alcune domande stereotipe sui ristoranti. Quindi è necessario passare a una settimana per domanda, qualora il programma dovesse essere così complesso (siamo incredibilmente generosi con Searle).

Ora Searle vi chiede di identificarvi con questo povero schiavo umano (in realtà non vi chiede di identificarvi con lui: ma sa che vi proietterete in questa persona, e al suo posto proverete l'incubo indescrivibilmente seccante di quella simulazione). Egli sa che la reazione sarà: "Questo non è comprendere la storia; è solo una specie di processo formale!". Ma attenzione: ogni volta che qualche fenomeno è considerato su una scala di un milione di volte diversa dalla sua scala usuale, non appare lo stesso risultato! Se dovessi sentire il mio cervello correre cento volte troppo lentamente (naturalmente questo è paradossale, ma è ciò che Searle vuole che io faccia), non solo sarebbe molto angoscioso, ma è presumibile che non riconoscerei neppure le sensazioni. Getta dentro ancora un altro fattore fra i mille e una persona non potrebbe nemmeno immaginare come si sentirebbe. Ora questo è ciò che Searle sta facendo. Egli vi invita a identificarvi con un non umano che facilmente diviene umano e vi chiede così di partecipare a un falso ragionamento. Egli usa sempre di più questo astuto, emotivo espediente, per

far sì che si sia d'accordo con lui, che certamente un intricato sistema di tubi non può pensare! Egli dimentica di dirvi che una simulazione a tubi del cervello prenderebbe, diciamo, alcuni milioni di tubi con alcuni milioni di operai in piedi presso i rubinetti a girarli quando è necessario, e dimentica di dirvi che rispondere a una domanda richiederebbe un anno o due. Si dimentica di dirlo perché, se voi ricordaste ciò, e poi per conto vostro immaginaste di visionare un film o di accelerarlo un milione di volte, e immaginaste di cambiare il vostro livello di descrizione dell'oggetto dal livello di rubinetti a quello del centro del bocchettone di emissione e via attraverso una serie di livelli sempre più alti fino a raggiungere un eventuale livello simbolico, allora potreste dire: "Evviva, se immagino a che cosa somiglierebbe questo intero sistema se percepito a questa scala di tempi e a questo livello di descrizione, posso vedere come potrebbe essere consapevole, dopo tutto!".

Searle è il rappresentante di una categoria di persone che hanno un orrore istintivo per ogni spiegazione definitivamente esaustiva dell'anima. Non so perché certe persone abbiano questo orrore, mentre altre, come me, trovano nel riduzionismo la suprema religione. Forse la mia esperienza, dopo tutta una vita nella fisica e nella scienza in generale, mi ha dato un profondo rispetto per come gli oggetti o le esperienze più concrete e familiari svaniscono, quando ci si avvicina alla scala infinitesimale, in un etere misteriosamente astratto, una miriade di effimeri vortici turbinanti di attività matematica quasi incomprensibile. Ciò evoca in me una specie di rispetto cosmico. Per me, il riduzionismo non diminuisce, anzi aggiunge mistero. Io so che non è qui il posto per commenti filosofici e religiosi, eppure mi sembra che quello che Searle e io abbiamo, sia, al livello più profondo, un disaccordo di carattere religioso, e dubito che, qualunque cosa io dica, possa mai cambiare la sua opinione. Egli insiste su cose che chiama "proprietà causali intenzionali" che sembrano svanire non appena si analizzano, si trovano per loro delle regole, o si simulano. Ma che cosa tali cose siano, se non epifenomeni, o qualità "innocentemente emergenti", questo non lo so.

Fenomeni mentali e comportamento

B. Libet
Dipartimento di Fisiologia, Univ. di California, San Francisco, Calif. 94143

Searle afferma che il principale assunto del suo articolo è diretto a dimostrare la sua seconda asserzione, che "instanziare un programma di computer non è mai in sé una condizione sufficiente di intenzionalità" (cioè di uno stato mentale che include opinioni, desideri e intenzioni). Egli lo prova con un esperimento per mostrare che perfino "un agente umano potrebbe

instanziare il programma e ancora non avere la intenzionalità relativa": cioè, Searle mostra, in maniera magistrale e convincente, che il comportamento del computer appropriatamente programmato potrebbe verificarsi in assenza di uno stato mentale cognitivo. Credo che sia anche possibile stabilire tale tesi per mezzo di un assunto basato sulla semplice logica formale.

Partiamo sapendo che stiamo trattando con due diversi sistemi: il sistema A è il computer, col suo programma appropriato; il sistema B è l'essere umano, in particolare il suo cervello. Anche se il sistema A potesse essere programmato a comportarsi e perfino ad apparire come il sistema B, in un modo che potrebbe renderli tali da non distinguersi a un osservatore esterno, il sistema A deve essere almeno internamente diverso da B. Se A e B fossero identici, sarebbero entrambi esseri umani e non ci sarebbe alcun problema da discutere.

Accettiamo la proposta che, su una base input-output, il sistema A e il sistema B possano essere portati a comportarsi in modo simile, proprietà che possiamo raggruppare insieme sotto la categoria X. Il possesso degli stati mentali relativi (includendovi il comprendere, le opinioni, i desideri, le intenzioni e simili) può essere chiamato proprietà Y. Sappiamo che il sistema B ha la proprietà Y. Ricordando che i sistemi A e B sono notoriamente differenti, è un errore di logica sostenere che, poiché i sistemi A e B hanno entrambi la proprietà X, debbano anche avere entrambi la proprietà Y.

Quanto detto precedentemente conduce a una dichiarazione più generale: che nessun comportamento di computer, prescindendo da quanto riesca a simulare il comportamento umano, è mai sufficiente prova di per sé di alcuno stato mentale. In effetti, Searle sembra sostenere questo caso più generale quando nella discussione osserva: a) far sentire ai computers dolore o innamoramento non sarebbe né più difficile né più facile che far loro avere cognizione; b) per la simulazione, quello che è necessario è il giusto input e output e un programma che trasformi il primo nel secondo; c) confondere la simulazione con la duplicazione è lo stesso errore, sia che si tratti di dolore, amore o cognizione.

D'altro lato, Searle sembra non mantenere questa asserzione generale con coerenza. Nella sua discussione sulla "replica della combinazione" (al suo esempio analitico o esperimento di pensiero), Searle dichiara: "Se potessimo costruire un robot il cui comportamento non si distinguesse dal comportamento umano, troveremmo razionale e davvero irresistibile attribuirgli intenzionalità, eccezion fatta per qualche ragione contraria". Sulla base del mio assunto, non si dovrebbe sapere che il robot ha un programma formale che spiega bene il suo comportamento, al fine di non dover attribuire l'intenzionalità. Tutto quel che dobbiamo sapere è che l'apparato di controllo interno del robot non è fatto nello stesso modo e dello stesso materiale del cervello umano, per rifiutare la tesi che il robot deve possedere gli stati

mentali dell'intenzionalità.

Ora è vero che né la mia tesi né quella di Searle escludono la possibilità che un computer, appropriatamente programmato, *possa* anche avere stati mentali (proprietà Y); la tesi afferma semplicemente che non è appropriato sostenere che il robot *debba* avere stati mentali (Y). Searle comunque contribuisce a fornire un'apprezzabile analisi del perché tanta gente ha creduto che i programmi di computer attribuiscano un tipo di processi o stati mentali al computer. Searle nota che, tra altri fattori, un residuo comportamentismo o operazionalismo soggiace alla volontà di accettare modelli input-output come sufficienti per postulare stati mentali umani in computers appropriatamente programmati. Vorrei aggiungere che ci sono ancora molti psicologi e forse filosofi che sono ugualmente gravati di comportamentismo residuo o operazionalismo anche quando trattano i criteri per l'esistenza di una consapevole esperienza soggettiva in soggetti umani (cfr. Libet 1973, 1979).

La risposta funzionalista

William G. Lycan
Dipartimento di Filosofia, Ohio State Univ., Columbus, Ohio 43210

La maggior parte delle versioni di comportamentismo filosofico hanno avuto la conseguenza che, se un organismo o dispositivo D supera il test di Turing, nel senso di manifestare sistematicamente tutte le stesse disposizioni comportamentali esterne di un normale essere umano, D ha tutti gli stessi tipi di stati intenzionali degli esseri umani; alla luce di controesempi abbastanza ovvi a questa tesi, i filosofi materialisti della mente hanno rifiutato il comportamentismo in favore di una tesi più sciovinistica: il fatto che D manifesti tutti gli stessi tipi di comportamento che ci sono propri, non basta, da solo, perché D abbia stati intenzionali: è necessario in aggiunta che D produca un comportamento in seguito a stimoli pressapoco nel modo che facciamo noi, che l'organizzazione funzionale interna di D sia non dissimile dalla nostra e che D sviluppi lo stimolo input con procedimenti interni analoghi. In base a questa teoria "funzionalista", essere in uno stato mentale di tale genere significa incorporare un componente funzionale o sistema di componenti di questo tipo che è in un certo stato che si distingue in sé e per sé. I "componenti funzionali" sono individuati secondo i ruoli che giocano all'interno della loro generale organizzazione funzionale.[12]

Searle illustra un certo numero di casi di entità con comportamenti che noi associamo agli stati intenzionali, ma che abbastanza chiaramente non hanno tali stati.[13] Accetto i suoi giudizi intuitivi sulla maggior parte di questi casi. Searle più un manuale di istruzioni più carta e matita presumibilmente non capisce il cinese, né lo capisce memorizzando le regole o munendosi di una

telecamera, e non lo capisce nemmeno il robot con Searle dentro. Né il mio stomaco, né il fegato di Searle, né un termostato, né un interruttore hanno opinioni e desideri. Ma nessuno di questi casi è un controesempio alla ipotesi funzionalista. I sistemi appena citati non sono funzionalmente isomorfi al relativo livello degli esseri umani che comprendono il cinese: un cinese impegnato in una conversazione usa procedimenti suoi, non quelli propri di un modello rappresentato dal personaggio Searle, di lingua inglese, di cultura americana. Perciò quelli non sono controesempi di una teoria funzionalista della comprensione del linguaggio, e lasciano aperta la possibilità che un computer funzionalmente isomorfico a un individuo di madrelingua cinese possa capire effettivamente pure il cinese. Lo stomaco, i termostati e simili, a causa della loro semplicità bruta, sono addirittura più chiaramente dissimili dagli umani (lo stesso è presumibilmente vero per i programmi di comprensione del linguaggio di Schank).

Io spero in una versione sofisticata del simulatore di cervello (o della macchina a combinazioni) che Searle illustra col suo esempio dell'idraulico. Immaginiamo un sistema idraulico che replica, forse non la precisa neuroanatomia di un cinese, ma tutto quello che è relativo all'organizzazione funzionale più elevata del cinese: condutture dell'acqua individuali sono raggruppate in sistemi analoghi a quelli trovati nel cervello, e il congegno sviluppa un input linguistico proprio allo stesso modo di chi parla cinese (non simula o descrive semplicemente questo processo). Il sistema inoltre è automatico e fa tutto senza l'intervento di Searle o di alcun altro *deus in machina.* A queste condizioni e dato un contesto sociale adatto, penso che sarebbe plausibile accettare la conseguenza funzionalista che il sistema idraulico comprende il cinese. L'articolo di Searle suggerisce due obiezioni a questa tesi.

Primo, "dov'è la capacità di comprendere all'interno di questo sistema?" Tutto quello che Searle vede sono tubi e valvole e acqua che scorre. Risposta: se guardiamo intorno ai dettagli della struttura del sistema, tu sei troppo piccolo per vedere che il sistema comprende le frasi cinesi. Se tu fossi un osservatore minuscolo, a dimensione di cellula, all'interno di un cervello di un nativo cinese, tutto quello che vedresti sarebbero neuroni, stupidamente, meccanicamente trasmettenti cariche elettriche, e nello stesso tono chiederesti: "Dov'è la capacità di comprendere in questo sistema?". Ma sbaglieresti a concludere che il sistema che stavi osservando non comprende il cinese: in modo uguale tu puoi sbagliare sul congegno idraulico.[14]

Secondo, anche se un computer dovesse replicare tutta l'organizzazione funzionale relativa al cinese madrelingua, tutto quello che il computer sta facendo realmente è eseguire operazioni di calcolo su elementi specificati formalmente. Un elemento caratterizzato solo formalmente o sintatticamente non ha significato o contenuto in sé, ovviamente, e nessuna manipolazione

sintattica lo doterà di significato. Risposta: la premessa è corretta e mostra che nessun computer ha o potrebbe avere stati intenzionali semplicemente per il fatto che esegue operazioni sintattiche su elementi caratterizzati formalmente. Ma questo non basta a provare che nessun computer può avere stati intenzionali. Gli stati del nostro cervello non hanno dei contenuti solo per il fatto di avere proprietà puramente formali.[15] Uno stato mentale descritto "sintatticamente" non ha significato o contenuto di per sé: in virtù di che cosa, allora, lo acquisisce? Secondo una recente teoria, il contenuto di una rappresentazione mentale non è determinato dall'interno (cfr. Putnam 1975a; Fodor 1980); piuttosto, è determinato in parte dagli oggetti dell'ambiente che attualmente figurano nell'eziologia della rappresentazione e, in parte, da fattori sociali e contestuali di diversi altri tipi (Stich, in preparazione). Ora, i computers di oggi vivono in ambienti altamente artificiali e deprimenti. Essi ricevono un input accuratamente e appositamente preselezionato: il loro software è manipolato da programmatori sconsiderati che sono isolati in laboratori e uffici, privi di ogni normale interazione con una cornice sociale naturale o appropriata.[16] Per questa ragione e per molte altre, Searle è sicuramente nel giusto quando dice che i computers attuali non hanno gli stati intenzionali che noi tendiamo ad attribuire loro. Ma nulla di ciò che Searle ha detto inficia la tesi che se un sofisticato computer del futuro non solo replicasse l'organizzazione funzionale umana ma avesse rappresentazioni interne come risultato del giusto tipo di storia causale e fosse collocato entro una favorevole cornice sociale, noi potremmo correttamente attribuirgli stati intenzionali. Questo punto può provocare, o meno, un conforto permanente alla comunità di IA.

Convinzioni, macchine e teorie

John McCarthy
Laboratorio di Intelligenza Artificiale, Stanford Univ., Stanford, Calif. 94305

John Searle confuta la replica di Berkeley ("Il sistema capisce il cinese") sostenendo che una persona (chiamiamolo Mr. Hyde) compie nella sua testa un processo (chiamiamolo Dr. Jekyll) per effettuare una conversazione scritta in cinese: Mr. Hyde non capisce il cinese. E tutti sembrano d'accordo con Searle. Ma io controbatterei, e suppongo che anche gli interlocutori di Berkeley lo farebbero, che date certe condizioni per la comprensione, il Dr. Jekyll capisce il cinese. Nella storia di Robert Louis Stevenson, Dr. Jekyll e Mr. Hyde si alternano nell'uso dello stesso corpo, mentre nel caso di Searle uno interpreta un programma specificando l'altro.

Il rifiuto di Searle dell'idea che ai termostati possano essere attribuite convinzioni è basato su un malinteso: non c'è una nozione panteistica che

tutte le macchine, inclusi telefoni, interruttori elettrici e calcolatori, abbiano convinzioni. Una convinzione può utilmente essere ascritta solo a sistemi intorno ai quali il bagaglio di nozioni di qualcuno può al meglio essere espresso con l'attribuire opinioni che soddisfano assiomi come quelli di McCarthy (1979). I termostati sono talvolta come tali sistemi. Dire a un bambino: "Se tieni la candela sotto il termostato, saresti sciocco a pensare che il meccanismo di accensione si spenga perché la stanza è troppo calda" fa un uso corretto del repertorio di concetti intenzionali del bambino. Formalizzare una convinzione richiede di saper trattare casi semplici quanto quelli più complicati. Attribuire convinzioni ai termostati è analogo all'includere 0 e 1 nel sistema numerico anche se non avremmo bisogno di un sistema numerico per trattare l'insieme vuoto o insiemi con solo un elemento: in effetti non avremmo neppure bisogno del concetto di insieme.

Tuttavia, un programma che comprende, non dovrebbe essere considerato come una teoria della capacità di comprendere più di quanto un uomo che comprende sia considerato una teoria. Un programma può solo essere un'illustrazione di una teoria, e una teoria utile conterrà molto di più dell'asserzione che "il seguente programma comprende quel che si riferisce ai ristoranti". Non so decidere se quest'ultimo difetto si applica a Searle o solo ad alcuni dei ricercatori che egli critica.

Intelligenza Artificiale: la cosa reale

John C. Marshall
Unità neuropsicologica, Dipartimento di Neurologia Clinica, Ospedale Radcliffe, Oxford, Inghilterra

Searle vorrebbe convincerci che coloro che attualmente popolano rispettabili università americane hanno soggiaciuto al sogno faustiano (Mefistofele: "Che cosa è allora?" Wagner: "Un uomo è sulla via di essere fatto"). Egli ci assicura, guardandoci dritto negli occhi, che alcuni studiosi contemporanei pensano che un "computer appropriatamente programmato è realmente una mente" e che tali creature artificiali hanno "letteralmente stati cognitivi". Ma chi lo crede davvero? Voglio dire, perché non dare allora una decente sepoltura al proprio IBM ormai obsoleto e recitare il De Profundis in sua memoria? E anche se certa gente a Yale, Berkeley, Stanford e così via organizza queste strane correnti di opinione, quale interesse *scientifico* potrebbe avere tutto ciò? Ma immaginiamo pure che essi abbiano ragione e che i loro computers realmente percepiscano, comprendano e pensino. Tutto quello che i nostri produttori di Golem hanno fatto sulla storia di Searle è creare un'altra mente ancora. Se il solo scopo è di "riprodurre" (termine di Searle, non mio) fenomeni mentali, non c'è certamente bisogno di comperare

un computer.

Francamente, non mi interessa proprio quel che alcuni membri della comunità di IA pensano sullo stato ontologico delle loro creazioni. Quello che mi interessa è se qualcuno possa produrre ben fondate spiegazioni rivelanti la percezione di musica tonale (Longuet-Higgins 1979) e le proprietà della stereovisione (Marr e Poggio 1979) e l'analisi delle frasi di lingua naturale (Thorne 1968). Tutti quelli che io conosco e che maneggiano computers, lo fanno perché hanno un'attraente teoria di una certa capacità psicologica e desiderano esplorare algoritmicamente certe conseguenze della teoria. Searle si riferisce a tale attività come relativa all'ipotesi dell'IA debole, ma io avrei pensato che la costruzione della teoria e la verifica fosse una delle iniziative più forti che uno scienziato si può permettere. Chiaramente, ci deve essere qui un malinteso radicale.

Il problema sembra stia nello strano uso del termine "teoria" che fa Searle (o chi lo informa sulla IA). Così scrive Searle: "Secondo l'ipotesi dell'IA forte, i computers appropriatamente programmati hanno letteralmente stati cognitivi, e perciò i programmi sono teorie psicologiche". Ignorando per il momento quel "e perciò", che introduce un *non sequitur*, come potrebbe un programma essere una teoria? Come Moor (1978) mette in evidenza, una teoria è, quanto meno, una collezione di proposizioni correlate che possono essere vere o false, laddove un programma è (o era) una serie di cartelle ordinate e forate. Per quello che so io, forse i computers "letteralmente" non hanno stati cognitivi, ma anche se li avessero, che cosa autorizzerebbe la conclusione che il programma di per sé sia una teoria psicologica? Che cosa direbbe uno di un'analoga affermazione applicata alla fisica invece che alla psicologia? "Computers appropriatamente programmati hanno letteralmente stati fisici, e perciò i programmi rappresentano una teoria della materia". Tale idea non suona come una conclusione valida, a mio parere. L'esposizione di Moor della distinzione tra programma e teoria è particolarmente chiara e degna di essere citata nei dettagli.

> Un programma deve essere interpretato al fine di generare una teoria. Nel corso della interpretazione è probabile che una parte del programma sia messo da parte come irrilevante dal momento che sarà dedicato ai particolari tecnici che rendono il programma accettabile al computer. In più, le restanti parti del programma devono essere organizzate in qualche maniera coerente con un'ampia serie di programmi di computer tesi a rappresentare processi specifici. Estrarre una teoria dal programma non è una cosa semplice, poiché diversi aspetti del programma possono generare diverse teorie. Perciò, allo stesso modo in cui il programma, inteso come modello, incorpora una teoria, esso può altrettanto bene incorporare molte teorie (Moor 1978, p. 215).

Searle riferisce che alcuni suoi referenti credono che i programmi siano altre menti, sebbene artificiali; se le cose stessero così, non potrebbero, questi studiosi, tentare di costruire teorie di menti artificiali, proprio come noi lo

facciamo per quelle naturali? Considerevole confusione sorge allora quando gli informatori di Searle ignorano la loro propria tesi e usano i termini "riprodurre" e "spiegare" come sinonimi: "Il progetto è quello di riprodurre e spiegare la mente disegnando programmi". È facile constatare come questo è indubbiamente deviante, in quanto proietta tale tesi dall'ambito dei computer agli individui umani. Così ho notato che molti degli stati mentali di mia figlia hanno una marcata rassomiglianza con i miei; ciò è avvenuto, senza dubbio, perché parte del mio piano genetico è stato usato per costruire la sua struttura e perché ho preso parte alla responsabilità di "programmarla". Tutto bene, ma sarebbe pura credulità considerare mia figlia come colei che mi "spiega", cioè come una "teoria" di me.

Quello che si vorrebbe piuttosto è una delucidazione del senso per cui programmi, computers e altre macchine figurano o no nella spiegazione del comportamento (Cummins 1977, Marshall 1977). È un peccato che Searle sottovaluti tali questioni nella discussione sull'uso quotidiano del vocabolario mentale, impresa che è meglio lasciare ai lessicologi. Searle scrive: "Lo studio della mente comincia da questo fatto: gli esseri umani hanno opinioni; mentre termostati, telefoni e macchine calcolatrici, no". Bene, forse inizia qui, ma non è una ragione per supporre che debba finire qui. Come si trasformerebbe tale "tesi" in filosofia naturale? "Lo studio della fisica comincia da questi fatti: le tavole sono oggetti solidi senza buchi, mentre il formaggio Gruyère…". In questo caso, Searle potrebbe ancora dire: "Se ottieni una teoria che nega questo punto, hai prodotto un controesempio alla teoria e la teoria è errata"? Naturalmente l'"opinione" di un termostato che la temperatura dovrebbe essere un po' più alta non è lo stesso della mia opinione su questo punto. Sarebbe del tutto irrilevante se fossero la stessa cosa. Certo chi teorizza la relazione tra le due opinioni, sta scavando verso un parallelo più profondo: ha visto un'analogia che può illuminare certi aspetti del controllo e della regolazione di sistemi complessi. La nozione di reazione positiva e negativa è quella che rende i termostati così attraenti per Alfred Wallace e Charles Darwin, per Claude Bernard e Walter Cannon, per Norbert Wiener e Kenneth Craig. L'osservazione di regolatori e termostati li ha posti in grado di vedere, al di là delle apparenze, a un livello in cui ci sono profonde affinità tra animali e manufatti (Marshall 1977).

È Searle, non lo studioso, che non prende veramente sul serio l'impresa. Secondo Searle, "quello che vogliamo sapere è ciò che distingue la mente dai termostati e dai fegati". Sì, ma non è tutto; noi vogliamo anche sapere a che livelli di descrizione ci sono forti rassomiglianze tra fenomeni disparati.

Nei paragrafi iniziali del *Leviathan,* Thomas Hobbes (1651, p. 8) dà chiara espressione alla filosofia meccanicistica:

> La natura, l'arte per mezzo della quale Dio ha fatto e governa il mondo, è imitata, come molte altre

> cose, dall'arte dell'uomo: egli infatti può così costruire un animale artificiale… Perché il cuore non è che una molla, e i nervi tante corde, e le giunture tante ruote che danno moto all'intero corpo, così come fu inteso dall'inventore supremo.

Che cosa è la nozione di "imitazione" che Hobbes usa qui? Ovviamente non l'idea dell'*esatta* imitazione o *copia.* Nessuno confonderebbe un nervo con un pezzo di corda, il cuore con la molla principale di un orologio, o la caviglia con una ruota. Non c'è ragione di ingannare l'occhio. I lavori dello scienziato non sono in quel senso *riproduzioni* della natura: piuttosto sono tentativi di vedere dietro al mondo fenomenologico una realtà nascosta. Fu Galileo, naturalmente, che articolò con molta forza questo paradigma: la scultura, nota Galileo, sembra "più vicina alla natura" che la pittura per il fatto che il sostrato materiale manipolato dallo scultore condivide con la materia manipolata dalla natura la qualità della tridimensionalità. Ma davvero questo fatto torna a vantaggio della scultura? Al contrario, dice Galileo, esso "diminuisce il suo merito" grandemente: "Che cosa ci sarà di così meravigioso nel fatto che la scultura stessa imiti la scultrice Natura?" E conclude: "L'imitazione più artistica è quella che rappresenta il tridimensionale col suo opposto, che è il piano" (Panofsky 1954, p. 97). Galileo riassume così la sua posizione: "Più lontano è il mezzo dell'imitazione dalle cose da imitare, più degna di ammirazione l'imitazione sarà" (Panofsky 1954). In una nota al passo, Panofsky sottolinea "l'affinità di base tra lo spirito di questa sentenza e l'illimitata ammirazione per Aristarco e Copernico 'perché essi si fidarono della ragione piuttosto che dell'esperienza sensoria'" (Panofsky 1954).

Ora Searle ha perfettamente ragione nel mettere in evidenza che in IA si cerca di modellare stati cognitivi e le loro conseguenze (la cosa reale) attraverso l'uso di una sintassi formale, l'interpretazione della quale esiste solo all'occhio dell'osservatore. Precisamente qui sta la bellezza e il significato dell'impresa, nel cercare di fornire una controparte a ogni reale distinzione mediante una distinzione sintattica. Questo è essenzialmente considerare lo studio delle relazioni fra transazioni fisiche e operazioni simboliche come un saggio in criptanalisi (Freud 1895; Cummins 1977). Sorge allora l'interessante questione se c'è un'unica progettazione per quanto riguarda gli elementi formali del sistema e i loro "significati" (Householder 1962). Searle, comunque, sembra suggerire che stiamo abbandonando completamente il metodo galileiano e quello "linguistico" al fine di copiare semplicemente cognizioni. Evidentemente vorrebbe farci cercare la mente solo nei "neuroni con neuroassoni e dendriti", sebbene ammetta, come possibilità empirica, che tali oggetti potrebbero "produrre consapevolezza, intenzionalità e tutto il resto usando tipi di principi chimici diversi da quelli usati dagli uomini". Ma questa ammissione rinuncia all'intero gioco. Come

potrebbe Searle costruire una mente a base di silicio (piuttosto che la nostra mente a base di carbonio), se non avendo un'appropriata e *astratta* (cioè, non materiale) caratterizzazione di quello che le due forme di vita hanno in comune? Searle risolve abilmente questo problema semplicemente "attribuendo" stati cognitivi a sé stesso, ad altra gente, e a una varietà di animali domestici: "Nelle 'scienze cognitive' si presuppone la realtà e conoscibilità del mentale allo stesso modo che nelle scienze fisiche si deve presupporre la realtà e conoscibilità degli oggetti fisici". Ma tutto ciò non funziona davvero: noi siamo, dopo tutto, ben lontani dall'avere alcuna prova convincente che cani e gatti abbiano "stati cognitivi" come l'uso del termine da parte di Searle farebbe supporre (cfr. *Cognition and Consciousness in Nonhuman Species*, BBS 1(4) 1978).

Thomas Huxley (1874, p. 156) pone la questione nella sua parafrasi della frase cartesiana ortodossa di Nicholas Malebranche: "Che prova c'è che i bruti sono altro da una razza superiore di marionette, che mangiano senza piacere, piangono senza dolore, non desiderano nulla, non conoscono nulla, e soltanto simulano intelligenza come un'ape simula un matematico?". L'amico e corrispondente di Descartes, Marin Mersenne, aveva pochi dubbi sulla risposta a questo genere di domanda; nella sua discussione sulle capacità percettive degli animali egli direttamente nega la intenzionalità alle bestie:

> Gli animali non hanno conoscenza di questi suoni, ma solo una rappresentazione, e non sanno se quello che apprendono è un suono o un colore o qualcosa di diverso, così si può dire che essi non agiscono come ci si aspetta da loro, e che gli oggetti fanno un'impressione sui loro sensi, da cui necessariamente segue la loro azione, come le rotelle di un orologio necessariamente seguono il peso o la molla che le guida (Mersenne 1636).

Per Mersenne, allora, il programma all'interno degli animali è un *calculus* non interpretato, una sintassi senza una semantica (cfr. Fodor, *Methodological Solipsism*, BBS 3(1) 1980). Searle, d'altro lato, sembra credere che scimmie e cani hanno realmente stati mentali perché "sono fatti di una materia simile a noi" e hanno occhi, naso e pelle. Non mi riesce di vedere come il dato sostenga la conclusione. Si dovrebbe pensare che qualche ragionamento intricato e una sottile sperimentazione sia necessaria per giustificare l'attribuzione dell'intenzionalità agli scimpanzé (Marshall 1971; Woodruff e Premack 1979). Che gli scimpanzé somiglino del tutto a noi è un fatto abbastanza debole per costruirci sopra una conclusione importante.

Quando Jacques de Vaucanson — il più grande teorico di IA — ebbe completato la sua anitra artificiale, la mostrò, in tutta la sua nuda gloria di legno, corda, acciaio e fili. Per quanto il suo pubblico possa avere preferito una creatura più morbida e carezzevole, Vaucanson ha fermamente resistito alla tentazione di accontentarlo.

> Forse certe signore, o certa gente, che ama solo l'esterno degli animali, avrebbe preferito vedere il tutto coperto: cioè *l'Anitra* con le *penne.* Ma, oltre al fatto che io ho desiderato rendere ogni cosa evidente, non avrei voluto che si pensasse di me che mi volevo imporre agli spettatori per mezzo di uno stratagemma segreto o da giocoliere (Fryer e Marshall, 1979).

Per Vaucanson, la *cosa reale* è la teoria che ha incorporata nel modello anitra.

Riconoscimenti

Ringrazio il Dr. J. Loew per i suoi commenti a versioni precedenti di questo lavoro.

Intenzionalità: hardware, non software

Grover Maxwell
Centro di Filosofia della Scienza, Univ. del Minnesota, Minneapolis, Minn. 56455

È un raro e piacevole privilegio commentare un articolo che è sicuramente destinato a diventare, quasi immediatamente, un classico. Ma, ahimè, perché si richiedono i commenti? Seguendo le istruzioni, resisterò alla fortissima tentazione di spiegare, come fa Searle, esattamente i giusti punti centrali sostenendoli con i giusti argomenti; lo lascerò fare a quelli che, per una ragione o per un'altra, ancora dissentono con la sua precisa volontà di richiamare l'attenzione su alcune possibili debolezze, anche se con un errore o due, nel trattare alcuni dei suoi assunti ausiliari. Quello che tenterò di fare sarà di esaminare brevemente — e perciò in maniera approssimativa e inadeguata — quelle che mi sembrano alcune implicazioni dei suoi risultati per il più generale problema del rapporto mente-corpo. Molto prudentemente, proprio per la brevità del suo articolo, Searle lascia totalmente intatte alcune questioni centrali concernenti le relazioni mente-cervello. In particolare, il suo punto d'attacco principale sembra compatibile con l'interazionismo, l'epifenomenismo, e con almeno alcune versioni della tesi d'identità. Egli ragiona pesantemente contro il materialismo eliminativo, e, in modo ugualmente importante, rivela il "funzionalismo" (o "materialismo funzionale") come è comunemente ritenuto e interpretato (per esempio da Hilary Putnam e David Lewis) per essere appunto un'altra varietà di materialismo eliminativo. Searle correttamente annota che un funzionalismo di questo genere (e IA forte, in generale) è una specie di dualismo. Ma non è un dualismo mentale-fisico: è un dualismo forma-contenuto, tale, per giunta, per cui la forma è la realtà e il contenuto non importa! (cfr. Fodor, *Methodological Solipsism*, BBS 3(1) 1980).

Ora devo ammettere che al fine di trovare queste implicazioni nei risultati di Searle, ho letto in essi un po' di più di quello che contengono esplicitamente. Nello specifico, ho dato per scontato che gli stati intenzionali sono genuinamente mentali nel senso del "mentale" di Nagel (1974) secondo quello che suppongo sia ovvio: che il materialismo eliminativo, cioè, cerca di "eliminare" il genuinamente mentale. Ma a me sembra che non sia necessario leggere tra le righe per vedere che Searle è in consonanza con le mie supposizioni. Per esempio, egli parla in effetti di "sistemi genuinamente mentali" e dice (Searle 1979c) di credere che "solo gli esseri capaci di stati *consci* sono capaci di stati intenzionali", sebbene dica che non sa come dimostrare tale tesi (come si potrebbe infatti dimostrarlo? Come si potrebbe dimostrare che il fuoco brucia?). La prova che Searle dà per la conclusione che solo le macchine possono pensare (possono avere stati intenzionali),

sembra abbia due premesse nascoste: 1) gli stati intenzionali devono sempre essere prodotti causalmente, e 2) ogni rete causale (con un certo grado di organizzazione e completezza, o alcune condizioni simili) è una macchina. Accetto per gli scopi di questo commento le sue premesse e la sua conclusione. Poi voglio chiedere: quale genere di hardware deve incorporare una macchina pensante? (per "macchina pensante" intendo naturalmente una macchina che abbia pensieri genuinamente mentali: una tale macchina, ripeto, avrà anche stati genuinamente mentali o eventi che organizzano sensazioni, emozioni, e così via in tutta la loro ricchezza soggettiva, qualitativa, consapevole, esperienziale). Per continuare questa linea di ricerca voglio impiegare una "ontologia dell'evento" scartando comunque la metafisica della sostanza (Maxwell (1978) fornisce uno schizzo di alcuni dei dettagli e dei punti di contestazione che la fisica contemporanea, in modo del tutto indipendente dalla filosofia della mente, apporta a una tale ontologia). Un evento è qualcosa di simile alla presentazione di una proprietà o alla realizzazione concreta di uno stato. Una rete causale, allora, consiste interamente in un gruppo di eventi e nei collegamenti causali che li interconnettono. *A fortiori*, la nostra "macchina" consisterà interamente di eventi e connessioni causali. In altre parole, l'hardware di questa macchina (o di ogni altra macchina, per esempio di un frigorifero) consiste nei suoi eventi costituenti e la macchina non consiste di nessun'altra cosa (tranne dei collegamenti causali). La nostra macchina pensante nella sola forma che conosciamo oggi è sempre un cervello (o, se preferite, un intero corpo umano o di altro animale) che, come abbiamo spiegato, presenta solo una certa rete causale di eventi. La teoria dell'identità mente-cervello, nella versione che io difendo, dice che alcuni degli eventi in questa rete non sono altro che eventi genuinamente mentali (esempi di stati intenzionali, o di dolore, o simili). L'epifenomenismo dice che gli eventi mentali "dipendono" dalla rete principale (il cervello) attraverso connessioni causali che si svolgono sempre in una sola direzione (l'epifenomenismo è, credo ovviamente, sebbene contingentemente, falso). L'interazionismo dice che ci sono connessioni causali in entrambe le direzioni, ma che gli eventi mentali sono in qualche modo in un ambito diverso dagli eventi del cervello.

Supponendo che Searle accetti l'ontologia dell'evento, se non altro per amore di discussione, egli potrebbe sostenere che gli eventi mentali, in generale, e le istanze degli stati intenzionali, in particolare, sono parti della macchina, oppure la sua posizione è piuttosto che essi sono appunto prodotti della macchina? Cioè: Searle sarebbe incline ad accettare la tesi di identità o piegherebbe piuttosto verso l'epifenomenismo o l'interazionismo? A mio avviso, in tale contesto, la teoria dell'identità sembra di gran lunga l'ipotesi più plausibile, elegante ed economica. Per essere sicura, essa deve affrontare problemi seri e, allo stato attuale, completamente irrisolti come l'"obiezione

del granello” (cfr., per esempio, Maxwell 1978) e la tendenza verso il panpsichismo (cfr., per esempio, Popper e Eccles 1977) ma credo che l’epifenomenismo e l’interazionismo abbiano di fronte difficoltà anche più severe.

Prima di procedere dovrei dare importanza al fatto che la conoscenza scientifica contemporanea non solo ci conduce a un’ontologia dell’evento, ma prova anche la falsità del realismo ingenuo e “gentilmente persuade” ad accettare quello che ho (in qualche modo erroneamente) chiamato “realismo strutturale”. Conformemente a questo, virtualmente tutta la nostra conoscenza del mondo fisico è conoscenza della struttura (includendo la struttura spaziotempo) delle reti causali che lo costituiscono (cfr., per esempio, Russell 1948 e Maxwell 1976). Questo vale pienamente per quanto riguarda la conoscenza del cervello (eccetto per uno specialissimo genere di conoscenza, che discuteremo presto). Noi siamo perciò ignoranti per quanto riguarda le proprietà intrinseche (non strutturali) della “materia” (o di ciò che la fisica contemporanea lascia di essa); in particolare, se soltanto conoscessimo di più la neurofisiologia, conosceremmo la struttura dell’immensa rete causale che costituisce il cervello, ma non conosceremmo il suo contenuto: cioè, non conosceremmo ancora che cosa sono tali eventi costituenti. La teoria dell’identità va un passo più in là e considera che alcuni di questi eventi sono appunto istanze dei nostri stati intenzionali, delle nostre sensazioni, delle nostre emozioni e così via, in tutta la loro ricchezza genuinamente mentale, come sono conosciute “per esperienza”. Questa è la “specialissima conoscenza” menzionata sopra, e se la teoria dell’identità è vera, è conoscenza di quello che sono alcuni (probabilmente solo una ristrettissima serie) degli eventi che costituiscono il cervello. In questa limitata serie di eventi noi conosciamo proprietà intrinseche e insieme strutturali.

Ritorniamo a una delle domande poste da Searle: “Potrebbe un manufatto, una macchina fatta come un uomo, pensare?”. La risposta che egli dà è, penso, la migliore possibile, dato il nostro stato presente di misurata ignoranza nelle scienze neurologiche, ma vorrei elaborarla un po’ di più. Poiché ho sostenuto prima che i pensieri e gli altri eventi mentali sono parte dell’hardware delle “macchine pensanti”, tale hardware deve in qualche modo essere entrato in ognuna di tali macchine che costruiamo. Ora non abbiamo la più vaga idea di come questo possa essere fatto. La migliore scommessa potrebbe essere, come indica Searle, di “costruire” una macchina (da un protoplasma) con neuroni, dendriti e assoni come i nostri, e poi sperare che, dall’hardware iniziale, venisse misteriosamente generato “hardware mentale”. Ma questa scommessa mi sembra estremamente implausibile. Comunque, non concludo che la costruzione di una macchina pensante sia impossibile. Concludo, piuttosto, che dobbiamo imparare molto di più sulla fisica, sulla neurofisiologia, neuropsicologia, neuropsicofisiologia e così via: non solo più

dettagli, ma molto di più sui veri fondamenti della nostra conoscenza teoretica in questi ambiti, prima che possiamo perfino solo riflettere sensatamente sulla costruzione delle macchine pensanti (ho mostrato in Maxwell 1978 che i fondamenti della fisica contemporanea sono in un tale stato che dovremmo sperare in mutamenti veramente rivoluzionari nelle teorie fisiche e che tali mutamenti possono aiutare immensamente a risolvere i problemi mente-cervello, e che riflessioni in neurofisiologia e forse anche in psicologia possono benissimo fornire spunti proficui per il fisico nel suo rinnovamento dei fondamenti della teoria spazio-temporale). In tale situazione Searle ha mostrato la totale futilità della direzione su cui si muove l'IA forte rispetto all'Intelligenza Artificiale genuina.

La penna è più potente del computer?

E.W. Menzel Jr.
Dipartimento di Psicologia, State Univ. of New York at Stony Brook, Stony Brook, NY 11794

L'area dell'IA differisce da quella dell'intelligenza naturale in almeno tre aspetti. Primo, in IA si è per forza limitati all'uso di dati comportamentali formalizzati o "output" come base per fare inferenze sui propri argomenti (la situazione non è diversa, comunque, nei settori della storia e dell'archeologia). Secondo, per convenzione, in IA si deve assumere, fino a prova contraria, che un argomento non ha più mentalità di una roccia; e inoltre che, fino a prova contraria, un argomento può essere considerato come soggetto alla sensibilità. Terzo, in IA l'analisi è ordinariamente limitata a questioni riguardanti la "struttura" dell'intelligenza, mentre un'analisi completa dell'intelligenza naturale deve anche considerare questioni di funzione, sviluppo ed evoluzione.

Per altri aspetti, comunque, mi sembra che i problemi dell'inferire le capacità mentali siano gli stessi nelle due aree. E lo scopo del test di Turing (o dei molti simili che costituiscono un pilastro nella psicologia comparativa) è di escogitare una chiara serie di regole per determinare lo status dei soggetti di ogni specie, sul cui possesso di una data capacità siamo incerti. Questo è un gioco e non può esser liberato da tutti i suoi possibili aspetti arbitrari. Inoltre, a meno che non si vada d'accordo con le regole del gioco, non c'è alcun modo di provare il proprio caso per (o contro) una data capacità, con assoluta certezza. Per quel che vedo io, Searle rifiuta semplicemente di fare tali giochi, almeno secondo le regole proposte dall'LA. Egli si assegna il ruolo di un giudice che conosce in anticipo, nella maggior parte dei casi, quale dovrebbe esser la corretta decisione. E, a parer mio, non ci fornisce alcuna regola per i restanti (e più interessanti) casi indecisi; ci dà solo regole di comune buon

senso (le cui trappole e ambiguità sono forse la maggior ragione per organizzare test oggettivi che sono basati su realizzazioni effettive piuttosto che su caratteristiche fisiche come la specie, la razza, il sesso e l'età).

Sarò più dettagliato. Prima di tutto, la discussione sul cervello e su certi processi del cervello è non solo vaga, ma sembra spostare e complicare i problemi che dichiara di risolvere. Nel dire questo non intendo affermare che i dati fisiologici siano irrilevanti; dico soltanto che la loro rilevanza non è chiarita, e il problema di decidere dove il cervello cessa e il non cervello comincia, non è così facile come sembra.

In effetti, dubito che molti neuroanatomisti cercherebbero anche solo di tracciare una linea retta e inalterabile che stabilisca esattamente dove nel regno animale il cervello emerge dal sistema nervoso centrale; e sospetto che qualcuno di loro chiederebbe: perché isolare il cervello e dichiararlo come indispensabile alla mente o alla intenzionalità? Perché non il sistema nervoso centrale o il DNA o (per diventare più restrittivi piuttosto che più liberali) il cervello umano o il cervello caucasiano? Problemi analoghi sorgerebbero nel cercare di specificare per una singola specie come quella umana precisamente quali parti del cervello o quali processi del cervello devono esser presi in considerazione e quando un processo del cervello cessi e quando ne cominci un altro. Del tutto incidentalmente, sarei curioso di sapere se Searle giudicherebbe possibile che un neurofisiologo potesse notare delle affinità tra i processi del cervello di Searle durante il corso del suo ipotetico esperimento e i processi del cervello di un professore cinese. Pure, sono curioso di sapere quale stato mentale assegnerebbe, diciamo, a un verme che striscia sulla terra.

Secondo, mi sembra che, specie nel campo della psicologia, ci siano sempre innumerevoli modi per tosare un gatto, e che questi non siano necessariamente commisurati, specialmente quando si discute di due diverse specie o culture o ere. Così, per esempio, sarei disposto a concedere che per fare calcoli non c'è bisogno della forza intellettuale di Newton o di Leibnitz, che inventarono il calcolo. Ma come proporrebbe Searle di quantificare i relativi "poteri causali" che sono qui implicati, o come altrimenti stabilirebbe la relativa somiglianza degli "effetti"? Il problema è difficile specialmente quando Searle parla di argomenti che presentano un livello zero di comprensione; perché non possediamo scale assolute o razionali in questo settore, ma solo scale relative. In altre parole, possiamo affermare, per definizione, che un dato oggetto può essere preso come criterio di "comprensione zero" e fissare la competenza di altri soggetti confrontandoli rispetto a questa norma; ma altri sono sempre liberi di invocare altre norme. Così, per esempio, Searle usa sé stesso come punto di confronto e asserisce che possiede una comprensione zero del cinese. A meno che l'esecuzione di Searle non sia peggiore di quella di un cane, mi sembra che lo studioso di IA potrebbe ribattere che la comprensione di Searle deve essere più grande di

zero e che il suo ipotetico esperimento non è perciò conclusivo: cioè che un computer, che opera come lui, non si può necessariamente dire che abbia un livello zero di comprensione.

In aggiunta a questi problemi, l'esperimento ipotetico di Searle è basato sull'assunto che IA sarebbe provata "falsa" se si potesse dimostrare che perfino un singolo soggetto in una singola occasione potrebbe superare un test di Turing, malgrado egli possieda quello che si può credere sia una comprensione zero. Questo, a mio parere, è un assunto errato. Nessuno studioso di IA, che io sappia, proclama l'infallibilità. Le sue predizioni sono, nel migliore dei casi, probabilistiche o statistiche e, anche a prescindere da problemi come quello dell'ingannare, ci si devono attendere errori di classificazione sulla base della sola probabilità. Turing, per esempio, predisse che entro l'anno 2000 i computers potranno prendersi gioco di chi li interroga su un test di Turing, e che una macchina potrà esser presa per una persona almeno 30 volte su 100. In breve, sarei d'accordo con Searle se avesse detto che la posizione dell'IA forte non si può provare con assoluta certezza; ma secondo i suoi criteri non si può provare alcuna teoria nella scienza empirica, e perciò respingo la sua tesi che l'Intelligenza Artificiale è falsa.

Forse la questione centrale, sollevata dall'articolo di Searle, è tuttavia: dove si trova l'intelligenza? Searle ci dice che l'intelligenza dei computers sta solo nei nostri occhi. Einstein, tuttavia, era solito dire: "La mia matita è più intelligente di me", e questa massima mi sembra che si avvicini alla verità almeno quanto la posizione di Searle. È vero che senza un cervello che guidi e interpreti il suo output le qualità di una matita o di un computer o degli altri nostri "strumenti del pensiero" non sarebbero molto incisivi. Ma parlando per me, confesso che dovrei assumere la stessa prudente opinione riguardo alle mie qualità. In altre parole, sono del tutto sicuro che non avrei nemmeno potuto "avere" i pensieri espressi nel presente commento senza l'assistenza di vari mezzi per "esternarli e oggettivarli" e renderli accessibili non solo per un ulteriore esame, ma per la loro stessa formulazione. Presumo che ci siano connessioni e corrispondenze causali fra ciò che è ora sulla carta (o è soltanto negli occhi del lettore?) e quello che è successo nel mio cervello o mente: ma è una questione aperta che cosa siano queste connessioni e corrispondenze causali. Inoltre, è solo se si confonde presente e passato, ed eventi interni ed esterni, e li si considera come una "cosa" singola che il pensare o il potere causale dietro il pensiero possono essere attribuiti a un singolo "luogo" o entità. Assicuro che è metaforico se non ridicolo dare alla mia matita credito o biasimo per la qualità dei pensieri di questo commento Ma sarebbe non meno metaforico e ridicolo — almeno nella scienza, a differenza di quanto avviene nella vita quotidiana — dare credito o biasimo al mio cervello, come tale. Qualunque fossero i processi del cervello o i processi mentali implicati nello stendere questo articolo, essi comunque sono stati terminati da lungo tempo.

In un futuro non troppo lontano nemmeno “io” come corpo esisterò più. Questo significa forse che il lettore del futuro non avrà una base valida per stimare se io fui (o, come figura letteraria, sono) più intelligente di una roccia o no? Sono curioso di vedere come Searle risponderebbe a questa domanda. In particolare, vorrei sapere se egli dedurrebbe mai da un manufatto o documento soltanto che il suo autore aveva un cervello o certi processi del cervello. Se è così, come si differenzia questo processo dal fare inferenze sulla mentalità in base solo agli outputs di un soggetto?

Menti decentralizzate

Marvin Minsky
Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass. 02139

Nel suo saggio Searle asserisce: “Lo studio della mente incomincia da un dato: gli esseri umani hanno opinioni, mentre i termostati, i telefoni e le macchine calcolatrici non ne hanno. Se hai una teoria che nega ciò, la teoria è falsa.” No, lo studio della mente non è lo studio delle opinioni: è il tentativo di scoprire concetti — siano essi vecchi o nuovi — che aiutano a spiegare perché certe macchine o animali sanno fare tante cose che altri non sanno fare. Devo contestare che concetti tradizionali, di ogni giorno, come il vedere e il comprendere, non siano abbastanza potenti o robusti per sviluppare o discutere tale punto. Nei secoli passati i biologi discussero sulle macchine e la vita proprio come oggi Searle sui programmi e la mente; si sarebbe potuto udire: “Lo studio della biologia comincia col fatto che gli uomini hanno la vita, mentre le locomotive e i telegrafi no. Se hai una teoria che nega ciò, la teoria è falsa.” Eppure oggi la scienza biologica è basata sul procedere dell’informazione: nessuna nozione di “vivo” appare nella teoria, né noi la rimpiangiamo. Il potere della teoria viene dalla sostituzione della nozione di “autoriproduzione” con nozioni più sofisticate sulla codificazione, traduzione e ricombinazione per spiegare la riproduzione di animali che non si “riproducono” nel comune senso della parola. Similmente, nella scienza della mente, sebbene germi di idee prescientifiche come *credere*, *conoscere* e *significare* siano utili nella vita di ogni giorno, sembrano tecnicamente troppo rozzi per sostenere teorie sofisticate: abbiamo bisogno di sostituire, piuttosto che cercar di sostenere e spiegare tali teorie. Per reali che “Sé” o “comprendere” possano sembrarci oggi, non sono (come il latte o lo zucchero) cose oggettive che le nostre teorie devono accettare e spiegare; sono solo primi passi verso concetti migliori. Non sarebbe appropriato qui render pubbliche le mie idee su come procedere; consideriamo piuttosto l’ipotesi che i nostri successori ripetano per filo e per segno il nostro dibattito

attuale: “L’antico concetto di ’opinione’ si è dimostrato inadeguato fino a quando non è stato sostituito da un *continuum* in cui risultò che le pietre totalizzarono punti zero e i termostati segnarono punti 0.52. Il risultato umano più alto misurato finora è di 67.9. Poiché è teoricamente possibile che qualche cosa sia creduto con l’intensità di 3600 punti, siamo rimasti delusi nel constatare che gli uomini hanno un punteggio così scarso. Né del resto, sono molto esperti se si considera la scala assoluta dell’intendere e avere in mente un progetto specifico. Nondimeno, sono sufficientemente separati dal livello dei termostati” (Olivaw, R.D. (2063), *Robotic reflections*, Phenomenological Science 67.60). Uno scherzo, naturalmente: dubito che esista un concetto monodimensionale come questo che possa dare un contributo determinante. Capire come le parti di un programma possano collegarsi a cose al di fuori di esso — o ad altre parti dentro — è complicato, non solo a causa delle complessità di ipotetiche intenzioni e significati, ma perché diverse parti della mente fanno diverse cose — sia rispetto all’esterno che l’una rispetto all’altra. Ciò solleva un’altra questione: “Nell’impiegare formule come ’A crede che B significhi C’, i nostri precursori filosofi erano inconsciamente intrappolati nell’“errore del singolo agente”, cioè la convinzione che dentro a ciascuna mente ci sia un singolo che crea la convinzione. È strano quanto a lungo quest’idea è sopravvissuta, perfino dopo che Freud pubblicò le sue prime goffe descrizioni delle nostre inconsistenti costituzioni mentali. Certo, quel mito del “Sé” è indispensabile sia per scopi sociali, che per ogni tentativo di un bambino di realizzare modelli semplificati della struttura della sua mente. Invece non è stato di grande utilità nella moderna teoria cognitiva applicata, quale è il nostro lavoro di preservare, riaggiustare e ricombinare quegli aspetti delle parti della mente di un cervello che sembrano avere valore” (Byerly, S. (2080), *New hope for the Dead*, Reader’s Digest, March 13).

Searle parla di lasciare “che l’individuo interni tutti questi elementi del sistema” e poi si lamenta che “non c’è nulla nel sistema che non sia in lui”. Proprio come i nostri predecessori cercarono la “vita” nello studio della biologia, Searle ancora cerca un “quid” nello studio della mente, e ritiene che la IA forte sia impotente a trattare la fenomenologia del comprendere. Poiché questo è un argomento così soggettivo, sento che non è fuori luogo introdurre un po’ di fenomenologia mia personale. Leggendo dell’ipotetico personaggio di Searle che racchiude nella sua mente senza capire l’ipotetico processo di *squiggle squoggle*, che appare comprendere il cinese, ho trovato che la mia esperienza personale aveva qualche qualità della doppia esposizione; “Il testo ha senso per qualche parte della mia mente ma, per altre parti della mia mente, suona molto come se esso stesso fosse scritto in cinese”. Comprendo la sua sintassi, posso analizzare le frasi e posso seguire le deduzioni tecniche. Ma i termini e le ipotesi stesse che parole quali “intendere” e “significare” intendono e significano, mi sfuggono. Sospetto che esse siano come i

“simboli formalmente specificati” di Searle, perché il loro principale significato impegna alcune parti più vecchie della mia mente che non sono in armonia e in contatto con quelle parti più nuove, meglio in grado di trattare tali argomenti (precisamente perché sanno come rafforzare i nuovi concetti dell’Intelligenza Artificiale forte).

Searle considera tali internalizzazioni — quelle non pienamente integrate nell’intera mente — come controesempi o *reductiones ad absurdum* poiché mettono a punto programmi che sono in certo qual modo separati dalle menti.

Io li considero come in grado di illustrare l’usuale condizione della mente normale in cui differenti frammenti di struttura interpretano — e fraintendono — i frammenti di attività che vedono in altri. Non c’è alcun motivo per cui anche i programmi non possano contenere tali contrasti.

In realtà, l’eccessiva semplicità dell’esempio di Searle aumenta la sua forza: il “cinese” dell’uomo non ha contatti con l’altra sua conoscenza, mentre perfino le parti di un computer difficilmente oggi sono unite insieme in modo così semplice.

Nel caso di una mente così separata in due parti, di cui una semplicemente esegue certa gestione spicciola per l’altra, dovrei supporre che ciascuna parte — il computer delle regole cinesi e il suo ospite — avrebbe allora la propria fenomenologia distìnta — forse lungo diverse gradazioni di tempo. Nessuna meraviglia, allora, che l’ospite non possa “capire” il cinese molto scorrevolmente — e qui io sono d’accordo con Searle. Ma (per la lingua, per non dire del camminare o del respirare) certamente le sfumature più essenziali dell’esperienza di intendere e capire emergono non da una nuda serie di dati ma dalle interazioni, dalle consonanze e dai conflitti fra diverse reazioni all’interno di varie immagini parziali di sé.

Che cosa ha a che fare tutto questo con la tesi di Searle? Ebbene, se si considera l’intenzionalità come un attributo esclusivo che ogni macchina ha o non ha, allora l’idea di Searle che l’intenzionalità emerga da qualche principio semantico fisico potrebbe sembrare plausibile. Ma a parer mio, questa idea di intenzionalità come semplice attributo risulta da un’ultrasemplificazione, e una cruda simbolizzazione di attività introspettive complesse e sottili. In breve, quando costruiamo i modelli semplificati delle nostre menti, abbiamo bisogno di termini per rappresentare intere classi di tali consonanze e conflitti — e, penso, questo è il motivo per cui creiamo termini generali come “significare” e “intendere”. È possibile che solo una macchina che presenti “parti” fra loro collegate come una mente umana abbia davvero qualcosa di molto simile a una fenomenologia umana. Eppure anche questo non può avallare alcuna tesi simile a quella di Searle, che cioè il carattere della mente dipende da un processo di informazione più che astratto — come, per esempio, le “proprietà causali particolari” delle sostanze del cervello secondo cui quei processi sono organizzati. E qui io trovo le tesi di Searle più difficili

da seguire. Egli critica il dualismo, eppure si lamenta di antagonisti fittizi che suppongono che la mente sia concreta come lo zucchero. Egli deride l'"operazionalismo residuo" — eppure egli continua a insistere che, in qualche modo, la chimica di un cervello può contribuire alla qualità o al profumo della sua mente senza alcun effetto osservabile sul suo comportamento.

I sostenitori della IA forte non dicono, come suggerisce Searle, che "quello che è specificamente mentale intorno alla mente non ha alcuna intrinseca connessione con le effettive proprietà del cervello". Essi invece avanzano un'ipotesi scientifica molto più discriminante: tali proprietà causali sono importanti per processi simili alla mente e, precisamente, per proprietà che sostengono la computazione. Così, quello che Searle erroneamente vede come una differenza di carattere generale è piuttosto una differenza di dettaglio specifico. La differenza è importante perché quello che potrebbe apparire a Searle come negligente disattenzione ai caratteri vitali è effettivamente una strategia scientifica deliberata e probabilmente benefica. Poiché, come Putnam nota: "La vera domanda è: *che cosa è la nostra forma intellettuale?* e non *che cosa è la materia?*" Si possono spiegare i piccoli effetti nei termini dell'attuale fisica del cervello. Ma quando non si è nemmeno raggiunto il livello di una descrizione *idealizzata* dell'organizzazione funzionale del cervello, parlare dell'importanza delle piccole perturbazioni sembra decisamente prematuro.

Ora, molti sostenitori dell'IA forte insistono nel postulare che l'organizzazione funzionale è l'unica dipendenza di questo genere, ed è questa ben marcata tesi che porta direttamente alla conclusione che Searle sembra tanto avversare: che le macchine non organiche possano avere gli stessi generi di esperienza che gli esseri umani hanno. Questo mi sembra giusto. Non posso però capire perché Searle sia così contrario all'idea che un mucchio di roba vecchia possa avere sentimenti come i nostri. Egli non presenta alcuna prova contro ciò, semplicemente cerca di descrivere come è assurdo immaginare macchine con menti come le nostre — con le intenzioni cioè e tutto il resto — fatte di pietra e carta invece che di elettroni e automi. Ma sono stupito di come Searle, sbarazzatosi del tanto disprezzato dualismo e operazionalismo, propone di distinguere le autentiche intenzioni dei composti di carbonio dalle loro imitazioni identiche nel comportamento, ma mentalmente contraffatte.

Mi sembra di aver trattato quindi delle tesi sul cinese e di quelle sulla sostanzialità. Eppure resta l'impressione che ci sia qualcosa di profondamente errato in tutte queste discussioni intorno alle nostre menti: di esse, nulla mai sembra essere definitivo. Dalle menti migliori, da tutti i punti di vista, emergono pensieri e metodi di bassa qualità e scarso valore.

Certo questo sorge da un fardello di idee tradizionali inadeguate nei

confronti di questa impresa tremendamente difficile. Perfino la nostra logica può essere sospetta. Così potrei perfino convenire con Searle che le moderne idee computazionali sono di poca importanza a questo punto, se, come lui, potessi giudicare quelle idee in base alla loro coerenza e consistenza rispetto a precedenti costruzioni filosofiche. In ogni caso, una volta che si sospetta che ci siano altre cattive mele nel serbatoio logistico, la rigorosa consistenza diventa una regola troppo fragile — e noi dobbiamo umilmente volgerci a quell'evidenza che possiamo trovare. Così, poiché questo è ancora un periodo formativo per le nostre idee sulla mente, suggerisco che dobbiamo rimanere particolarmente sensibili al potere empirico che ogni nuova idea può darci nell'approfondire ulteriormente l'argomento.

E, come neanche Searle riesce a negare, il computazionalismo è la principale fonte delle nuove macchine e dei programmi che hanno prodotto per noi le prime imitazioni, per quanto limitate e povere, di un'attività simile alla mente.

La fonte primaria dell'intenzionalità

Thomas Natsoulas
Dipartimento di Psicologia, Univ. di California, Davis, Calif. 95616

Ho condiviso la convinzione di Searle: il livello di descrizione che i programmi dei computer esemplificano non è adeguato alla spiegazione della mente. Le mie critiche sono apparse nelle discussioni di teorie percettive che fanno ben scarso, se non nullo, riferimento ai programmi di computer di per sé (Natsoulas 1974; 1977; 1978a; 1978b; 1980). Come Searle opta per la base materiale della mente ("Gli effettivi fenomeni mentali umani dipendono dalle effettive proprietà fisico-chimiche degli attuali cervelli umani"), io ho affermato che la particolare natura concreta delle consapevolezze percettive, come capitano in un certo sistema percettivo, è essenziale per i riferimenti che esse fanno agli oggetti, eventi, o situazioni nell'ambiente di stimolo.

In opposizione a Gibson (per esempio, 1966; 1967; 1972) la cui teoria percettiva è equivalente alle ipotesi concernenti il raccogliere da parte dei sistemi percettivi entità astratte chiamate "invarianti informazionali" dal flusso degli stimoli, io ho affermato:

> I sistemi percettivi funzionano con le loro proprie modalità per attribuire le proprietà scoperte, che sono specificate informazionalmente, all'ambiente fisico reale intorno a noi. Gli invarianti informazionali a cui un sistema percettivo reagisce sono definiti astrattamente (da Gibson) in modo tale che il processo stesso di risonanza può esemplificarli. Ma il processo di risonanza *non* è astratto in sé. E la caratterizzazione di esso al livello di invarianti informazionali non basta alla teoria della percezione. È fondamentale per la teoria della percezione che gli invarianti informazionali si riferiscano a *modalità concrete* che sono caratteristiche dell'organismo (Natsoulas 1978b, p. 281).

L’ultima affermazione è fondamentale per la teoria percettiva, ho sostenuto, se tale teoria deve spiegare l’intenzionalità della consapevolezza percettiva (cfr. anche Ullman, *Against Direct Perception*, BBS 3(3) 1980).

E proprio come Searle sommariamente rifiuta il tentativo di eliminare l’intenzionalità, dicendo che non dà alcun vantaggio “fingere un’anestesia”, io ho sostenuto contro il tentativo di Dennett (1969; 1971; 1972) di trattare l’intenzionalità come una semplice sovrapposizione euristica sulla teoria estensionale del sistema nervoso e dei movimenti corporei, la tesi seguente:

> Nel sapere che siamo soggetti conoscenti, c’è una cosa che sappiamo: che siamo consapevoli degli oggetti in un modo diverso dal modo “spento” in cui qualche volta pur pensiamo a essi. La presenza per esperienza degli oggetti rende difficile, se non impossibile, sostenere che le percezioni interessano solo l’acquisizione dell’informazione… È questo genere di presenza che rende la specificità percettiva qualcosa di più che una “interpretazione” o una “sovrapposizione euristica” da cui ci si può astenere una volta che una giustificazione abbastanza completa sia disponibile. L’“essere qualitativamente presente” degli oggetti e delle scene… è almeno così dubbio quanto la nostra stessa esistenza (Natsoulas 1977, pp. 94-95; cfr. Searle, 1979b, p. 261).

Comunque, non so in che cosa consiste la specificità percettiva. Dò dei suggerimenti e credo, con Sperry (1969; 1970; 1976), che una descrizione oggettiva di un’esperienza soggettiva sia possibile in termini di funzioni mentali. Una tale descrizione dovrebbe includere quel carattere o quei caratteri di consapevolezza percettiva che la fanno essere (o la fanno apparire, negli stati di allucinazione) consapevolezza percettiva di un oggetto, un evento, o una situazione nell’ambiente fisico o nel corpo del percettore al di fuori del sistema nervoso. Se la descrizione non includesse questo carattere sarebbe incompleta, a parer mio, e necessiterebbe di un ulteriore sviluppo.

In un altro articolo recente sull’intenzionalità, Searle (1979a), a proposito dell’inevitabilità del “cerchio intenzionale”, sostiene che ogni spiegazione dell’intenzionalità che possiamo raggiungere presuppone una comprensione dell’intenzionalità: “Non c’è analisi della intenzionalità in condizioni logicamente necessarie e sufficienti della forma ’X è uno stato intenzionale S se e solo se p, q, e r’, dove ’p, q, e r’ non fanno uso di nozioni di intenzionalità” (p. 195). Questo per dire che l’intenzionalità degli stati mentali non è riducibile; ma non penso che questo implichi di per sé la possibilità che l’intenzionalità possa essere una proprietà di certi processi del cervello. Searle potrebbe ancora accettare l’opinione che è piuttosto una delle loro proprietà “base” ancora sconosciute.

Ma l’accurata caratterizzazione di Searle della relazione tra cervello e processi mentali come causale, dove i processi mentali consistentemente sono detti prodotti da processi del cervello, dà un’impressione diversa. Naturalmente i processi del cervello producono altri processi del cervello, ma se egli avesse inteso includere i processi mentali fra i secondi, avrebbe egli

scritto delle sole proprietà causali del cervello in una discussione sulla base materiale dell'intenzionalità? Si è tentati di credere che Searle parlerebbe in favore di qualche forma di interazionismo riguardo alla relazione mente-cervello. Penso che la sua analogia dei processi mentali con i prodotti dei processi biologici, come zucchero e latte, fosse tesa a illuminare la base causale dei processi mentali, e non la loro natura. La sua affermazione che l'intenzionalità è "un fenomeno biologico" è preceduta da "qualunque altra cosa l'intenzionalità possa essere" e seguita da una ripetizione della sua tesi precedente riguardante la base materiale della mente (i processi mentali come prodotti dai processi del cervello). E sono ben sicuro che Searle non considererebbe i processi mentali uguali a un altro degli effetti del cervello e, precisamente, ai comportamenti (cfr. Searle 1979b).

Sebbene possa essere allettante costruire la posizione di Searle come una forma di dualismo, ci resta l'alternativa più probabile che egli ha semplicemente deciso di non prendere, in questi scritti recenti, una posizione sulla questione ontologica. Egli ha scelto di trattare solo con quegli elementi che gli sembrano già chiari per quanto riguarda il problema dell'intenzionalità. Comunque questa enfasi sulla base materiale dell'intenzionalità sembrerebbe essere un'indicazione che egli avanza passo passo verso una posizione sicura sulla questione ontologica e che questa posizione ha importanza ai fini della comprensione dell'intenzionalità. Insisto su quest'ultimo aspetto per ciò che Searle ha scritto sulla "forma di realizzazione" degli stati mentali ancora in un altro articolo sull'intenzionalità. In questo articolo (Searle 1979c) egli dichiara che "i problemi ontologici tradizionali sugli stati mentali sono per la maggior parte semplicemente irrilevanti rispetto ai caratteri intenzionali" (p. 81). Non importa come uno stato mentale è realizzato. Conoscere che cosa è uno stato intenzionale richiede solo che conosciamo il suo "contenuto rappresentativo" e il suo "modo psicologico".

Ma questo non ci direbbe realmente che cosa è lo stato, ma solo quale è, o di che genere è. Per esempio, può darsi che io sappia che lo stato mentale che mi è appena capitato di avere fosse un pensiero di passaggio: dovuto al fatto che in questo momento a Londra piove. Quello che mi è venuto in mente è un genere di pensiero del tipo "piove-a-Londra-giusto-ora". Sapere che capita questo pensiero e molti altri, pure il conoscere i loro contenuti rappresentativi e modi psicologici, non sarebbe conoscere che cosa è il pensiero, e che cosa è quell'accadimento mentale che è il pensiero di passaggio.

Di più, perché uno stato mentale o accadimento abbia i suoi caratteri intenzionali, deve avere una forma della realizzazione che esso dà a essi. Searle ha affermato: "Non importa come uno stato intenzionale si realizzi, finché la realizzazione è una realizzazione della sua intenzionalità" (1979c. p. 81). La seconda parte di questa frase è un'ammissione che effettivamente

importa *come* si realizza. La spiegazione dell'intenzionalità rimane incompleta in assenza di una giustificazione della sua origine e della sua relazione con quell'origine. Questo punto diventa indicativo quando è considerato nel contesto di un'altra discussione contenuta nello stesso articolo.

In una parte dell'articolo Searle ha prestato un po' di attenzione a quello che egli ha chiamato "la forma primaria dell'intenzionalità", e precisamente la percezione (cfr. Searle 1979b, pp. 260-61). L'esperienza visiva che si ha di una tavola è una "presentazione" della tavola, come opposta a una rappresentazione di essa. Ancora, tali presentazioni sono intrinsecamente intenzionali, perché ogni volta che si presentano, la persona con quel mezzo percepisce o ha allucinazioni su una tavola. L'esperienza visiva è di una tavola, anche se non è una rappresentazione di una tavola, perché è soddisfatta dalla presenza fisica di una tavola, là dove la tavola sembra visivamente essere collocata.

Concludendo questa discussione, Searle aggiunge:

> Dire che ho un'esperienza visiva ogni volta che percepisco la tavola visivamente, è fare un'affermazione ontologica, un'affermazione sulla forma di realizzazione di uno stato intenzionale, mentre dire che tutte le mie opinioni hanno un contenuto rappresentazionale, non è fare un'affermazione ontologica (1979c, p. 91).

Dal momento che Searle direbbe che lui e la gente che legge questo articolo, e gli animali e così via, hanno esperienze visive, la domanda a cui è necessario che egli risponda, da quel teorico dell'intenzionalità che è, risulta la seguente: qual è l'affermazione ontologica che egli fa nel fare così? O qual è la "forma di realizzazione" delle nostre esperienze che Searle afferma quando le attribuisce a noi?

La stanza degli scacchi: ulteriore smitizzazione di IA forte

Roland Puccetti
Dipartimento di Filosofia, Dalhousie University, Halif., Nova Scotia, Canada

Sul terreno che ha circoscritto, che è considerevole, Searle mi sembra completamente vittorioso. Quello che farò qui è di alzare le vedute del suo argomento e di provarle su un obiettivo ancora più largo e molto invogliante.

Supponiamo di avere un umano intelligente di una cultura che non conosce gli scacchi, per esempio del Ciad nell'Africa Centrale, e di introdurlo in una stanza degli scacchi. Qui egli si trova di fronte una console di calcolatore su cui sono esposti i numeri 1-8, le lettere R, N, B, K, Q e P più le parole WHITE e BLACK. Gli viene detto che WHITE muove per primo e poi BLACK, alternandosi finché le luci della console si spengono. C'è

naturalmente nella macchina una rappresentazione binaria della scacchiera che impedisce le mosse illegali, ma egli non ha bisogno di sapere nulla di questo. Gli viene insegnato a identificarsi con WHITE, a muovere per primo e che la combinazione lettera-numero P-K4 è un buon inizio, così egli schiaccia P-K4 e aspetta.

Sulla console appare BLACK, seguito da tre combinazioni lettere-numero alternative P-K4, P-QB4, P-K3. Se questo fosse un programma *depth/first*, ciascuna di queste tre repliche sarebbe cercata due mosse più avanti e verrebbe fornita di una valutazione statica. Così contro P-K4 di BLACK, WHITE potrebbe provare o N-KB3 o B-B4. Se la scelta cade su N-KB3, la replica di BLACK, N-QB3 o P-Q3, conducono entrambe a una valutazione di +1 per WHITE; mentre se essa cade su B-B4, le repliche di BLACK, B-B4 o N-KN3, producono una valutazione rispettivamente di +0 e +3. Poiché il nostro ciadiano è stato istruito a respingere ogni combinazione lettera-numero che dà una valutazione inferiore a +1, egli non sceglierà B-B4, ma sceglierà N-KB3 a meno che non compaia una valutazione più alta. E infatti fa così. La risposta di BLACK P-QB4 permette N-KB3, e contro questa mossa le migliori contromosse di BLACK, P-Q3, P-K3 e N-QB3 producono valutazioni di +7, +4 e +8. Dall'altra parte, se questo fosse un programma *breadth/first* (in cui tutti i nodi a uno stesso livello sono esaminati prima dei nodi di livello più profondo) le continuazioni di WHITE procederebbero più staticamente: ma di nuovo questo non sarebbe un problema per il ciadiano nella stanza degli scacchi, che, instanziando un qualunque tipo di programma, non ha la benché minima nozione di quello che sta facendo.

Dobbiamo essere perfettamente chiari su quello che questo implica: ambedue i programmi descritti qui giocano a scacchi (Frey 1977) e il secondo con sorprendente successo in una recente competizione, quando è stato eseguito su un calcolatore più potente, un Cyber 170 della Control Data di dimensioni più ampie (Frey 1977, Appendice). Ciononostante non c'è la minima ragione per credere che uno dei due programmi capisca il gioco degli scacchi. Ciascuno esegue "operazioni computazionali su elementi specificati in modo puramente formale" ma così farebbe il ciadiano che non capisce nella nostra stanza degli scacchi, anche se, naturalmente, in modo molto più lento (potremmo usarlo solo per partite per lettera, per questa ragione). Queste operazioni non possono, da sole, costituire una comprensione del gioco, non importa quanto intelligentemente sia giocato.

È sorprendente che questo non sia stato notato prima. Per esempio, gli autori del programma che ha avuto più successo fino a questo momento (Slate & Atkin 1977) scrivono che la funzione valutativa del CHESS 4.5 comprende che è sbagliato fare una mossa che lascia uno dei pezzi sotto attacco e indifeso, che è bene fare una mossa che mette sotto forchetta due pezzi nemici, è bene fare una mossa che previene una manovra a forchetta

dell’avversario (p. 114).

Eppure in una situazione in cui lo stesso programma gioca con WHITE, e può muovere solo il re in KB5, e BLACK ha il re in KR6, e il solo pedone di BLACK avanza da KR4 a una possibile nuova regina, la valutazione iniziale delle sei mosse legali di WHITE è quella che segue:

Mossa	**Punteggio PREL**
K-K4	-116
K-B4	-114
K-N4	-106
K-K5	-121
K-K6	-129
K-B6	-127

In altre parole, con una ricerca a un passo il programma dà una smilza preferenza alla mossa di WHITE in K-N4 perché uno dei suoi valutatori incoraggia il re a essere vicino al pedone nemico ancora vivo, ed N4 è la posizione più vicina che può prendere legalmente il re di WHITE. Questo punteggio preliminare non differisce molto da quello delle altre mosse, poiché, come ammettono gli autori, “la funzione di valutazione non comprende che il pedone sarà catturato due mosse più tardi” (p. 111). È solo dopo una iterazione a due passi e poi a tre passi di K-N4 che il programma trova che tutte le possibili repliche sono state prese in considerazione. Gli autori, candidamente, concludono:

> L’intera ricerca a 3 passi qui fu conclusa in circa 100 millisecondi. In un torneo la ricerca sarebbe andata avanti per circa 12 passi per avere lo stesso risultato, poiché il programma non riesce a vedere che, poiché WHITE non può forzare una posizione in cui tutto il materiale è andato, la partita è necessariamente patta (p. 113).

Ma allora se CHESS 4.5 non capisce nemmeno questo degli scacchi, perché dire che esso “comprende” la manovra a forchetta, e così via? Tutto quello che ciò può significare è che il programma ha dentro di sé dei valutatori predefiniti che lo scoraggiano dall’andare dentro le forchette dell’avversario, e lo incoraggiano a cercare modi per mettere sotto forchetta l’avversario. Questo non è comprendere, poiché, come abbiamo visto, il nostro ciadiano nella stanza degli scacchi potrebbe faticosamente raggiungere lo stesso risultato alla console, nella beata ignoranza della scacchiera, delle posizioni degli scacchi, o addirittura di come il gioco viene giocato. Naturalmente in questo modo viene simulato un gioco di scacchi intelligente, ma non per

questo viene duplicata la comprensione degli scacchi.

Fino a circa la metà di questo secolo, le macchine che giocavano a scacchi erano automi con giocatori umani intelligentemente nascosti dentro di essi (Carroll 1975). Ora abbiamo automi molto più complessi, e mentre i programmi che eseguono sono dentro di loro, non hanno l'intenzionalità verso le mosse degli scacchi che fanno, che i minuscoli umani avevano nelle burle di ieri. Semplicemente non sanno quello che fanno.

Searle senza necessità guasta il suo argomento verso la fine del suo articolo offrendo l'osservazione, forse con lo scopo di disarmare i più strenui difensori della IA forte, che noi umani siamo "macchine pensanti". Ma sicuramente, se lui era nel giusto nell'invocare significati letterali contro i proclami che, per esempio, i termostati hanno convinzioni, è in errore nel dire che gli uomini sono macchine.

Letteralmente, non c'erano macchine su questo pianeta 10.000 anni fa, mentre la specie dell'*homo sapiens* è arrivata qui da almeno 100.000 anni, e perciò non può essere che gli uomini siano macchine.

Il "potere causale" delle macchine

Zenon W. Pylyshyn
Centro per lo studio avanzato nelle Scienze del Comportamento, Stanford, Calif. 94305
Dipartimento di Psicologia, Università di Western Ontario, London, Canada N6A, 5C2

A che genere di materiale può alludere? Searle vorrebbe che noi credessimo che i computer, in qualità di manipolatori di simboli formali, necessariamente mancano della qualità dell'intenzionalità, o della capacità di capire e di riferire, perché hanno "poteri causali" diversi dai nostri. Sebbene non sia esplicitamente dichiarato che cosa significa avere poteri causali diversi (a parte il non essere capace d'intenzionalità), appare almeno che i sistemi che sono funzionalmente identici, non hanno bisogno di avere gli stessi "poteri causali". Così la relazione di equivalenza riguardo ai poteri causali è un raffinamento della relazione di equivalenza riguardo alla funzione. Quello che Searle vuole affermare è che solo i sistemi che sono equivalenti agli umani in questo senso più forte possono avere intenzionalità. La sua tesi è basata sulla supposizione che l'intenzionalità è strettamente legata a proprietà specifiche e che è letteralmente causata da esse. Da questo punto di vista sarebbe estremamente improbabile che ogni sistema non fatto di protoplasma — o qualcosa di essenzialmente identico al protoplasma — possa avere intenzionalità. Così, se sempre più cellule nel vostro cervello dovessero essere sostituite da circuiti integrati, programmati in modo tale da mantenere la

funzione input-output di ciascuna unità identica a quella dell'unità che viene sostituita, in tutta verosimiglianza parlereste esattamente come fate ora, eccetto che vi asterreste dal volere attribuire un significato a tutto ciò. Quello che noi osservatori esterni potremmo ritenere siano parole, diventerebbe per voi soltanto certi rumori che i circuiti vi hanno fatto fare.

Searle presenta una varietà di metafore seducenti e fa appello all'intuizione in appoggio a questa opinione piuttosto stupefacente. Per esempio, egli chiede: perché dovremmo trovare così sorprendente l'opinione che l'intenzionalità è legata a proprietà dettagliate della composizione materiale del sistema, quando accettiamo tranquillamente l'affermazione parallela nel caso della secrezione lattea? Certamente è ovvio che solo un sistema con certi poteri causali può produrre latte; ma perché allora non dovrebbe essere vero lo stesso per l'abilità di riferirsi? Perché questo esempio è così importante per Searle? Il prodotto della lattazione è una sostanza, il latte, le cui proprietà che essenzialmente lo definiscono sono, naturalmente, fisiche e chimiche (sebbene nulla impedisca la produzione di latte sintetico usando un processo che è materialmente molto diverso dalla lattazione del mammifero). Searle ci sta allora proponendo che l'intenzionalità è una sostanza secreta dal cervello, e che un possibile esame per l'intenzionalità potrebbe interessare, diciamo, il tessuto del cervello che ha realizzato alcuni degli episodi mentali?

Allo stesso modo Searle sostiene che il fatto d'avere un programma non può essere una condizione sufficiente per l'intenzionalità poiché si può mandare a effetto quel programma su una macchina di Turing consistente in un "rotolo di carta igienica e un mucchio di piccole pietre". Una tale macchina non avrebbe intenzionalità perché tali oggetti "sono il genere sbagliato di materiale" per avere intenzionalità. Ma qual è il giusto genere di materiale? Sono insiemi di cellule, neuroni individuali, protoplasmi, molecole di proteine, atomi di carbonio e idrogeno, particelle elementari? Lasciamo che Searle nomini il livello; esso può esser simulato perfettamente usando "il genere sbagliato di materiale".

Chiaramente non è il materiale che ha l'intenzionalità. Le tue cellule cerebrali non trasportano più di quanto non facciano le condutture dell'acqua, i pezzi di carta, le operazioni di computer, o l'homunculus nell'esempio della stanza cinese. Searle non presenta alcuna prova per l'affermazione che quello che crea la differenza tra l'esser capace di riferire e il non esser capace — o mostrare qualsiasi altra capacità — è una proprietà del sistema più fine di quelle che possono essere catturate da una descrizione funzionale. Inoltre è ovvio dalla tesi di Searle che la natura del materiale non può essere ciò che è rilevante, dal momento che l'inglese madrelingua che ha memorizzato le regole formali si suppone sia un esempio di un sistema fatto del materiale giusto, eppure esso manca ancora della relativa intenzionalità. Pur avendo detto tutto questo, si potrebbe ancora voler sostenere che in certi casi — forse

nel caso dell'esempio di Searle — potrebbe essere appropriato dire che nulla si riferisce a nulla, o che i simboli non vengono usati in un modo che si riferisca a qualcosa. Ma se volessimo negare che questi simboli si riferiscono a qualcosa, sarebbe appropriato chiedere che cosa ci autorizza a dire che un simbolo si riferisce a qualcosa.

Ci sono almeno tre approcci per rispondere alla domanda: il punto di vista di Searle che lo attribuisce alla natura dell'incorporamento del simbolo (della sostanza stessa del cervello); la tesi funzionalista tradizionale che indica il ruolo funzionale che il simbolo gioca nel comportamento complessivo del sistema; e la tesi associata a filosofi come Kripke e Putnam che indica la natura della connessione causale che il simbolo ha con certi eventi passati. Gli ultimi due sono compatibili nella misura in cui specificare il ruolo funzionale di un simbolo nel comportamento di un sistema non preclude la specificazione delle sue interazioni causali con un ambiente. Va notato che Searle non considera nemmeno la possibilità che un modello computazionale puramente formale possa costituire una parte essenziale di una teoria adeguata, mentre gli altri due approcci contenevano anche una considerazione dei traduttori del sistema, e una considerazione di come i simboli vengano ad acquistare il ruolo che hanno nel funzionamento del sistema.

Funzionalismo e referenza

La tesi funzionalista è generalmente quella dominante sia in IA che nella psicologia del trattamento dell'informazione. Nel passato, il mentalismo spesso sosteneva che la referenza era stabilita da relazioni di similarità; un'immagine poteva essere riferita a un cavallo solo se assomigliava sufficientemente a un cavallo. Il comportamentismo mediazionale considerò la referenza come una semplice traccia della percezione: un evento mentale è riferito a un certo oggetto se condivide alcune proprietà di quell'oggetto quando è percepito. Ma la psicologia dell'analisi dell'informazione ha optato per un livello di descrizione che tratta degli aspetti informazionali o codificati degli effetti dell'ambiente sull'organismo. In base a questa tesi è stato sostenuto che quello che un simbolo rappresenta può essere visto esaminando come il simbolo entra in relazione con altri simboli e con trasmettitori. È questa posizione che Searle usa del tutto specificamente. La mia opinione è che, sebbene Searle abbia ragione nell'indicare che alcune versioni della risposta funzionalista sono in un certo senso incomplete, egli è fuori strada sia nella diagnosi di dove sta il problema, che nella prognosi di come la posizione cognitivista debba stabilire un'ipotesi impoverita del funzionamento mentale (cioè l'ipotesi debole di IA).

Il senso in cui una risposta funzionalista potrebbe essere incompleta è che ha mancato di fare il passo successivo nello specificare che cosa nel sistema

garantisce l'attribuzione agli stati funzionali (o alle espressioni simboliche che esprimono quello stato) di un particolare contenuto semantico piuttosto che un altro contenuto logicamente possibile. Una teoria cognitiva dichiara che il sistema si comporta in un certo modo perché certe espressioni rappresentano certe cose (cioè, hanno una certa interpretazione semantica). In più, è essenziale che facciano così: altrimenti non saremmo in grado di assumere certe classi di comportamenti regolari in una singola generalizzazione del genere "il sistema fa X perché lo stato S rappresenta ciò" — per esempio, la persona corse fuori dall'edificio perché credeva che fosse in fiamme (per una discussione di questa tesi, cfr. Pylyshyn 1980b). Ma la particolare interpretazione appare estranea alla teoria nella misura in cui il sistema si comporterebbe esattamente nello stesso modo senza l'interpretazione. Così Searle conclude che siamo solo noi, i teorici, che prendiamo un'espressione per rappresentare, diciamo, l'edificio in fiamme. Non si considera che il sistema rappresenti alcunché, perché, letteralmente, esso non sa a che cosa si riferisce l'espressione: solo noi teorici lo sappiamo. Stando così le cose, non si può dire che il sistema si comporti in un certo modo in grazia di quello che rappresenta. Questo è in contrasto col modo in cui il nostro comportamento è determinato: noi effettivamente ci comportiamo in un certo modo a causa di ciò su cui i nostri pensieri vertono. E ciò, secondo Searle, aggiunge per IA debole la giustificazione funzionalista in cui analogie formali stanno al posto di contenuti mentali ma esse, in sé, né hanno né spiegano tali contenuti mentali.

Gli ultimi pochi passi sono, comunque, senza una prosecuzione. Il fatto che eravamo noi, i teorici, a fornire l'interpretazione delle espressioni non significa in sé e per sé che una tale interpretazione è semplicemente una questione di convenienza, o che c'è un senso in cui l'interpretazione è nostra piuttosto che del sistema. Naturalmente è possibile che l'interpretazione sia solo nella mente del teorico e che il sistema si comporti nel modo che fa per ragioni interamente diverse. Ma anche se ciò fosse, non ne seguirebbe semplicemente il fatto che il teorico di IA è quello che ne raggiunge l'interpretazione. Molto dipende dalle ragioni per cui si raggiunge quella interpretazione.

In ogni caso, la questione se l'interpretazione semantica risieda nella testa del programmatore o nella macchina è la domanda sbagliata da fare. Una domanda più pertinente sarebbe: che cosa fissa l'interpretazione semantica degli stati funzionali, o che libertà ha il teorico nell'assegnare un'interpretazione semantica agli stati del sistema? Quando un computer è considerato come un congegno autocontenuto per trattare simboli formali, abbiamo gran libertà nell'assegnare interpretazioni semantiche agli stati. Infatti, abitualmente cambiamo la nostra interpretazione degli stati funzionali del computer, talvolta considerandoli come numeri, talvolta come caratteri

d’alfabeto, talvolta come parole o descrizioni di una scena, e così via. Anche dove è difficile pensare a un’interpretazione coerente che sia diversa da quella che il programmatore aveva in mente, tali alternative sono sempre possibili in linea di massima. Comunque, se attrezziamo la macchina di trasduttori e le permettiamo di interagire liberamente in contesti sia naturali che linguistici, e se la dotiamo del potere di produrre conclusioni, sintatticamente specificato, è ben ovvia la libertà che il teorico (che sa come i trasduttori operano e perciò sa a che cosa rispondono) ancora avrebbe nell’assegnare un’interpretazione coerente agli stati funzionali in modo tale da catturare regolarità psicologicamente rilevanti nel comportamento.

Il ruolo delle intuizioni

Supponiamo che le connessioni tra il sistema e il mondo menzionate sopra (e possibilmente altre considerazioni che nessuno ha ancora esaminato) limitassero unicamente le possibili interpretazioni che potrebbero essere poste negli stati rappresentazionali. Risolverebbe ciò il problema di giustificare l’attribuzione di particolari contenuti semantici a questi stati? Qui io sospetto di incontrare differenze di opinione che possono essere irrisolvibili, semplicemente perché sono fondate su intuizioni diverse. Per esempio, sorge immediatamente la questione se possediamo un’interpretazione privilegiata dei nostri pensieri che deve avere la precedenza su tali analisi funzionali. E se è così, c’è l’ulteriore domanda se la consapevolezza è ciò che permette l’accesso privilegiato: e da qui la domanda su cosa si debba fare riguardo all’evidente necessità di considerare come un fatto reale i processi mentali inconsci. Per quello che posso vedere, la sola cosa che rafforza tale particolare opinione è l’intuizione che, qualunque cosa possa essere vera riguardo le altre creature, io almeno so a che cosa si riferiscono i miei pensieri perché ho un accesso esperienziale diretto ai referenti dei miei pensieri. Anche se noi avessimo forti intuizioni su questi casi, c’è una buona ragione per credere che tali intuizioni dovrebbero essere considerate non più che fonti secondarie di vincoli, la cui validità dovrebbe essere giudicata in base al modo in cui operano i sistemi teorici basati su di esse. Non possiamo considerare come sacre le intuizioni (per esempio, se un’altra creatura ha l’intenzionalità) specialmente quando tali intuizioni si fondano (come quelle di Searle, per sua propria ammissione) sulla conoscenza di che cosa è fatta la creatura o la macchina. Searle è pronto ad ammettere che altre creature possono avere l’intenzionalità se “noi possiamo vedere che esse sono fatte di materia simile a noi stessi”. Chiaramente le intuizioni basate su tale sciovinismo antropocentrico non possono costituire la fondazione di una scienza della cognizione (cfr. *Cognition and Consciousness in Nonhuman Species*, BBS 1(4) 1978).

Un problema essenziale nella scienza — specialmente in una scienza in sviluppo come la psicologia cognitiva — è quello di decidere quali generi di fenomeni "vanno insieme", nel senso che essi ammetteranno un insieme uniforme di principi esplicativi. Le teorie del trattamento dell'informazione hanno ottenuto successo nel giustificare aspetti della soluzione di problemi, del trattamento del linguaggio, della percezione e così via, mascherando deliberatamente la distinzione conscio-inconscio, raggruppando in categorie comuni una larga classe di processi, governati da regole, necessari a giustificare il fatto che si funziona, indipendentemente dal fatto che la gente sia o non sia al corrente dei medesimi. Queste teorie hanno messo da parte questioni quali: che cosa costituisce un'esperienza consapevole o una "pura e semplice sensazione" che tratta solo con alcuni dei suoi correlati funzionali attendibili (come il credere che uno sia in pena, opposto all'esperienza della pena stessa); si è così in larga misura deliberatamente evitata la domanda su cosa dà ai simboli la loro semantica. Poiché l'IA ha scelto di evidenziare i fenomeni in questo modo, persone come Searle sono portate a concludere che quello che si sta facendo è IA debole, o il modellare la struttura funzionale astratta del cervello senza riguardo per quello che i suoi stati rappresentano. Pure non c'è ragione di pensare che questo programma non porta all'IA forte come soluzione. Non c'è alcuna ragione di dubitare che, per esempio, quando e se un robot è costruito, l'attribuzione dell'intenzionalità alle macchine programmate sarà garantita come la sua attribuzione alle persone, e per ragioni che non hanno assolutamente nulla a che fare con la questione della consapevolezza.

Quello che si trascura frequentemente nelle discussioni sull'intenzionalità è che non possiamo stabilire con qualche grado di precisione che cosa è che ci autorizza a dichiarare che le persone sappiano riferirsi (sebbene ci siano una o due idee generali, come quelle discusse sopra), e perciò gli argomenti contro l'intenzionalità dei computers tipicamente si riducono ad "argomenti derivati dall'ignoranza". Se sapessimo che cosa è che garantisce il nostro dire che le persone si riferiscono, potremmo essere anche in grado di dichiarare che l'attribuzione del contenuto semantico alle espressioni computazionali formali — sebbene sia compiuta in pratica da una "inferenza rispetto alla migliore spiegazione" — è alla fine garantita esattamente allo stesso modo. L'umiltà, se non altro, dovrebbe spingerci ad ammettere che non sappiamo come dovremo descrivere le capacità dei futuri robot e delle altre macchine computazionali, anche quando sappiamo come operano i loro circuiti elettronici.

La replica comportamentista

Howard Rachlin

Dipartimento di Psicologia, Univ. Statale di New York, Stony Brook, NY 11794

È facile trovarsi d'accordo con l'argomento negativo che Searle usa sulla mente e l'IA nel suo stimolante articolo. Quello che è difficile accettare è la concezione della mente che ha Searle.

Il suo argomento negativo è che la mente non può mai essere un programma di computer. Naturalmente è quello che i comportamentisti hanno detto sempre (nonostante il "comportamentismo residuo" nelle menti dei ricercatori di IA). Il suo argomento positivo è che la mente è (lo stesso che) il cervello. Ma questo è chiaramente falso quanto la posizione dell'IA forte che egli critica.

Forse il punto di vista comportamentista può essere meglio capito con due esempi, uno considerato da Searle e uno (per quanto discretamente ovvio) non considerato. L'esempio della combinazione del robot è essenzialmente di tipo comportamentista. Un robot si comporta esattamente come un uomo. Pensa anche? Searle dice: "Se il robot appare e si comporta come noi, allora dovremmo supporre, fino a prova contraria, che *deve avere stati mentali come i nostri*" (corsivo mio). Naturalmente lo supporremmo. Ed è chiaro quello che a questo robot sarebbe richiesto di fare. Potrebbe rispondere a domande su una storia che sente, ma dovrebbe anche ridere e piangere al momento giusto: dovrebbe essere in grado di dire quando la storia è finita. Se è una storia morale, il robot potrebbe cambiare il suo successivo comportamento in situazioni simili a quelle che la storia descrive. Il robot potrebbe fare domande sulla storia stessa, e le risposte che riceve potrebbero cambiare il suo comportamento più tardi. La lista dei comportamenti tipicamente umani in "risposta" alle storie è infinita. Con un numero finito di esperimenti non possiamo comunque essere totalmente certi che il robot abbia capito la storia. Ma la prova contraria può venire in un solo modo, dal comportamento successivo del robot. Cioè, può essere che il robot provi che non ha capito una storia raccontata a lui in un tempo X, perché fa o perché dice qualcosa di diverso rispetto a quanto farebbe un essere normale che udisse una storia simile, in simili condizioni. Se supera tutte le nostre prove comportamentali, diremmo che, restando possibile una futura smentita basata sul suo comportamento, il robot ha capito la storia. E noi diremmo questo anche se dovessimo aprire il robot e trovare un uomo che traduce il cinese, un computer, un cane, una scimmia, o un pezzo di formaggio senza sapore. Il test appropriato consiste nel vedere se il robot, udendo la storia, si comporta come un normale essere umano. Come si comporta un normale essere umano quando gli si racconta una storia? Questa è la domanda fondata: alla quale i comportamentisti sono stati interessati e alla quale Searle e i suoi compagni mentalisti potrebbero anche profittevolmente dedicare la loro attenzione

quando cessano di fantasticare su ciò che succede dentro la testa. La mitologia neurale che Searle suggerisce non è migliore della mitologia del programma di computer dei ricercatori di IA.

Searle è disposto ad abbandonare la tesi dell'intenzionalità (in un robot) non appena scopre che un computer la stava dopo tutto gestendo. Qui c'è un perfetto esempio di come i concetti cognitivi possano servire come maschera per l'ignoranza. Si dice che il robot pensa, finché non scopriamo come lavora. Allora si dice che non pensa. Ma supponiamo, contrariamente alle supposizioni di qualcuno, che tutte le proprietà funzionali del cervello umano fossero scoperte. Allora il "robot umano" sarebbe smascherato, e noi potremmo altrettanto bene abbandonare la tesi dell'intenzionalità pure per gli umani. È solo il comportamentista, sembra, che è intenzionato a mantenere termini come pensiero, intenzionalità e simili (come modelli di comportamento). Ma non ci sono "stati mentali sottostanti… il comportamento" nel modo che uno scheletro sta sotto la nostra struttura corporea. Il modello del comportamento è lo stato mentale. Questi modelli sono i risultati di fattori interni ed esterni nel presente e nel passato — non di uno stato mentale che controlla — anche se identificato con il cervello.

Che l'identificazione della mente col cervello non regga è ovvio dalla considerazione di un altro esempio che oserei dire sarà portato in campo da altri commentatori — anche ricercatori di IA — tanto è ovvio. Chiamiamolo "La risposta del cervello di Donovan (Hollywood)". Un cervello viene tolto da un umano adulto normale. Il cervello è posto dentro uno scaffale per computer con il solito congegno input-output — memorie, tastiera, video e così via. Il cervello è collegato al congegno da una serie di meccanismi che possono stimolare tutti i nervi che il corpo può stimolare nel misurare lo stato di tutti i nervi che influenzano il movimento muscolare. Il cervello, destinato a interagire con un corpo, non farà certamente meglio (anzi farà probabilmente molto peggio) nell'operare, rispetto a un meccanismo standard di computer designato a tale proposito. Questo "robot" si conforma al criterio di Searle relativo a una macchina pensante — in effetti è una macchina pensante ideale dal suo punto di vista. Ma sarebbe ridicolo dire che potrebbe pensare. Una macchina che non può comportarsi come un essere umano non può, per definizione, pensare.

Il misticismo come filosofia dell'Intelligenza Artificiale

Martin Ringle
Dipartimento di Scienza del computer, Vassar College, Poughkeepsie, NY 12601

Searle identifica un punto debole nella metodologia di Intelligenza

Artificiale che è certamente degno di essere analizzato. Egli rileva che focalizzando l'attenzione con un alto livello di analisi cognitiva, l'Intelligenza Artificiale ignora il ruolo fondamentale che le proprietà fisiche hanno nella determinazione dell'intenzionalità. Il caso può essere definito così: negli esseri umani l'analisi dei caratteri percepiti del mondo include la diretta attività fisica delle strutture e sottostrutture nervose tanto quanto le interazioni causali tra il sistema nervoso e i fenomeni fisici esterni. Quando definiamo un "programma" come una spiegazione (o, al minimo, come una descrizione) di un processo cognitivo, noi astraiamo gli elementi di informazione a un livello arbitrario di risoluzione e presupponiamo le costrizioni e i contributi come dati a livelli inferiori. L'Intelligenza Artificiale sbaglia, secondo Searle, a dimenticare la forza di questo presupposto e a sostenere perciò che la messa in opera da parte del computer del programma stabilito mostrerà da sé le proprietà intenzionali del fenomeno umano originale.

La teoria dell'Intelligenza Artificiale, naturalmente, sostiene che le proprietà di livello inferiore sono irrilevanti rispetto al carattere dei processi cognitivi di livello più alto — seguendo così la vecchia grande tradizione inaugurata da Turing (1964) e Putnam (1960).

Se questo è infatti il nocciolo della disputa tra Searle e l'Intelligenza Artificiale, allora è di interesse filosofico relativamente piccolo. Poiché finisce col non dire nulla di più del fatto che ci possono essere importanti processi di informazione che avvengono ai livelli intranervosi e sottonervosi, e questa questione può essere decisa solo empiricamente. Se risulta che tali processi non esistono, allora gli approcci attuali in Intelligenza Artificiale sono giustificati; se, d'altro lato, l'affermazione di Searle è corretta, allora l'Intelligenza Artificiale deve usare processi di livello più basso nei suoi modelli cognitivi. In sostanza, la simulazione dei processi subnervosi su scala abbastanza larga per essere sperimentalmente significativa potrebbe dimostrarsi impossibile (almeno con la tecnologia che è attualmente disponibile). Questo è tutto troppo probabile e, se si prova che sia vero, suonerebbe come una sentenza metodologica per l'Intelligenza Artificiale, almeno così come noi la conosciamo ora. Tuttavia, questo avrebbe scarsa importanza filosofica poiché l'incapacità di modellare l'interfaccia tra i processi subnervosi e internervosi complessi costituirebbe un fallimento tecnico e non teorico. Ma Searle vuole molto più di questo. Egli basa un rifiuto dell'adeguamento dei modelli di Intelligenza Artificiale sulla fiducia che le proprietà fisiche dei sistemi nervosi sono tali che, per principio, non possono essere simulate da un sistema di computer non protoplasmico. Questo è il motivo per cui Searle si rifugia in quello che può solo essere definito come misticismo.

Searle si riferisce alle proprietà privilegiate dei sistemi nervosi protoplasmici come a "poteri causali". Posso scoprire almeno due plausibili

interpretazioni di questo termine, ma nessuna delle due soddisferà l'argomentazione di Searle. La prima interpretazione di "potere causale" riguarda il collegamento diretto del sistema nervoso con i fenomeni fisici del mondo esterno. Per esempio, quando un essere umano analizza le immagini visuali, la ricchezza dell'informazione interna risulta dall'interazione fisica diretta col mondo. Quando un computer analizza una scena, non è necessario alcun legame effettivo tra i chiari fenomeni del mondo e una "rappresentazione" interna della macchina. Poiché la rappresentazione interna è il risultato di un programma stabilito, si potrebbe (e spesso si fa in IA) mettere dentro la "rappresentazione" a mano, cioè, senza alcun apparecchio fisico e visivo. In tal caso, il legame causale tra gli stati del mondo e gli stati interni della macchina è semplicemente stipulato. Andando avanti di un altro passo, possiamo dedurre che senza un tale legame causale, gli stati interni non possono essere considerati come stati cognitivi, poiché mancano di qualunque contenuto semantico indipendente da un programma. Gli addetti all'Intelligenza Artificiale potrebbero tentare di rimediare la situazione introducendo trasduttori sensori appropriati e meccanismi mobili (come le "mani") entro i loro sistemi, ma temo che Searle potrebbe ancora confermare la sua opinione col sostenere che i poteri causali di tale sistema mancherebbero ancora di rispecchiare i precisi poteri causali del sistema nervoso umano. La premessa che Searle usa a proprio vantaggio, comunque, è che "nient'altro che un sistema che partecipasse delle proprietà fisiche dei nostri sistemi mostrerebbe precisamente lo stesso genere di legami causali".

Eppure, se la causalità alla quale Searle è interessato non include nulla più che la diretta connessione tra processi interni e stati motosensori, sembrerebbe che egli parli realmente di proprietà funzionali, non fisiche. Egli non può asserire che una cellula fotoelettrica è incapace di catturare lo stesso genere di informazione di un filamento organico o cono in una retina umana a meno che non possa specificamente identificare una deficienza di fondo del primo rispetto al secondo. E questo non lo fa. Possiamo riassumere dicendo che i "poteri causali", in questa interpretazione, presuppongono una locazione fisica, ma che nessun abbellimento fisico particolare è richiesto per un corpo. Connettere effettivi meccanismi sensomotori a un processore interno dovrebbe, perciò, soddisfare requisiti di causalità di questo genere (rimuovendo il carattere stipulazionale degli stati interni).

Nella seconda intepretazione il termine "poteri causali" si riferisce alle capacità dei neuroni protoplasmici di produrre stati fenomenici, come le sensazioni, i dolori e simili. Qui Searle sostiene che cose come le automobili e le macchine da scrivere, a causa della loro composizione fisica inorganica, sono categoricamente incapaci di causare sensazioni, e che questo aspetto di consapevolezza è fondamentale per la intenzionalità.

Ci sono due risposte a questa tesi. Primo, ragionando con Dennett, Schank

e altri, potremmo dire che Searle è in errore nella sua opinione che l'intenzionalità necessariamente richiede sensazioni, che infatti i componenti funzionali delle sensazioni sono tutto ciò che è richiesto per un modello cognitivo. Ma se anche accettiamo la spiegazione di Searle dell'intenzionalità, la tesi sembra ancora insostenibile. Il semplice fatto che fenomeni mentali come le sensazioni sono stati, parlando storicamente, confinati a organismi protoplasmici, in nessun modo dimostra che tali fenomeni potrebbero non sorgere in un sistema nonprotoplasmico. Una tale asserzione è alla pari con l'affermazione (fatta originariamente) che solo creature organiche come uccelli o insetti potrebbero volare. Searle esplicitamente e ripetutamente annuncia che la intenzionalità "è un fenomeno biologico", ma non spiega mai che genere di fenomeno biologico è, né ci dà mai motivo di credere che c'è una proprietà o una serie di proprietà inerente la materia nervosa protoplasmica che non potrebbe, per principio, essere replicata in un sostrato fisico alternativo.

Si può solo concludere che la conoscenza della necessaria connessione tra l'intenzionalità e l'incorporamento protoplasmico è ottenuta attraverso un genere di rivelazione mistica. Questo, naturalmente, non dovrebbe essere troppo fastidioso per i ricercatori di Intelligenza Artificiale che, dopotutto, usano a loro vantaggio il misticismo nella stessa misura di chiunque nella scienza cognitiva oggigiorno. Così va.

Searle e i poteri speciali del cervello

Richard Rorty
Dipartimento di Filosofia, Princeton University, Princeton, NY 08544

Searle organizza la sua tesi come farebbe un cattolico tradizionale che difenda la transustanziazione. Supponiamo che un teologo demitizzante ci spinga a pensare all'eucarestia non in termini di mutamento sostanziale, ma piuttosto nei termini del significato per la vita dei fedeli. Il difensore dell'ortodossia risponderà che "la distinzione fra naturale e soprannaturale non può essere lasciata solo all'occhio dello spettatore ma deve essere intrinseca; altrimenti toccherebbe a ogni spettatore il compito di trattare una cosa come soprannaturale" (cfr. Searle sulla distinzione fra mentale e nonmentale, p. 420). La teologia, dicono gli ortodossi, prende l'avvio da fatti come quello che l'eucarestia cattolica è un evento soprannaturale mentre per un ministro unitariano che porge intorno bicchieri d'acqua, non lo è. Searle dice che "lo studio della mente comincia col fatto che gli esseri umani hanno opinioni, mentre i termostati… e le macchine calcolatrici non le hanno". In teologia, continuano gli ortodossi, si presuppone la realtà e conoscibilità del soprannaturale… Searle dice: "Nelle 'scienze cognitive' si presuppone la

realtà e conoscibilità del mentale". Gli ortodossi pensano che quelli che smitizzano scambiano proprio il soggetto, poiché sappiamo in anticipo che la distinzione tra il naturale e il soprannaturale è una distinzione tra due generi di entità aventi poteri causali speciali diversi. Sappiamo che non possiamo interpretare l'eucarestia "funzionalmente" in termini di utilità perché ci potrebbe essere una tale espressione senza quegli elementi che effettivamente mutano la sua forma sostanziale. Similmente Searle sa in anticipo che uno stato cognitivo "non potrebbe consistere solo di processi computazionali e del loro output perché i processi computazionali e il loro output possono esistere senza uno stato cognitivo". Sia i teologi ortodossi che Searle criticano i loro oppositori come "spudoratamente comportamentisti e operazionalisti". Searle usa l'esempio di essere addestrato per rispondere a domande in cinese al fine di superare un test di Turing. Il difensore della transustanziazione userebbe l'esempio di un laico travestito da prete che celebra la messa e che prende in giro i parrocchiani. La risposta iniziale dell'esempio di Searle è che se l'addestramento durasse per anni e anni, cosicché Searle diventasse capace di rispondere a tutte le possibili domande cinesi in cinese, allora egli capirebbe benissimo il cinese.

Se puoi prendere in giro la gente tutto il tempo, dicono comportamentisti e operazionalisti, questo non è più un prendere in giro. La risposta iniziale all'esempio del teologo ortodosso è che quando i preti anglicani celebrano i riti dell'eucarestia, quello che accade è funzionalmente identico a quello che accade nelle chiese cattoliche a dispetto del "difetto d'intenzione" negli ordini anglicani. Quando hai una massa di fedeli numerosa come la comunità anglicana che prende l'eucarestia senza che i necessari "poteri causali speciali" siano stati presenti, questo dimostra che quei poteri non erano essenziali al sacramento. La simulazione accettata abbastanza ampiamente è la cosa reale. I cattolici, comunque, risponderanno che un'ostia anglicana "consacrata" è il corpo di Cristo non più che un orsacchiotto di felpa è un orso, poiché i "poteri causali speciali" sono l'essenza della cosa. Similmente Searle sa in anticipo che "solo qualcosa che ha gli stessi poteri causali del cervello può avere intenzionalità".

Come sa questo Searle? Allo stesso modo, presumibilmente, che il teologo ortodosso conosce le cose. Searle sa che cosa "mentale" e "cognitivo" e termini simili significano, e così sa che non possono essere appropriatamente applicati in assenza del cervello — o, forse, in assenza di qualcosa che è molto simile a un cervello rispetto ai "poteri causali". Come diremmo che qualcosa è sufficientemente simile? chiedono comportamentisti e operazionalisti. Presumibilmente essi non riceveranno alcuna risposta finché non scopriamo abbastanza sul modo in cui il cervello lavora per distinguere l'intenzionalità da semplici simulazioni di intenzionalità. Come potrebbe una persona neutrale giudicare la disputa tra anglicani e cattolici romani sulla

validità degli ordini anglicani? Presumibilmente dovrà aspettare finché non scopriamo di più su Dio. Ma forse l'analogia è errata: noi moderni crediamo nel cervello ma non in Dio. Pure, anche se mettiamo da parte l'analogia teologica, possiamo avere difficoltà sapendo solo ciò che la ricerca del cervello si suppone cerchi. Dobbiamo scoprire il contenuto piuttosto che semplicemente la forma. Searle ce lo dice, poiché gli stati mentali sono "letteralmente un prodotto dell'operazione del cervello" e da ciò nessuna concepibile descrizione di programma (che dà semplicemente una forma, organizzatile da molti diversi tipi di hardware) andrà bene. I comportamentisti e gli operazionalisti, comunque, ritengono il contenuto e la forma e le distinzioni dell'hardware e del programma come puramente euristiche, relative e pragmatiche. Questo è il motivo per cui sono, se non turbati, almeno diffidenti, quando Searle dichiara che "i reali fenomeni mentali umani potrebbero essere dipendenti dalle reali proprietà fisico-chimiche dei cervelli umani reali". Se si deve prendere questa dichiarazione in un senso controverso, allora sembra proprio un dispositivo per assicurare che i segreti poteri del cervello si spingeranno sempre più indietro, fuori dalla vista, ogni volta che viene proposto un nuovo modello di cervello funzionante. Poiché Searle ci può dire che ogni modello è puramente una scoperta di modelli formali, e che il "contenuto mentale" ci è ancora sfuggito (egli potrebbe comprovare tale suggerimento citando Henri Bergson e Thomas Nagel sull'ineffabile intima natura perfino della creazione bruta). Non c'è dopo tutto gran differenza — per quanto si estenda la distinzione forma-contenuto — tra il costruire modelli per il comportamento degli umani e per quello dei loro cervelli. Senza ulteriori suggerimenti su come riferire il contenuto quando finalmente lo incontriamo, possiamo ben sentire che ogni ricerca

> è un arco attraverso il quale
> irraggia quel mondo non esplorato
> il cui margine svanisce sempre più
> quando mi muovo.
> (Tennison: *Ulysses*)

Le mie critiche a Searle non dovrebbero essere, comunque, prese come una prova di simpatia per l'Intelligenza Artificiale. Nel 1960 Putnam notò che l'analogia fra mente e programma non mostrava che possiamo usare computers per aiutare i filosofi a risolvere il problema mente-corpo, ma che non c'era alcun problema mente-corpo che i filosofi dovessero risolvere. Il meglio del lavoro degli ultimi venti anni di IA ha rafforzato la tesi di Putnam. Né, purtroppo, si è fatto alcunché per aiutare i neurofisiologi — cosa che effettivamente si sarebbe potuta fare. Forse valeva la pena di vedere se programmare i computers poteva produrre utili modelli almeno del cervello, se non del "pensiero" o della "mente". Forse, tuttavia, il denaro speso a fare i

giochi di Turing con computers dispendiosi avrebbe dovuto essere usato per pagare filosofi relativamente a buon mercato, come Searle e me. Per ora avremmo potuto definire esattamente di quali tipi di operazionalismo e comportamentismo ci dovevamo vergognare e di quali no. Concesso che certe prime forme dogmatiche di queste teorie erano un po' grossolane, Peirce aveva ragione nel dire che qualcosa di quel tipo doveva essere preso per vero se dobbiamo minimizzare le dispute sulla transustanziazione. Se l'attuale atteggiamento pre-wittgensteiniano di Searle viene normalmente accettato, il valido lavoro di Ryle e Putnam verrà annullato e "il mentale" recupererà il suo splendore cartesiano. E questo tornerà a vantaggio dell'Intelligenza Artificiale. Gli "scienziati cognitivi" insisteranno nel dire che solo molta più simulazione e denaro riverseranno luce empirica su questi profondi misteri "filosofici". Certo Searle non vuole questo.

Capire Searle

Roger C. Schank
Dipartimento di Computer Science, Yale University, New Haven, Conn. 06520

Che cosa è capire? Che cosa è consapevolezza? Che cosa è significato? Che cosa significa pensare? Queste, certo, sono domande da filosofo. Sono il pane quotidiano della filosofia. Ma che dire del ruolo di tali domande in Intelligenza Artificiale? Non dovrebbero i ricercatori di IA essere ugualmente interessati a tali questioni? Credo che la risposta sia sì e no.

Secondo la distinzione tra IA debole e forte, mi dovrei collocare nel campo dell'ipotesi dell'Intelligenza Artificiale debole con la volontà di muovermi verso il lato forte. In una nota a piè pagina Searle ricorda che egli non dice che io sono necessariamente impegnato nelle due "tesi di Intelligenza Artificiale" che cita. Egli afferma che le dichiarazioni che i computers possono capire le storie o che i programmi possono spiegare il comprendere umano non sono sostenute dal mio lavoro. Egli ha certamente ragione in quella affermazione. Nessun programma che abbiamo scritto si può ancora dire che capisca veramente. Per questo motivo nessun programma che abbiamo scritto "spiega la capacità umana di capire". Sono d'accordo con Searle su ciò per due ragioni. Primo, noi non abbiamo assolutamente finito nel costruire macchine che capiscono. I nostri programmi sono, in questa fase, parziali e incompleti. Non si può dire che veramente capiscano: per questo motivo essi non possono essere nulla più che spiegazioni parziali delle capacità umane.

Naturalmente io capisco che Searle fa un'affermazione più ampia di questa. Searle afferma che i nostri programmi non saranno mai in grado di capire o di

spiegare le capacità umane. La sua ultima tesi è chiaramente del tutto errata. I nostri programmi hanno fornito notevoli supporti a teorie che furono più tardi esaminate su soggetti umani. Tutto il lavoro sperimentale in psicologia finora ha dimostrato, per esempio, che la nostra nozione di *script* (Schank e Abelson 1977) è una spiegazione di capacità umane (cfr. Nelson e Gruendel 1978; Gruendel 1980; Smith, Adams e Schorr 1978; Bower, Black e Turner 1979; Graesser e altri 1979; Anderson 1980).

Tutti gli studi menzionati consistono in resoconti di esperimenti su soggetti umani che sostengono la nozione di uno *script.* Naturalmente Searle può dire che erano le nostre teorie piuttosto che i nostri programmi a spiegare le capacità umane in quell'esempio. In quel caso, posso solo tentare di spiegare accuratamente la mia opinione di quello di cui l'Intelligenza Artificiale si occupa realmente. Non possiamo avere teorie separate dalle nostre implementazioni su computer di quelle teorie stesse. La serie dei fenomeni da spiegare è troppo ampia e dettagliata perché sia coperta da una teoria scritta in inglese. Noi possiamo solo sapere che le nostre teorie della comprensione sono plausibili se possono funzionare essendo sottoposte a verifica su una macchina.

Si lascia Searle con le sue obiezioni contro gli esperimenti psicologici in sé e per sé come test adeguati di teorie di capacità umane. Egli considera la psicologia come irrilevante? L'evidenza suggerisce di sì, sebbene egli non sia così esplicito su questo punto. Questo mi riporta indietro alla sua prima tesi: "Può una macchina capire?". O, per dirla in un altro modo, può un modello processuale di comprensione dirci qualcosa sul comprendere? La questione è pertinente se il bersaglio d'attacco è l'Intelligenza Artificiale o la psicologia.

Per rispondere a questa domanda tenterò di tracciare un'analogia. Provate a spiegare che cosa è la "vita". Possiamo dare varie spiegazioni biologiche della vita. Ma alla fine, io chiedo, qual è l'essenza della vita? Che cosa è che distingue un corpo morto che è fisicamente intatto da un corpo vivo? Sì, certo, i processi sono in corso nel vivo e non lo sono nel morto. Ma come riavviarli di nuovo? Con la scossa di elettricità del Dr. Frankenstein? Qual è l'avvio? Che cosa produce la vita?

I biologi possono dare varie spiegazioni del processo della vita, ma alla fine, quell'elusivo "avvio della vita" rimane misterioso. E così è con il comprendere e la coscienza. Noi attribuiamo la comprensione, la coscienza e la vita ad altri per il motivo che noi stessi abbiamo queste disponibilità. Veramente non sappiamo se qualcun altro "capisce", "pensa" o perfino è "vivo". Lo supponiamo sulla base nient'affatto scientifica che dal momento che noi siamo tutte queste cose, altri lo devono essere a loro volta.

Non possiamo dare spiegazioni scientifiche per alcuno di questi fenomeni. Sicuramente le risposte, formulate in termini chimici, non soddisferebbero Searle. Trovo difficile credere che quello che i filosofi hanno perseguito per

secoli fossero spiegazioni chimiche per quei fenomeni che pervadono la nostra vita. Eppure quella è la posizione che Searle si costringe ad adottare, perché, oltre alla spiegazione chimica, che cosa resta? Abbiamo bisogno di spiegazioni in termini umani, in termini delle entità che incontriamo e con cui trattiamo nella nostra vita quotidiana, che soddisferanno il nostro bisogno di conoscere su queste cose.

Ora ritorno alla mia analogia. Possiamo arrivare alla spiegazione processuale della "vita"? Sì, certo, potremmo costruire un modello che funzioni "come se fosse vivo", un robot. Sarebbe vivo? Lo stesso si può dire riguardo alla coscienza e alla comprensione. Potremmo costruire programmi che funzionano come se capissero o avessero libero pensiero consapevole. Sarebbero consapevoli? Capirebbero realmente? Io ho di questi interrogativi opinioni un po' diverse da quelle della maggior parte dei miei colleghi in Intelligenza Artificiale. Io non attribuisco opinioni ai termostati, ai motori di macchine o ai computers. Le mie risposte alle domande di cui sopra sono negative. Un robot non è vivo. I nostri sistemi che capiscono le storie non capiscono nel senso del termine che significa vera e autentica empatia di sentimento ed espressione.

Possiamo mai sperare di far sì che i nostri programmi "capiscano" a quel livello? Possiamo mai creare la "vita"? Queste sono, dopo tutto, domande empiriche. Infine, la mia obiezione agli appunti di Searle può essere formulata in questo modo. Comprende il cervello? Certo, noi umani comprendiamo, ma comprende davvero quel pezzo di materia che definiamo come il nostro cervello? Tutto quello che accade là sono tante reazioni chimiche e impulsi elettrici, proprio tanti simboli cinesi.

Comprendere significa trovare il sistema dietro i simboli cinesi, se scritto per cervelli o per computers. La persona che scrive le regole perché Searle le usi per produrre i corretti simboli cinesi al tempo giusto — quello è un linguista degno di essere assunto. Il linguista "capisce" nel senso profondo come funziona la lingua cinese. E le regole che scrive incorporano quel comprendere. Searle vuole chiamare in questione l'impresa e l'assunto dell'Intelligenza Artificiale, ma alla fine anche lui deve apprezzare il fatto che le regole per manipolare i simboli cinesi sarebbero un grande risultato. Scriverle richiederebbe una maggior comprensione della natura del linguaggio. Tali regole soddisferebbero molti degli interrogativi della filosofia, della linguistica, della psicologia e dell'Intelligenza Artificiale. Searle, che usa quelle regole, capisce? No. Capisce la configurazione dell'hardware del computer? No. Capisce la configurazione dell'hardware del cervello? No. Chi capisce allora? La persona che scrive le regole, naturalmente. E chi è? È quello che è chiamato un ricercatore di Intelligenza Artificiale.

Come trasformare un processore d'informazione in un soggetto che comprende

Aaron Sloman e Monica Croucher
Scuola di Scienze Sociali, Università del Sussex, Brighton BN I 90N, Inghilterra

Il saggio, deliziosamente chiaro e provocatorio, contiene un sottile errore, che viene spesso commesso anche dai ricercatori di Intelligenza Artificiale che usano un linguaggio mentalistico familiare per descrivere i loro programmi. L'errore è dovuto a una mancanza di distinzione tra forma e funzione.

Il fatto che un meccanismo o processo abbia proprietà che, in un contesto adatto, lo metterebbero in grado di eseguire qualche funzione, non implica necessariamente che già esegua quella funzione. Perché un processo sia intelligente o pensante o qualunque altra cosa, non è sufficiente che esso replichi parte della struttura dei processi del comprendere, del pensare e così via. Deve anche adempiere le funzioni di quei processi. Questo richiede che sia causalmente collegato a un sistema più ampio nel quale altri stati e processi esistono. Searle ha ragione perciò a dare importanza ai poteri causali. Comunque non sono i poteri causali delle cellule del cervello ciò che abbiamo bisogno di considerare, ma piuttosto i poteri causali dei processi computazionali. La ragione per cui i processi che descrive non hanno a che fare col comprendere, non è che essi non sono prodotti di cose che abbiano i giusti poteri causali, ma che essi non hanno i giusti poteri causali, dal momento che non sono integrati nel giusto tipo di sistema totale.

Che certe operazioni su simboli che capitano in un computer, o perfino nella mente di un'altra persona, siano isomorfe a certe operazioni formali nella vostra mente, non causa necessariamente che essi svolgono la stessa funzione nell'economia della vostra mente. Quando leggete una frase avviene un processo complesso, generalmente inconscio, di analisi sintattica e semantica, insieme con varie inferenze, alterazioni della vostra memoria di lungo termine, forse mutamenti nei vostri piani attuali o perfino nei vostri gusti, contrarietà o stati emotivi. Qualcun altro che legga la frase condividerà al massimo una parte di questi processi. Anche se c'è una parte di manipolazioni simboliche formali comuni a tutti quelli che odono la frase, l'esistenza di quei processi formali non costituirà, in assoluto, la comprensione della frase. La comprensione può avvenire solo in un contesto nel quale il processo ha l'opportunità di interagire con eventi come opinioni, motivazioni, percezioni, inferenze e decisioni — perché è incluso in modo appropriato in un sistema generale appropriato. Questo potrebbe sembrare quello che Searle chiama "La replica del robot" attribuita a Yale. Comunque,

non è sufficiente dire che i processi devono avvenire in un sistema fisico intorno al quale ci si muove, si fanno rumori, e così via. Noi dichiariamo che non deve nemmeno essere un sistema fisico: le proprietà del sistema più largo richiesto per l'intenzionalità sono computazionali, non fisiche (questo, diversamente dalla posizione di Searle, spiega perché risulta accettabile da parte della gente normale attribuire stati mentali ad anime disincarnate, angeli e così via, se non anche a dei termostati).

Quale genere di sistema più ampio si richiede? Non è facile rispondere. C'è l'inizio di un'esplorazione dei risultati nei capitoli 6 e 10 di Sloman (1978) e in Sloman (1979) (cfr. anche Dennett 1978). Uno dei problemi centrali è quello di specificare le condizioni alle quali potrebbe essere corretto descrivere un sistema computazionale, incorporato in un cervello umano o no, come avente desideri propri, preferenze, gusti, e altre motivazioni. L'ipotesi che attualmente esploriamo è che tali motivazioni sono tipicamente instanziate in rappresentazioni simboliche di stati, eventi, processi, o criteri di selezione, che giocano un ruolo nel controllare le operazioni del sistema, includendo operazioni che cambiano i contenuti dell'archivio delle motivazioni, come accade appunto quando facciamo in modo (spesso con difficoltà) di cambiare i nostri gusti, o quando un'intenzione viene abbandonata perché si trova in conflitto con un principio. Più generalmente, le motivazioni controlleranno la collocazione delle risorse, includendo la direzione dell'attenzione nei processi percettivi, la creazione di scopi e sottoscopi, il genere di informazioni che sono processate e accumulate per un uso futuro, e le inferenze che vengono fatte, controlleranno anche le azioni esterne se il sistema è connesso a un insieme di "motori" (come i muscoli) sensibili ai segnali trasmessi durante l'esecuzione di piani e strategie. Alcune motivazioni saranno capaci di interagire con le opinioni per produrre i complessi disturbi caratteristici degli stati emotivi, come la paura, l'ira, l'imbarazzo, la vergogna e il disgusto. Una condizione preliminare perché il sistema abbia propri desideri e scopi è che le sue motivazioni devono evolvere come risultato di un processo di reazione durante una lunga sequenza di esperienze, in cui opinioni e serie di concetti si sviluppano. Questo, a sua volta, richiede che il sistema di motivazioni abbia una struttura a numerosi livelli, che noi tenteremo di analizzare ulteriormente qui.

Questa tesi risulta circolare perché usa una terminologia mentalistica, ma la nostra tesi, ed è una tesi non considerata da Searle, è che un'elaborazione ulteriore di queste idee può condurre a una specificazione puramente formale dell'architettura computazionale del sistema richiesto. Frammenti possono già essere trovati in sistemi operanti esistenti (guidati in parte da priorità o interazioni), e nei programmi di Intelligenza Artificiale che interpretano immagini, fabbricano e revisionano programmi, fanno ed eseguono piani. Ma nessun sistema esistente si avvicina a combinare tutte le complicazioni

richieste prima che i processi mentali usuali possano avvenire. Alcune delle forme ci sono già, ma non ancora le funzioni.

L'esperimento di Searle, in cui esegue operazioni incomprese che riguardano simboli cinesi, non comprende operazioni collegate a un sistema appropriato nel modo appropriato. La notizia, data in cinese, che la sua casa è in fiamme, non lo manderà in tutta fretta a casa, anche se in qualche modo egli opera correttamente con i simboli. Ma, ugualmente, nessuno dei cosiddetti programmi prodotti fin qui è collegato a un sistema appropriato più ampio di opinioni e decisioni. Così, per quanto riguarda i significati ordinari delle parole, non è corretto dire che ogni programma esistente di Intelligenza Artificiale comprende, crede, impara, percepisce o risolve problemi. Naturalmente, si potrebbe obiettare (per quanto non da parte nostra) che essi hanno già il potenziale per essere così collegati — hanno una forma che è adeguata alla funzione in questione. Se fosse così, potrebbero forse essere usati come estensioni di facoltà umane — per esempio, come aiuti per i sordi o i ciechi o gli handicappati mentali, e potrebbero allora essere parte di un sistema che comprende.

Si potrebbe obiettare che il linguaggio mentalistico dovrebbe essere esteso a inglobare tutti i sistemi con la potenzialità di essere ben collegati entro una mente completa. Cioè, si potrebbe obiettare che i significati di parole come "capire", "percepire", "intendere", "credere" dovrebbero avere le loro precondizioni funzionali alterate, come se dovessimo cominciare a chiamare le cose "cacciaviti" o "tachimetri" se capitasse loro di avere la struttura appropriata per eseguire le rispettive funzioni, indipendentemente dal fatto che siano mai usati o anche destinati a essere usati con le caratteristiche funzioni di cacciaviti e tachimetri. La giustificazione per estendere l'uso del linguaggio intenzionale e di altri linguaggi mentali in questo modo sarebbe la scoperta che alcuni aspetti dell'architettura più ampia (come la presenza di meccanismi di inferenza) sembrano essere richiesti all'interno di tali sottosistemi isolati in modo da render loro possibile soddisfare anche le precondizioni formali. Comunque, la nostra polemica contro Searle non dipende dalla alterazione dei significati delle parole familiari.

È necessario che un sistema mentale sia capace di controllare le operazioni di un corpo fisico o sia collegato a sensori fisici capaci di ricevere informazioni sull'ambiente fisico? Questo si lega alla domanda se una persona totalmente paralizzata, sorda, cieca, senza alcun organo di senso funzionante potrebbe nondimeno essere consapevole, con pensieri, speranze e timori (notate che questo non è troppo diverso dallo stato in cui le persone normali entrano temporaneamente ogni notte). Vorremmo obiettare che non c'è alcuna ragione (eccezion fatta per insostenibili considerazioni comportamentali) per negare che questa è una possibilità logica. Comunque, se l'individuo non avesse mai interagito col mondo esterno nel modo

normale, non potrebbe pensare al presidente Carter, a Parigi, alla battaglia di Hastings, o anche al proprio corpo: nel migliore dei casi i suoi pensieri ed esperienze si riferirebbero a entità non esistenti in un mondo immaginario. Questo perché una referenzialità corretta presuppone relazioni causali che non reggerebbero nel caso in cui la nostra mente non fosse basata su connessioni.

Si potrebbe pensare che abbiamo travisato il punto essenziale della tesi di Searle dal momento che, indipendentemente dal tipo di architettura computazionale che noi alla fine ipotizziamo per una mente, sia esso connesso o sconnesso, si potrà sempre ripetere il suo esperimento per mostrare che un sistema manipolatore di simboli puramente formali, con quella struttura, non avrebbe necessariamente motivazioni, opinioni od oggetti di percezione. Infatti esso potrebbe eseguire tutti i programmi da sé (almeno per principio) senza avere alcuno dei presunti desideri, opinioni, percezioni, emozioni, o qualsiasi altra cosa.

A questo punto la tesi delle “altre menti” prende una svolta curiosa. Searle si considera un’autorità decisiva nelle questioni, come se quello che accade nelle sue attività mentali includa il vedere (il credere di vedere) elefanti rosa, il pensare al teorema di Pitagora, l’avere paura di essere bruciato al rogo, o il comprendere frasi in cinese. In altre parole, afferma, senza discussione, che è impossibile che un’altra mente sia basata sui suoi processi mentali senza che lui stesso lo sappia. Tuttavia noi dichiariamo (cfr. la discussione della consapevolezza in Sloman 1978, cap. 10) che se egli esegue davvero fedelmente tutti i programmi, a condizione che vi sia un’adeguata distribuzione di tempo tra sottosistemi paralleli dove è necessario, allora avverrà una collezione di processi mentali, della cui natura egli sarà ignorante, se tutto quello che pensa di star facendo è di manipolare simboli senza significato. Egli non avrà più motivi per negare l’esistenza di tali processi mentali rispetto a quelli che avrebbe se questi fossero presentati nei dettagli del lavorio di basso livello della mente di un’altra persona, e se lui stesso capisse solo in termini di processi elettrici e chimici, o forse in sequenze di modelli astratti inseriti in tali processi.

Se le istruzioni che Searle sta eseguendo richiedono che si usino informazioni su cose che egli percepisce nell’ambiente come base per scegliere alcune delle operazioni formali, sarebbe allora perfino possibile per un “visitatore” acquisire informazioni su Searle (col fare deduzioni dal comportamento di Searle e da quello che altre persone dicono su di lui) senza che Searle capisca che cosa succede. Forse questo non è troppo dissimile da quello che accade in alcuni casi di personalità multiple.

Giochi di simulazione

William E. Smythe

Dipartimento di Psicologia, Università di Toronto, Toronto, Ontario, Canada
S IAI

Un uso estensivo di idiomi intenzionali è ora comune nelle discussioni delle capacità e del funzionamento dei sistemi di Intelligenza Artificiale. Spesso queste descrizioni devono essere considerate non più concretamente di quanto avviene in molta dell'ordinaria programmazione dove si potrebbe dire, per esempio, che un programma di regressione "vuole" minimizzare la somma dei quadrati degli scarti o "crede" di avere trovato la funzione più adatta quando ha fatto così. Questa pratica richiede almeno un certo coinvolgimento rispetto alla tesi che afferma che si possono ottenere stati intenzionali in una macchina proprio in virtù del fatto che essa esegue certi calcoli. L'articolo di Searle serve come indicatore valido e tempestivo di alcuni dei tranelli che accompagnano una tale tesi. Se certi sistemi di IA devono possedere intenzionalità, mentre altri sistemi computazionali no, dovrebbe essere in virtù di una qualche serie di princìpi puramente computazionali. Tuttavia, come Searle indica, nessuno di tali princìpi è già emerso in Intelligenza Artificiale. Inoltre, c'è ragione per credere che non lo sarà mai. Un sunto di tali tesi è il seguente: gli stati intenzionali sono, per definizione, "diretti a" oggetti e stati del mondo. Donde il primo requisito per ogni teoria su di essi sarebbe di specificare la relazione tra gli stati e il mondo "intorno" al quale essi sono. Tuttavia è precisamente questa relazione che non è parte della considerazione computazionale degli stati mentali (cfr. Fodor 1980).

Un sistema computazionale può essere comparato e relazionato con un ambiente esterno, qualunque modo un utente umano possa scegliere. Non c'è dipendenza di questa relazione da alcuna storia ontogenetica o filogenetica di interazione con l'ambiente circostante. Infatti la relazione tra sistema e ambiente può essere nulla senza compromettere i calcoli eseguiti su simboli che intenzionalmente si riferiscono a essa. Questo fatto provoca considerevoli riserve e dubbi riguardo al problema se una teoria puramente computazionale di intenzionalità sia possibile. Searle cerca di stabilire una conclusione anche più forte: la sua tesi è che la realizzazione computazionale degli stati intenzionali è, in effetti, impossibile su basi a priori. La tesi è basata su un "gioco di simulazione" della specie del gioco di imitazione di Turing nel quale l'uomo imita il computer. Nel gioco di simulazione un agente umano instanzia un programma di computer eseguendo operazioni puramente sintattiche su simboli senza senso. Il punto della dimostrazione è che la semplice applicazione di regole per l'esecuzione di tali operazioni non è sufficiente per manifestare il giusto genere di intenzionalità. In particolare, una data serie di regole potrebbe creare un'effettiva imitazione di qualche attività intelligente senza per questo portare l'agente, che segue le regole, più vicino ad avere stati intenzionali pertinenti all'ambito in questione.

Una difficoltà con questa tesi è che non distingue tra due modi fondamentalmente diversi di instanziare un programma di computer o altro esplicito procedimento in un sistema fisico. Un modo è di includere il programma in un sistema che è già capace di interpretare e seguire regole. Questo richiede che il procedimento sia espresso in un "linguaggio" che il sistema inglobante possa già "capire". Un secondo modo è di instanziare il programma direttamente realizzando le sue "regole" come primitive operazioni di hardware. In questo caso una regola è seguita, non con l'"interpretarla", ma con l'eseguire semplicemente qualunque procedimento la regola comporti. Il gioco di simulazione di Searle è appropriato e pertinente al primo genere di instanziazione, ma non al secondo. Seguire regole in lingua naturale (come il gioco di simulazione richiede) coinvolge la mediazione di altri stati intenzionali e così è necessariamente un esempio di instanziazione indiretta. Per imitare un'instanziazione diretta di un programma, d'altro lato, i relativi primitivi dovrebbero essere realizzati non mediamente, ma nell'attività del singolo. Se tale imitazione fosse possibile, ciò avverrebbe solo a costo di essere incapace di notare nel sistema la mancanza di stati intenzionali, se davvero non ne avesse alcuno.

La distinzione tra procedimenti computazionali direttamente e indirettamente instanziati è importante perché entrambi i tipi di processi sono richiesti per specificare completamente un sistema computazionale. Il primo tipo contiene la sua architettura o insieme di costituenti primitivi, e il secondo comprende gli algoritmi che il sistema può applicare (Newell 1973, 1980). Perciò la tesi di Searle è una sfida all'ipotesi dell'IA forte, quando quell'ipotesi è presentata in termini di capacità dei programmi, ma non quando è racchiusa (come, per esempio, in Pylyshyn 1980a) nei termini di sistemi computazionali. La dichiarazione che questi ultimi non possono avere stati intenzionali deve perciò procedere lungo linee ben diverse. L'approccio considerato prima, per esempio, richiamava l'attenzione sull'arbitraria relazione tra simbolo computazionale e referente. Altrove questo argomento è stato proposto più dettagliatamente di quanto non lo sia una nozione troppo restrittiva di simbolo che crea le più serie difficoltà per la teoria computazionale (Kolers e Smythe 1979, Smythe 1979). La nozione di un elemento soggetto solo a manipolazioni sintattiche formali non è né una sufficiente caratterizzazione di ciò che è un simbolo, né è ben motivato nell'ambito della cognizione umana. Per quanto ragionevole sia questa tesi, non è la conclusione definitiva che il gioco di simulazione di Searle tenta di dimostrare. Comunque, il gioco di simulazione getta veramente luce su un altro punto. Come mai l'opinione che i sistemi di Intelligenza Artificiale sono veramente costitutivi degli eventi mentali è così dilagante? Una risposta è che molta gente sembra svolgere una versione del gioco di simulazione diversa da quella che Searle raccomanda.

I simboli della maggior parte dei sistemi di Intelligenza Artificiale e di simulazione cognitiva sono raramente quel genere di simboli senza significato che il gioco di simulazione di Searle richiede. Al contrario, essi sono spesso manifestati ed estrinsecati in forme che portano una notevole quantità di extrasignificato all'utente, sopra e al di sopra dell'identità procedurale del sistema stesso, come iscrizioni pittoriche e linguistiche, per esempio. Questo genere di realizzazione dei simboli può portare a seri problemi teorici.

Per esempio, sistemi come quello di Kosslyn e Schwartz (1977) danno l'impressione di operare ampiamente su immagini mentali perché le rappresentazioni interne "hanno l'aspetto" delle immagini proiettate da un tubo a raggi catodici. Non è chiaro che si possa dire che il sistema manipoli immagini in alcun altro senso. C'è un problema simile con i sistemi di comprensione linguistica. La semantica di tali sistemi è spesso stabilita per mezzo di un procedimento informale che Hayes (1977, p. 559) chiama "fingi-che-sia-inglese". Questo significa che conclusioni disorientanti sulle capacità di questi sistemi possono risultare dalla rassomiglianza superficiale che le loro rappresentazioni interne presentano rispetto ad affermazioni in lingua naturale. Un'importante virtù della tesi di Searle è che essa specifica come svolgere correttamente il gioco di simulazione. La realizzazione procedurale dei simboli è tutto quello che dovrebbe importare in una teoria computazionale; la loro apparenza esterna dovrebbe essere irrilevante. Il gioco, giocato in questo modo, può non stabilire chiaramente che i sistemi computazionali mancano di intenzionalità. Tuttavia, almeno scalza una potente motivazione tacita per supporre che l'abbiano.

Il termostato e il professore di filosofia

Donald O. Walter
Istituto di ricerche sul cervello e Dipartimento di Psichiatria, Università di California, Los Angeles, Calif. 30024

Searle — L'uomo certamente non capisce il cinese, e neppure i tubi dell'acqua, e se siamo tentati di accettare quella che penso sia un'opinione assurda, che cioè in qualche modo la combinazione di uomo e tubi dell'acqua possono capire…

Walter — La striscia bimetallica di per sé certo non mantiene la temperatura entro i limiti, e neppure il forno di per sé, e se siamo tentati di accettare l'opinione che in qualche modo un sistema striscia bimetallica più forno manterrà la temperatura entro i limiti, o (parafrasando Hanson 1969, o altri) la retina sinistra di Searle non vede, e neppure la destra, e né l'uno o l'altro dei nervi ottici; possiamo perfino immaginare una "sindrome di

disconnessione" nella quale la corteccia ottica di Searle non è più connessa col resto del suo cervello, e così concludere che la sua corteccia ottica non vede, ancora se poi concludiamo che, poiché nessuna parte vede, per questo egli non può vedere, mostriamo in tal modo coerenza o manchiamo di vedere qualche cosa sui nostri propri concetti?

Searle — Nessuno suppone che le simulazioni del computer di un incendio bruceranno il vicinato… Perché mai supporre che una simulazione da parte del computer di una comprensione potrebbe realmente comprendere qualcosa?

Walter — Nessuno suppone che la descrizione da parte di un romanziere di un incendio brucerà il vicinato: perché si dovrebbe supporre che un romanziere che scriva sul comprendere veramente lo comprenda?

Searle — Se sapessimo come render conto del suo comportamento indipendentemente da tali considerazioni, non attribuiremmo intenzionalità a esso, specialmente se sapessimo che ha un programma formale (Hofstadter 1979, p. 601). C'è un "teorema" sul progresso in Intelligenza Artificiale: una volta che una funzione mentale è programmata, la gente subito cessa di considerare che è un ingrediente essenziale del "pensare reale". L'inevitabile centro di intelligenza è sempre quella cosa successiva che non è stata ancora programmata.

Walter — Searle sembra essere certo che un programma è formale (sebbene egli giochi, a suo vantaggio, sull'ambiguità tra "forma adeguatamente definibile" e "forma completamente definibile"), mentre "intenzionalità", "poteri causali" e "effettive proprietà" sono cose radicalmente diverse, che sono indiscutibilmente presenti in ogni cervello umano (normale sveglio?), e forse nei bizzarri cervelli dei "marziani" (se fossero "vivi", almeno nel senso che noi non capissimo che cosa succede dentro di loro). Queste cose radicalmente diverse sono pure non definibili nei termini della loro forma, ma del loro contenuto. Egli asserisce ciò ripetutamente, senza rendere esplicito alcunché di questa vitale alternativa. Tocca a Searle stabilire la comunicazione con i lettori, cosa che egli non ha fatto col suo articolo. Speriamo che nella sua risposta ci renda più esplicita l'alternativa menzionata, ma non descritta.

Computers, cognizione e filosofia

Robert Wilensky
Dipartimento di Ingegneria Elettronica e Computer Science, Università di California, Berkeley, Calif. 94720

Le tesi di Searle sulla plausibilità della comprensione da parte di un computer contengono parecchie incrinature logiche, semplici ma fatali. Posso

trattare qui solo delle difficoltà più importanti. Tuttavia è un attacco alla sostanza delle osservazioni di Searle piuttosto che alle pecche tecniche dei suoi assunti che motiva questo commento. L'articolo di Searle suggerisce che anche la miglior simulazione di un comportamento intelligente non spiegherebbe nulla sulla cognizione, e produce argomenti a sostegno di questa tesi.

Poiché vorrei dichiarare che la simulazione del computer può produrre importanti intuizioni sulla natura dei processi cognitivi umani, è importante mostrare perché le ragioni di Searle non minacciano questo progetto. La mia principale obiezione all'argomento di Searle egli l'ha definita "La replica del sistema di Berkeley". La tesi afferma che lo "scenario dell'uomo nella stanza" non presenta alcun problema per un sostenitore di IA forte, che sostiene che il comprendere è una proprietà di un sistema di trattamento dell'informazione. L'uomo nella stanza col registro, funzionante nella maniera prescritta dal teorico cognitivo che instanziò il suo comportamento, costituisce un tale sistema. L'uomo funzionante nella sua normale maniera quotidiana è un altro sistema. Il sistema dell'"uomo ordinario" può non conoscere il cinese, ma questo non dice nulla sulle capacità del sistema dell'"uomo nella stanza", che deve perciò rimanere almeno un candidato per la considerazione di uno che comprende in vista delle sue capacità di analisi del linguaggio.

La risposta di Searle a questo argomento è di fare interiorizzare nell'uomo il sistema "uomo nella stanza" conservando tutte le regole e i calcoli nella testa. Egli ora comprende l'intero sistema. Searle ribatte che se l'uomo "non capisce, non c'è alcun modo per cui il sistema capisca, perché il sistema è solo una parte di lui". Tuttavia questo è proprio completamente errato. Tanti sistemi (in effetti, i sistemi più interessanti) sono incorporati in altri sistemi di capacità più deboli. Per esempio, la struttura di un computer può non essere capace di moltiplicare polinomi, o di analizzare il linguaggio naturale, anche se i programmi scritti per quei computers lo possono fare: i neuroni individuali probabilmente non hanno molta — o forse nessuna — capacità di comprensione, sebbene i sistemi che costituiscono possano capire in misura notevole.

La difficoltà nel comprendere la posizione del sistema nel caso del paradosso di Searle è nell'essere capace di vedere la persona come consistente di due sistemi separati. Le seguenti elaborazioni possono essere utili. Supponi che decidessimo di risolvere la questione una volta per tutte semplicemente col chiedere alla persona interessata se comprende il cinese. Noi diamo alla persona un pezzo di carta con caratteri cinesi che significano (liberamente tradotto): "Capisci il cinese?". Se il sistema dell'uomo-nella-stanza dovesse rispondere, facendo le appropriate manipolazioni di simboli, restituirebbe una striscia di carta col messaggio: "Certo che comprendo il cinese! Cosa pensi che io faccia? Stai scherzando?". Segue allora un dialogo caloroso, dopo il

quale ci scusiamo col sistema uomo-nella-stanza per le nostre scortesi insinuazioni. Subito dopo, ci accostiamo all'uomo stesso (cioè, gli chiediamo di smettere di giocare con i pezzi di carta e di parlare a noi direttamente) e gli chiediamo se per caso conosce il cinese. Egli negherà naturalmente di conoscerlo.

L'errore che fa Searle nell'identificare le esperienze di un sistema con quelle del suo sistema di programmazione è un errore che i filosofi spesso fanno quando si riferiscono ai sistemi di Intelligenza Artificiale. Per esempio, Searle dice che il sottosistema inglese sa che gli "hamburgers" si riferiscono agli "hamburgers", ma che il sottosistema cinese conosce solo i simboli formali. Ma è in realtà l'homunculus che è consapevole della manipolazione dei simboli, e non ha alcuna idea di quale sia il compito in cui egli è impegnato. Il sistema parassitario è implicato in questo compito a livello più alto, e non ha alcuna conoscenza del fatto che è compiuto per mezzo della manipolazione dei simboli più di quanto noi siamo consapevoli di come i nostri propri processi cognitivi vengono effettuati.

Quello che è insolito in questa situazione non è che un sistema è incluso in uno più debole, ma che il sistema di programmazione è tanto più potente di quello che è necessario sia. Cioè, l'homunculus è dotato di comprensione in modo del tutto autonomo, e opera usando solo una piccola percentuale della sua capacità di organizzare simboli. Se sostituiamo l'uomo con un congegno capace di eseguire solo queste operazioni, la tentazione di considerare i sistemi come identici diminuisce grandemente. È importante mettere in evidenza, contrariamente alla dichiarazione di Searle, che la posizione stessa del sistema non costituisce una tesi di Intelligenza Artificiale forte. Semplicemente mostra che se è possibile che un sistema diverso da una persona operante nella maniera standard possa capire, allora l'argomento uomo-nella-stanza non è affatto problematico. Se neghiamo questa possibilità, allora il delicato tema dell'uomo-nella-stanza non è necessario — un programma di computer è qualcosa di diverso da una persona che opera normalmente, e per ipotesi non sarebbe capace di comprendere.

Searle presenta pure una tesi sulla simulazione in generale. Egli afferma che, poiché la simulazione di una tempesta non ci lascia bagnati, perché dovremmo supporre che una simulazione del comprendere dovesse capire? Ebbene, la ragione è che mentre le simulazioni non conservano necessariamente *tutte* le proprietà di quello che simulano, conservano necessariamente *particolari* proprietà. Io potrei simulare una tempesta in laboratorio spruzzando acqua attraverso un tubo di gomma. Se sono interessato a studiare proprietà particolari, non devo rinunciare alle simulazioni: semplicemente devo stare attento a quali proprietà è probabile che la simulazione che costruisco conservi effettivamente.

Così tutto va a finire nella domanda: "Che genere di cosa è il

comprendere?” Se è una cosa inerentemente fisica, come il fuoco o la pioggia o la digestione, conservare le proprietà logiche del comprendere non conserverà in effetti la natura essenziale del fenomeno, e una simulazione del computer non comprenderà. Se invece comprendere è essenzialmente un tipo di attività logico e simbolico, conservare le sue proprietà logiche sarebbe sufficiente per avere comprensione, e una simulazione da parte del computer letteralmente comprenderà.

L’assunto di Searle è che il termine “comprendere” si riferisce a un fenomeno fisico, proprio allo stesso modo del termine “fotosintesi”. La sua tesi qui è strettamente un appello alle nostre intuizioni sul significato di questo termine. Le mie intuizioni semplicemente non coinvolgono le proprietà causali degli organismi biologici (anche se coinvolgono le loro proprietà logiche e comportamentali). Mi sembra che questo debba essere vero per la maggior parte della gente, poiché la maggior parte della gente potrebbe essere raggirata e indotta a pensare che una simulazione del computer realmente capisca, ma una simulazione della fotosintesi non indurrebbe nessuno a pensare che avesse effettivamente creato acqua e anidride carbonica.

Un tema importante nell’articolo di Searle è che l’intenzionalità è realmente alla base del problema. I computers falliscono nel trovare i criteri del vero comprendere perché appunto non hanno stati intenzionali, con tutto quello che ciò comporta. Ciò, secondo Searle, è in effetti quello che falsa le intuizioni nell’esempio uomo-nella-stanza. Comunque mi sembra che l’argomento di Searle non abbia niente a che fare con l’intenzionalità. Quello che causa difficoltà nell’attribuire stati intenzionali alle macchine è il fatto che la maggior parte di questi stati ha anche una natura soggettiva. Se questo è il caso, l’esempio uomo-nella-stanza di Searle potrebbe essere usato per simulare una persona avente un certo stato non intenzionale, ma soggettivo, e avere ancora il suo effetto desiderato. Ciò è precisamente quello che accade. Per esempio, supponi che simulassimo qualcuno sofferente di ansietà indiretta. È difficile credere che qualcosa — l’uomo che fa la simulazione o il sistema che effettua — faccia effettivamente esperienza di ansietà indiretta, anche se questo non è uno stato intenzionale. Inoltre, l’esperienza del disagio sembra proporzionale alla soggettività, ma indipendentemente dall’intenzionalità. Non disturba molto le mie intuizioni sentire che un computer può capire o conoscere qualcosa; che creda poi che qualcosa è un po’ più difficile da mandar giù, e che abbia amore, odio, rabbia, dolore e ansietà è molto più difficile. Nota che la soggettività sembra aumentare in ogni caso, ma l’intenzionalità rimane la stessa. Il punto è che l’argomento di Searle non ha nulla a che fare con l’intenzionalità di per sé, e non getta luce sulla natura degli stati intenzionali o sui tipi di meccanismi capaci di averli.

Vorrei ricapitolare dicendo un’ultima parola sull’esperimento dell’uomo-

nella-stanza usato da Searle, poiché questo forma la base per la maggior parte dei suoi argomenti successivi. Woody Allen descrive una bestia mitica chiamata il Grande Capriolo. Quest'ultimo ha la testa di un leone e il corpo di un leone, ma non dello stesso leone. L'esempio di Searle è realmente come un Grande Capriolo, con la testa e il corpo di uno che comprende, ma non della stessa persona che comprende. E proprio qui sta la difficoltà.

RISPOSTA DELL'AUTORE

Intenzionalità intrinseca

Sono lieto del grande interesse che il mio articolo ha sollevato e grato che una così alta percentuale dei commenti sia profonda ed efficace. In questa risposta tenterò di rispondere a ogni critica importante alla mia tesi. Per fare questo, tuttavia, è necessario che io renda pienamente espliciti alcuni dei punti che erano impliciti nell'articolo, poiché questi punti coinvolgono temi ricorrenti nei commenti ricevuti.

Intelligenza Artificiale forte. Uno dei pregi dei commenti è che essi rendono chiaro il carattere estremo della tesi di Intelligenza Artificiale forte. La tesi implica che fra tutti i tipi noti di processi specificamente biologici, dalla mitosi e meiosi alla fotosintesi, digestione, lattazione e secrezione di auxina, uno e un solo tipo è completamente indipendente dalla biochimica delle sue origini, e questo è appunto costituito dalla cognizione. La ragione per cui è indipendente è che la cognizione consiste interamente di processi computazionali, e poiché quei processi sono puramente formali, qualunque sostanza sia capace di instanziare il formalismo è capace di cognizione. Capita appunto che il cervello sia uno tra gli innumerevoli diversi tipi di computers capaci di cognizione, ma i computers fatti di tubi per l'acqua, carta igienica e pietre, fili elettrici — qualunque cosa di solido e resistente abbastanza per portare il giusto programma — avrà necessariamente pensieri, sentimenti, e il resto delle forme di intenzionalità, perché è tutto quello in cui consiste l'intenzionalità: cioè, instanziare i giusti programmi.

Il punto dell'Intelligenza Artificiale forte non è che se costruissimo un computer grande abbastanza o complesso abbastanza per portare gli effettivi programmi che il cervello presumibilmente instanzia, noi otterremmo l'intenzionalità come sottoprodotto (*contra* Dennett), ma piuttosto che l'intenzionalità non è nulla di diverso dall'instanziazione dei giusti programmi.

Ora io trovo la tesi di IA forte insostenibile in ogni senso della parola. Ma non è abbastanza trovare una tesi incredibile, si deve avere una ragione, e io offro una ragione che è molto semplice: instanziare un programma non può in sé e per sé essere costitutivo di intenzionalità, perché sarebbe possibile per un agente instanziare il programma e ancora non avere il giusto genere di intenzionalità. Questa è la base dell'esempio della stanza cinese. Molto di quello che segue interesserà la forza di tale argomento.

Intuizioni

Parecchi commentatori (Block, Dennett, Pylyshyn, Marshall) sostengono che l'argomento è basato su intuizioni mie, e che tali intuizioni, le cose che noi ci sentiamo inclini a dire, non potrebbero mai provare quel genere di cosa che io sto cercando di provare (Block), o che possono essere generate intuizioni contrarie e ugualmente valide (Dennett), e che la storia della conoscenza umana è piena della confutazione di tali intuizioni quali quella che la Terra è piatta o che la tavola è solida, per cui ne segue che le intuizioni qui non hanno alcuna forza.

Ma consideriamo. Quando dico che io in questo momento non capisco il cinese, questa dichiarazione non registra semplicemente una mia intuizione, qualcosa che io mi trovo propenso a dire. È un semplice fatto riguardante me il fatto che io non capisco il cinese. Inoltre, in una situazione in cui mi si dà una serie di regole per manipolare simboli cinesi non interpretati, regole che non consentono alcuna possibilità di attaccare alcun contenuto semantico a questi simboli cinesi, è ancora un fatto che riguarda me il fatto che io non capisco il cinese. In effetti, è proprio lo stesso fatto di prima. Ma, propone Wilensky, supponete che fra quelle regole per manipolare simboli ve ne siano alcune che sono: "Comprendi il cinese?" in cinese, e che in risposta a queste io restituisca i simboli cinesi per "Certo che capisco il cinese". Dimostra questo, come suggerisce Wilensky, che c'è un sottosistema in me che capisce il cinese? Fintantoché non c'è alcun contenuto semantico attaccato a questi simboli, resta il fatto che non c'è alcuna comprensione. La forma dell'argomentazione di Block sull'intuizione è che, poiché presumibilmente esistono dati empirici che mostrano che il pensare è appunto manipolazione simbolica formale, noi non potremmo confutare la tesi con intuizioni scarsamente elaborate. Si potrebbe allo stesso modo tentare di confutare l'opinione che la Terra è rotonda facendo ricorso alla nostra intuizione che è piatta. Ora Block ammette che non è un fatto di intuizione, ma un semplice fatto che i nostri cervelli sono la "sede" della nostra intenzionalità. Voglio aggiungere che è ugualmente un semplice fatto che io non capisco il cinese. Il mio articolo è un tentativo di esplorare le conseguenze logiche di questi e altri semplici fatti. Le intuizioni, nel senso deprecatorio che egli attribuisce loro, non hanno nulla a che fare con l'argomento. Una conseguenza è che le manipolazioni di simboli formali non potrebbero essere costitutive del pensare. Block non affronta mai gli argomenti per questa conseguenza. Egli semplicemente lamenta la debolezza delle nostre intuizioni. Dennett pensa che può generare controintuizioni. Suppone, nella "replica del robot", che il robot sia il mio proprio corpo. E allora? Non capirei io il cinese? Ebbene, il guaio è che il caso, come egli ce lo dà, è sottodescritto, perché non ci viene mai detto che cosa accade nella mente dell'agente (ricordate che, in queste

discussioni, dovete insistere sempre sul punto di vista della prima persona. Il primo passo alla destrezza di mano dell'operazionalista avviene quando cerchiamo di rappresentarci come potremmo sapere a cosa somiglierebbe per altri). Se descriviamo il caso di Dennett abbastanza esplicitamente, non è difficile vedere quali sarebbero gli effetti. Supponete che il programma contenga istruzioni come le seguenti: quando qualcuno mostra un segnale come *squiggle squiggle*, passagli il sale. Con tali istruzioni non ci vorrebbe molto perché uno capisca che *squiggle squiggle* probabilmente significa passare il sale. Ma ora l'agente comincia a imparare il cinese seguendo il programma. Tuttavia questa "intuizione" non urta contro i fatti che appunto indicavo, perché quello che l'agente fa in tal caso è attaccare un contenuto semantico a un simbolo formale e far così un passo verso la comprensione del linguaggio. Sarebbe ugualmente possibile descrivere un caso in modo tale che sarebbe impossibile attaccare qualche contenuto semantico, anche se è in questione il mio corpo, e in tal caso sarebbe impossibile per me imparare il cinese seguendo il programma. Gli esempi di Dennett non generano controindicazioni: semplicemente questi sono descritti così inadeguatamente che non possiamo dire, dalla sua descrizione, quali sarebbero stati i fatti.

In un punto Dennett e io abbiamo intuizioni contrarie. Egli dice: "Io capisco l'inglese, il mio cervello no". Io penso al contrario che, quando capisco l'inglese, è il mio cervello che fa quel lavoro. Non trovo assolutamente nulla di strano nel dire che il mio cervello capisce l'inglese, o anche nel dire che il mio cervello è consapevole. Trovo la sua dichiarazione altrettanto non plausibile quanto insistente. "Io digerisco la pizza; il mio stomaco e l'apparato digestivo no". Marshall sostiene che la tesi che i termostati non hanno opinioni è tanto confutabile quanto la tesi che le tavole sono solide. Ma notate la differenza. Nel caso delle tavole abbiamo scoperto fatti precedentemente sconosciuti sulla microstruttura di oggetti manifestamente solidi. Nel caso dei termostati i fatti importanti sono tutti già ben noti. Naturalmente fatti come quello che i termostati non hanno opinioni e che io non parlo cinese sono, come tutti i fatti empirici, soggetti a smentita. Potremmo, per esempio, scoprire che, contrariamente alle mie opinioni più profonde, io sono un competente parlatore di cinese colto. Ma pensate come stabiliremmo tale cosa. Al minimo, dovremmo stabilire che, del tutto inconsapevolmente, io conosco i significati di un grande numero di espressioni cinesi, e per stabilire che i termostati hanno opinioni, esattamente nello stesso senso in cui lo faccio io, dovremmo stabilire, per esempio, che, per qualche miracolo, i termostati hanno sistemi nervosi capaci di sostenere stati mentali, e così via. Insomma, sebbene in qualche senso l'intuizione figuri in ogni argomento, traviserete interamente la natura della presente discussione se pensate che sia una faccenda di intuizioni mie contro quelle di qualcun altro, o che qualche insieme di intuizioni contrarie abbia pari validità. La

dichiarazione che io non parlo cinese e che i miei termostati sono privi di opinioni, non sono cose che in qualche modo mi sento misteriosamente propenso a dire.

Infine, in risposta a Dennett (e anche Pylyshyn), io non penso naturalmente che l'intenzionalità sia un fluido. Nulla di quello che dico mi fa arrivare a quella opinione. Io penso, al contrario, che gli stati intenzionali, i processi e gli eventi, sono precisamente questo: stati, processi ed eventi.

Il punto è che essi sono prodotti e realizzati nella struttura del cervello. Dennett mi assicura che una tale opinione va contro "le correnti prevalenti della dottrina". Tanto peggio per le correnti prevalenti.

Intenzionalità intrinseca e attribuzioni di intenzionalità relative all'osservatore

Perché allora la gente si sente portata a dire che, in un certo senso almeno, i termostati hanno opinioni? Io penso che, al fine di capire che cosa accade quando si fanno tali dichiarazioni, è necessario che distinguiamo attentamente tra il caso di quella che io chiamerò intenzionalità intrinseca, che è il caso degli stati mentali effettivi, e quelle che io chiamerò attribuzioni di intenzionalità relative all'osservatore, che sono un modo che la gente ha di parlare intorno a quelle entità che figurano nelle nostre attività, ma sono prive di intenzionalità intrinseca. Possiamo illustrare questa distinzione con esempi che sono del tutto indiscussi. Se dico che ho fame o che Carter crede che può vincere le elezioni, la forma di intenzionalità in questione è intrinseca. Io sto discutendo, in modo vero o falso, di certi fatti psicologici su di me e Carter. Ma se dico che la parola "Carter" si riferisce al presidente, o che la frase "es regnet" (= "piove" in tedesco) significa che piove, io non sto attribuendo stati mentali alla parola "Carter" o alla frase "es regnet". Queste sono attribuzioni di intenzionalità fatte a entità che mancano di stati mentali, ma nei quali l'attribuzione è una maniera di parlare intorno all'intenzionalità degli osservatori. Si dice comunemente che la gente usa il nome Carter per riferirsi a una persona specifica, o che quando la gente dice letteralmente "es regnet", intende dire che sta piovendo realmente.

Le attribuzioni di intenzionalità relative all'osservatore sono sempre dipendenti dalla intenzionalità intrinseca degli osservatori. Non ci sono due tipi di stati mentali intenzionali; c'è solo un tipo in cui essi hanno intenzionalità intrinseca; ma ci sono attribuzioni di intenzionalità in cui l'attribuzione non attribuisce intenzionalità intrinseca al soggetto dell'attribuzione. Ora io credo che molto dell'attuale disputa si basi sull'errore commesso nell'operare questa distinzione. Quando McCarthy vigorosamente sostiene che i termostati hanno opinioni, egli confonde

attribuzioni di intenzionalità relative all'osservatore con attribuzioni di intenzionalità intrinseca. Per considerare questo punto, chiedetevi perché facciamo queste attribuzioni a termostati e simili. Non è perché supponiamo che abbiano una vita mentale in tutto simile alla nostra; al contrario, sappiamo che non hanno vita mentale per niente. È piuttosto perché li abbiamo designati (la nostra intenzionalità) a servire certi nostri scopi (più della nostra intenzionalità), a eseguire il genere di funzioni che noi eseguiamo sulla base della nostra intenzionalità. Io credo che sia ugualmente chiaro che la nostra attribuzione di intenzionalità avviene anche per le automobili, i computers e le macchine calcolatrici ed è relativa all'osservatore. Il funzionalismo, a proposito, è un intero sistema basato sulla mancanza di questa distinzione. Le attribuzioni funzionali sono sempre relative all'osservatore. Non c'è alcuna funzione intrinseca simile, come non ci sono stati intenzionali intrinseci.

Tipi naturali

Questa distinzione tra intenzionalità intrinseca e attribuzioni di intenzionalità relative all'osservatore potrebbe sembrare meno importante se potessimo, come parecchi commentatori propongono (Minsky, Block, Marshall) assimilare l'intenzionalità intrinseca a qualche più ampio tipo naturale che sussumerebbe sia i fenomeni mentali esistenti, sia altri fenomeni naturali sotto un apparato esplicativo più generale. Minsky dice che "germi di idee prescientifiche come 'credere'" non hanno posto nella scienza della mente del futuro (presumibilmente anche "mente" non avrà posto nella "scienza della mente" del futuro). Ma seppure questo è vero, ciò è assolutamente irrilevante rispetto al mio argomento, che è indirizzato alla scienza della mente del presente. Anche se, come Minsky sostiene, noi alla fine veniamo a parlare delle nostre opinioni presenti come se fossero in un *continuum* con cose che non sono affatto stati intenzionali, ciò non altera il fatto che noi abbiamo opinioni intrinseche e che i computers e i termostati non le hanno. Cioè, anche se una certa scienza futura raggiunge un livello che smentisce l'opinione, e così rende possibile porre i termostati e la gente in un unico *continuum*, ciò non altererebbe il fatto che sotto il nostro concetto attuale di opinione, la gente ha di fatto opinioni, mentre i termostati non le hanno. Né confuterebbe la mia diagnosi di tale errore l'attribuire stati mentali intrinseci ai termostati in quanto tale attribuzione sarebbe basata su una confusione tra intenzionalità intrinseca e attribuzione di intenzionalità relativa all'osservatore.

Minsky sottolinea inoltre che le nostre operazioni mentali sono spesso divise in parti che non sono pienamente integrate da alcun "sé" e di cui solo alcune eseguono l'interpretazione. Ed egli chiede, se questo è il modo in cui

ciò accade nelle nostre menti, perché non anche nei computers? La risposta è che anche se ci sono parti dei nostri processi mentali dove il ragionamento ha luogo senza alcun contenuto intenzionale, ci devono ancora essere altre parti che attribuiscono un contenuto semantico a elementi sintattici, se ci deve essere davvero comprensione. La base dell'esempio della stanza cinese è che le manipolazioni dei simboli formali non portano mai da sé alcun contenuto semantico, e così l'instanziare un programma di computer non è in sé e per sé sufficiente per comprendere.

Come funziona il cervello

Parecchi commentatori mi rimproverano perché non spiego come funziona il cervello per produrre intenzionalità, e almeno due (Dennett e Fodor), obiettano alla mia tesi che dove si tratta di intenzionalità — rispetto alle condizioni di soddisfazione dell'intenzionalità — quello che importa sono le cause interne e non le esterne. Ebbene, io non so come il cervello produce i fenomeni mentali, ed evidentemente nessun altro lo sa, ma che esso produce i fenomeni mentali e che le operazioni interne del cervello sono causalmente sufficienti per i fenomeni, questo è abbastanza evidente da quello che sappiamo. Consideriamo il caso seguente, in cui sappiamo qualcosa su come il cervello funziona. Da dove sono seduto, posso vedere un albero. La luce riflessa dall'albero in forma di fotoni colpisce il mio apparato ottico. Ciò innesca una serie di sequenze di scoppi neurali. Alcuni di questi neuroni nella corteccia visiva sono in effetti notevolmente specializzati per rispondere a certi generi di stimoli visivi. Quando l'intera serie di sequenze ha luogo, essa causa un'esperienza visiva, e l'esperienza visiva ha intenzionalità. È un evento mentale conscio con un contenuto intenzionale; cioè, le sue condizioni di soddisfazione sono interne a esso. Ora io potrei avere esattamente quell'esperienza visiva anche se non ci fosse alcun albero, purché soltanto succedesse qualcosa nel mio cervello che fosse sufficiente a produrre l'esperienza. In tale caso non vedrei l'albero, ma avrei un'allucinazione. In tal caso, perciò, l'intenzionalità è una questione di cause interne; se l'intenzionalità è soddisfatta, cioè se effettivamente vedo un albero invece di avere una allucinazione dell'albero, è di nuovo questione di cause esterne. Se io fossi un cervello in una vasca potrei avere esattamente gli stessi stati mentali che ho ora; solo che la maggior parte di essi sarebbe falsa o altrimenti insoddisfatta. Ora questo semplice esempio di esperienza visiva è designato per chiarire che cosa ho in mente quando dico che l'operazione del cervello è causalmente sufficiente per l'intenzionalità, e che è l'operazione del cervello e non l'impatto del mondo esterno che importa per il contenuto dei nostri stati intenzionali, almeno nel senso di "contenuto".

Alcuni dei commentatori sembra che suppongano che io consideri i poteri causali del cervello come un argomento contro l'Intelligenza Artificiale forte. Ma questo è un malinteso. È una questione puramente empirica se una qualunque data macchina ha poteri causali equivalenti al cervello. La mia tesi contro l'Intelligenza Artificiale forte è che instanziare un programma non è sufficiente a garantire che esso abbia quei poteri causali.

Attendi fino al prossimo anno

Molti autori (Block, Sloman e Croucher, Dennett, Lycan, Bridgeman, Schank) dichiarano che il programma di Schank non è abbastanza buono, ma che programmi migliori sconfiggeranno la mia obiezione. Io penso che in questa idea manchi il punto essenziale dell'obiezione. La mia obiezione vale contro ogni tipo di programma, in quanto programma di computer. Né giova alla disputa aggiungere la teoria causale della referenza, perché anche se i simboli formali nel programma hanno qualche connessione causale rispetto ai loro presunti referenti nel modo reale, finché l'agente non ha alcun modo di conoscere ciò, non aggiunge alcun tipo di intenzionalità ai simboli formali. Supponi, per esempio, che il simbolo per "uovo-foo-yung" nella stanza cinese sia effettivamente connesso causalmente a "uovo-foo-yung". Ancora, l'uomo nella stanza non ha alcun modo di conoscere ciò. Per lui questo rimane un simbolo formale non interpretato, con nessun contenuto semantico. Tornerò su quest'ultimo punto nella discussione di autori specifici, specialmente Fodor.

Seriatim

Ora mi volgo, con le solite scuse per la brevità, da queste considerazioni più generali a una serie di argomenti specifici.

Haugeland presenta un argomento che è genuinamente originale. Supponete che un cinese madrelingua abbia i neuroni rivestiti di un sottile strato che impedisce l'esplosione dei neuroni. Supponi che il "demone di Searle" ovvii a tale mancanza stimolando i neuroni come se fossero stati colpiti. Allora capirà il cinese anche se nessuno dei suoi neuroni ha i giusti poteri causali; il demone li ha, e lui comprende solo l'inglese. La mia obiezione è rivolta solo all'ultima frase. I neuroni del parlante hanno ancora i giusti poteri causali; hanno solo bisogno di un po' di aiuto da parte del demone. Più generalmente, se la stimolazione delle cause è a un livello abbastanza basso da riprodurre le cause e non semplicemente descriverle, la "simulazione" riprodurrà gli effetti. Se quello che il demone fa è riprodurre i giusti fenomeni causali, esso avrà

riprodotto l'intenzionalità che costituisce gli effetti di quel fenomeno. E non si dimostra, per esempio, che il mio cervello manca della capacità della consapevolezza se qualcuno deve svegliarmi al mattino massaggiandomi la testa.

La distinzione di Haugeland tra intenzionalità originale e derivata è in qualche modo come la mia distinzione tra l'intenzionalità intrinseca e le attribuzioni di intenzionalità relative all'osservatore. Ma egli è in errore se pensa che la sola distinzione è che l'intenzionalità originale è "sufficientemente ricca" nella sua "attività semantica"; l'attività semantica in questione è ancora relativa all'osservatore e quindi non sufficiente per l'intenzionalità. Il motore della mia macchina è, nel suo senso relativo all'osservatore, semanticamente attivo in tutti i modi, ma non ha intenzionalità. Un infante umano è semanticamente abbastanza inattivo, tuttavia ha intenzionalità. Rorty espone un argomento riguardante la transustanziazione che è formalmente parallelo al mio riguardante le attribuzioni di intenzionalità intrinseche e relative all'osservatore. Poiché le premesse dell'argomento della transustanziazione sono presunte false, il parallelo si suppone sia una obiezione alla mia tesi. Ma il parallelo è totalmente irrilevante. Qualunque valido argomento, da premesse vere a conclusioni vere, ha analoghi formali esatti da false premesse a false conclusioni. Parallelo al noto argomento che "Socrate è mortale" abbiamo che "Socrate è un cane". "Tutti i cani hanno tre teste", perciò "Socrate ha tre teste". La possibilità di tali paralleli formali non fa nulla per indebolire gli argomenti originali. Per mostrare che il parallelo era significativo, Rorty avrebbe dovuto mostrare che le mie premesse sono tanto infondate scientificamente quanto la dottrina della transustanziazione. Ma quali sono le mie premesse? Sono che le persone hanno stati mentali come opinioni, desideri ed esperienze visive, che hanno pure cervelli e che i loro stati mentali sono causalmente i prodotti dell'operazione dei loro cervelli. Rorty non dice nulla in ogni caso per mostrare che queste proposizioni sono false e io francamente non posso supporre che egli dubiti della loro verità. Auspicherebbe egli forse una prova? Egli conclude lamentando che se la mia opinione guadagna favore, il "buon lavoro" dei suoi favoriti autori comportamentisti e funzionalisti sarà "demolito". Questa non è una prospettiva che io trovi del tutto sviante poiché è implicita nel mio intero argomento l'opinione che le persone realmente hanno stati mentali e dire questo non è attribuire loro tendenze di comportamento o adottare un certo tipo di posizione nei loro confronti o suggerire spiegazioni funzionali dei loro comportamenti. Questo non dà al mentale un "luminoso splendore cartesiano", solo implica che i processi mentali sono reali quanto ogni altro processo biologico.

McCarthy e Wilensky entrambi si attribuiscono la "replica del sistema". La

più importante aggiunta fatta da Wilensky è di supporre che noi chiediamo al sottosistema cinese se parla cinese ed esso risponde sì. Ho già sostenuto che questo non aggiunge alcuna plausibilità in ogni modo alla tesi che avvenga qualche comprensione del cinese all'interno del sistema. Sia Wilensky che McCarthy non rispondono alle tre obiezioni che ho fatto alla replica dei sistemi.

1. Il sottosistema cinese non attacca alcun contenuto semantico ai simboli formali. Il sottosistema inglese sa che "hamburger" significa hamburger. Il sottosistema cinese sa solo che *squiggle squiggle* è seguito da *squoggle squoggle*.

2. La replica dei sistemi è totalmente immotivata. La sua sola motivazione è il test di Turing e il fare ricorso a quello significa precisamente fare la domanda considerando come vero ciò che è ancora in discussione.

3. La replica dei sistemi ha la conseguenza che ogni genere di relazioni input-output sistematiche (per esempio, la digestione) dovrebbe contare come comprensione, poiché garantisce tanta intenzionalità relativa all'osservatore quanta ne garantisce il sottosistema cinese (e non è, a proposito, una risposta a questo punto il fare ricorso all'impenetrabilità cognitiva della digestione, nel senso di Pylyshyn (1980a), poiché la digestione è cognitivamente penetrabile: il contenuto delle mie credenze può sconvolgere la mia digestione). Sembra che Wilensky obietti che altri generi di stati mentali oltre a quelli intenzionali potrebbero costituire il soggetto della disputa. Ma io condivido in pieno che avrei potuto fare una discussione su dolori, solletico e ansietà, ma questi sono: a) meno interessanti per me; b) meno discussi nella letteratura di Intelligenza Artificiale. Preferisco attaccare l'IA forte su quello che i suoi proponenti considerano sia il loro più forte principio fondante. Pylyshyn fraintende il mio punto. Io non offro alcuna prova a priori che un sistema a circuito integrato non potrebbe avere intenzionalità. Quella è, come dico ripetutamente, una questione empirica. Quello che sostengo è che, al fine di produrre intenzionalità, il sistema dovrebbe duplicare i poteri causali del cervello e che instanziare semplicemente un programma formale non sarebbe sufficiente per questo. Pylyshyn non offre risposta alle prove che io dò per queste conclusioni.

Poiché Pylyshyn non è il solo che ha questo fraintendimento, vale forse la pena porre l'accento solo su ciò che è pericoloso. La posizione dell'IA forte è che qualsiasi cosa, con il giusto programma, dovrebbe avere l'intenzionalità rilevante. Il circuito, nel suo esempio, avrebbe necessariamente intenzionalità e non importerebbe se fossero circuiti o condutture dell'acqua o ritagli di carta, purché instanziassero il programma. Ora io sostengo che non potrebbero avere intenzionalità solamente in grazia del fatto che instanziano il programma. Una volta che si veda che il programma non aggiunge necessariamente intenzionalità a un sistema, allora diventa una questione

empirica quali generi di sistemi abbiano realmente intenzionalità, e la condizione necessaria per questo è che devono avere poteri causali equivalenti a quelli del cervello. Penso che sia evidente che tutti i tipi di sostanze nel mondo, come condutture dell'acqua e carta, vanno a mancare di quei poteri, ma questa è una dichiarazione empirica da parte mia. Dal mio punto di vista è una dichiarazione empirica testimoniabile se nel riparare un cervello danneggiato potessimo duplicare la base elettrochimica dell'intenzionalità usando qualche altra sostanza, per esempio il silicio. Sulla posizione dell'IA forte non ci possono essere questioni empiriche sulle basi elettrochimiche necessarie per l'intenzionalità, poiché qualsiasi sostanza è sufficiente per l'intenzionalità se ha il giusto programma.

Io credo che Pylyshyn fraintenda anche la distinzione fra intenzionalità intrinseche e attribuzioni di intenzionalità relative all'osservatore. La questione rilevante non è quanto spazio ha l'osservatore nel fare attribuzioni relative all'osservatore, ma se c'è qualche intenzionalità intrinseca nel sistema al quale le attribuzioni possano corrispondere.

Sembrerebbe che Schank e io siamo in accordo su molti punti, ma c'è almeno un piccolo fraintendimento. Egli pensa che io voglia mettere in dubbio l'impresa dell'"Intelligenza Artificiale". Non è vero. Io sono tutto a favore dell'IA debole, almeno come programma di ricerca. Sono completamente d'accordo che se qualcuno sapesse scrivere un programma che desse il giusto input o output per storie cinesi, sarebbe un "grande risultato" che richiede una notevole comprensione della natura del linguaggio. Non sono nemmeno sicuro che possa essere fatto. La mia opinione è che instanziare il programma non è costitutivo del comprendere.

Abelson come Schank sottolinea che non è il risultato principale programmare computers che sanno simulare la comprensione di una storia. Ma, per ripetermi, questo è un risultato di quella che io chiamo IA debole, e io applaudirei entusiasticamente. Egli critica questo valido argomento insistendo che, dal momento che il nostro comprendere la maggior parte delle cose, l'aritmetica, per esempio, è molto imperfetto, "potremmo ben essere umili e dare al computer il beneficio del dubbio quando e se esegue altrettanto bene quanto noi". Temo che né questo, né le sue altre opinioni si accordino con i miei argomenti per mostrare che, per quanto umili possiamo essere, non c'è ragione di supporre che instanziare un programma formale nel modo in cui lo fa un computer non è assolutamente un motivo per attribuirgli intenzionalità.

Fodor è d'accordo con la mia tesi centrale che instanziare un programma non è una condizione sufficiente d'intenzionalità. Egli pensa comunque che se noi avessimo capito i corretti nessi causali tra i simboli formali e gli oggetti nel mondo, questo sarebbe sufficiente. Ora c'è un'obiezione ovvia a questa variante della "replica del robot" che ho fatto molte volte: lo stesso

esperimento di pensiero di prima si applica a questo caso. Cioè, qualunque impatto causale esterno ci sia sui simboli formali, questo non è sufficiente di per sé a dare ai simboli alcun contenuto intenzionale. Qualunque cosa può avere causato i simboli, ma l'agente ancora non capisce il cinese. Lascia che il simbolo "uovo-foo-yung" sia causalmente connesso a "uovo-foo-yung" in qualunque modo ti piaccia; quella connessione di per sé non metterà mai l'agente in grado di interpretare il simbolo come significante "uovo-foo-yung". Per fare questo dovrebbe avere, per esempio, qualche consapevolezza della relazione causale tra il simbolo e il referente; ma ora noi non stiamo più spiegando l'intenzionalità in termini di simboli e cause, ma in termini di simboli, cause e intenzionalità, e abbiamo abbandonato sia l'IA forte che la replica del robot. La sola risposta di Fodor è dire che ciò dimostra che non abbiamo ancora il giusto genere di connessione causale. Ma che cos'è il giusto genere, poiché l'argomento di cui sopra si applica a ogni genere? Afferma che non ce lo può dire, ma esiste tuttavia. Ebbene, io posso dirgli che cosa è: è ogni tipo di causazione sufficiente a produrre contenuto intenzionale nell'agente, sufficiente a produrre, per esempio, un'esperienza visiva, o un ricordo, o un credo o un'interpretazione semantica di qualche parola.

La variante di Fodor della replica del robot si trova perciò di fronte a un dilemma; se le connessioni causali sono appunto dati di fatto intorno alle relazioni tra i simboli e il mondo esterno, esse non daranno mai da sé alcuna interpretazione ai simboli; esse non porteranno da sé alcun contenuto intenzionale. Se, d'altro lato, l'impatto causale è sufficiente a produrre intenzionalità nell'agente, può essere solo perché c'è qualcosa di più per il sistema del semplice fatto dell'impatto causale e del simbolo, precisamente il contenuto intenzionale che l'impatto produce nell'agente. O l'uomo nella stanza non impara il significato del simbolo dall'impatto causale, nel qual caso l'impatto causale non aggiunge nulla all'interpretazione, o l'impatto causale gli insegna il significato della parola, nel qual caso la causa è rilevante solo perché produce una forma di intenzionalità che è qualcosa in aggiunta a sé e al simbolo. In nessuno dei casi il simbolo, o causa e simbolo, è costitutivo di intenzionalità.

Non è questo il luogo per discutere il ruolo generale dei processi formali nei processi mentali, ma non posso fare a meno di richiamare l'attenzione a una massiccia confusione implicita nella tesi di Fodor. Dal fatto che, per esempio, le regole sintattiche riguardano oggetti formali, non consegue che siano regole formali. Come altre regole che toccano il comportamento umano, esse sono definite dal loro contenuto, non dalla loro forma. Appunto così accade che in questo caso il loro contenuto riguarda le forme.

In quello che è forse il suo punto cruciale, Fodor suggerisce che dovremmo pensare al cervello o al computer come eseguenti operazioni formali su simboli interpretati e non solo su simboli formali. Ma chi opera

l’interpretazione? E che cosa è un’interpretazione? Se egli intende che per intenzionalità ci deve essere contenuto intenzionale in aggiunta ai simboli formali, allora io, naturalmente, sono d’accordo. Infatti due dei punti principali della mia tesi sono che nel nostro caso abbiamo l’“interpretazione”, cioè, abbiamo intenzionalità intrinseca, e che il programma del computer non potrebbe mai essere sufficiente di per sé a fare ciò.

Nel caso del computer noi facciamo attribuzioni di intenzionalità relative all’osservatore, ma ciò non dovrebbe essere scambiato per la cosa reale poiché il programma del computer non ha in sé alcuna intenzionalità intrinseca.

Sloman e Croucher sostengono che il problema nel mio esperimento di pensiero è che il sistema non è abbastanza capace. A colui che comprende la storia di Schank, essi aggiungerebbero ogni tipo di altre operazioni, ma essi sottolineano che queste operazioni sono computazionali e non fisiche. L’ovvia obiezione alla loro proposta è quella che essi anticipano: io posso ancora ripetere il mio esperimento di pensiero con il loro sistema per quanto capace esso sia. A ciò essi rispondono che io sostengo senza prova che “è impossibile per un’altra mente essere basata sul suo (mio) processo mentale senza il suo (mio) conoscere”. Ma non è quello che credo io. Per quello che so io, ciò può esser falso. Piuttosto, quello che credo è che non puoi capire il cinese se non conosci i significati di ognuna delle parole in cinese. Più in generale, a meno che un sistema possa attaccare contenuto semantico a una serie di elementi sintattici, l’introduzione degli elementi nel sistema non aggiunge alcuna intenzionalità. Questo vale per me e per tutti i piccoli sottosistemi che vengono organizzati all’interno di me stesso.

Eccles sottolinea correttamente che io non mi occupo mai di confutare la posizione dell’interazione dualista tenuta da lui e da Popper. Al contrario, io controbatto l’IA forte sulla base di quello che si potrebbe chiamare una posizione interazionista monista. La mia sola scusa per non attaccare a spada tratta la sua forma di dualismo è che questo articolo aveva veramente altri obiettivi. Io sono direttamente interessato all’IA forte e solo incidentalmente al “problema mente-cervello”. Egli ha ben ragione a pensare che i miei argomenti contro l’IA forte non sono di per sé contrari alla sua versione di interazionismo dualista, e sono lusingato di vedere che condividiamo l’opinione che “è proprio ora che l’IA forte sia screditata”.

Temo di non avere nulla di originale da dire sulla risposta comportamentista di Rachlin, e se la discutessi farei solo le solite obiezioni al comportamentismo estremo. In particolare ho ancora maggiori difficoltà col comportamentismo e il funzionalismo, perché non riesco a immaginare che qualcuno veramente creda a queste idee. So che certe persone dicono di crederci, ma che cosa devo pensare quando Rachlin dice che “non ci sono stati mentali sottostanti (…) il comportamento” e “il modello del

comportamento è lo stato mentale”? Non ci sono dolori sotto il comportamento di dolore di Rachlin? Per quanto riguarda me, devo confessare che ci sono purtroppo spesso dolori sotto il mio comportamento di dolore, e perciò concludo che la forma di comportamento di Rachlin non è vera in generale.

Lycan ci dice che i miei controesempi non sono tali per una teoria funzionalista della comprensione del linguaggio, perché l’uomo del mio controesempio userebbe i programmi sbagliati. Allora diteci quali sono i programmi giusti e noi programmeremo l’uomo con quei programmi e ancora produrremo un controesempio. Egli ci dice pure che le corrette connessioni causali determinano il contenuto appropriato da attaccare ai simboli formali. Io credo che la mia risposta a Fodor e ad altre versioni della risposta causale o del robot sia altrettanto appropriata rispetto ai suoi argomenti, per cui non la ripeterò.

Hofstadter in modo simpatico definisce il mio articolo come “uno degli articoli più errati e indisponenti che io abbia mai letto nella mia vita”. Credo che sarebbe stato meno (o forse più?) contrariato se si fosse preso il disturbo di leggere l’articolo in maniera veramente accurata. Sembra che la sua strategia generale sia che ogni volta che io affermo *p,* egli dice che non affermo *p.* Per esempio, io respingo il dualismo, ed egli dice che io credo nell’anima. Io penso che sia un semplice fatto di natura che i fenomeni mentali siano causati da fenomeni neurofisiologici, ed egli dice che io ho “grave difficoltà” ad accettare un’idea di tal genere. L’intero tono del mio articolo è quello di trattare la mente come una parte del mondo (fisico), come qualunque altra cosa, ed egli dice che io ho “un orrore istintivo” per ogni riduzionismo. Egli travisa le mie idee quasi a ogni punto e di conseguenza io trovo difficile prendere sul serio i commenti. Se il mio testo è troppo difficile, consiglio a Hofstadter di leggere Eccles, che recepisce correttamente il mio rifiuto del dualismo.

Inoltre, il commento di Hofstadter contiene il seguente *non sequitur*: dal fatto che l’intenzionalità “scaturisce” dal cervello e dalla premessa che “i processi fisici sono formali”, cioè “governati da regole”, egli deduce che i processi formali sono costitutivi del mentale, che noi siamo “sistemi formali di base”. Ma quella conclusione semplicemente non segue dalle due premesse. Non segue nemmeno data la sua strana interpretazione della seconda premessa: “A metterla in un altro modo, la premessa è che non c’è intenzionalità al livello delle particelle”. Posso accettare tutte queste premesse, ma esse semplicemente non comportano la conclusione. Esse comportano che l’intenzionalità è un “risultato di processi formali” nel banale senso che è un risultato di processi che hanno un livello di descrizione per il quale essi sono l’instanziazione di un programma di computer, ma la stessa cosa è vera per il latte e lo zucchero e infiniti altri “risultati di processi

formali”.

Hofstadter ipotizza pure che forse alcuni trilioni di tubi per l’acqua possano lavorare a produrre consapevolezza, ma poi evita di trattare direttamente l’elemento cruciale della mia tesi, che è che, anche se fosse questo il caso, dovrebbe esserlo perché il sistema dei tubi era la copia dei poteri causali del cervello e non solo perché instanziava un programma formale.

Ritengo di essere d’accordo con gli acuti commenti di Smythe eccetto forse in un punto. Sembra che egli supponga che se il programma è instanziato con “operazioni di hardware primitivo” le mie obiezioni non si applicherebbero. Ma perché? Lasciamo che l’uomo del mio esempio abbia il programma già delineato nel suo hardware. Con ciò egli non comprende ancora il cinese. Supponiamo che egli abbia una tale carica che automaticamente esca con frasi cinesi non interpretate in risposta a stimoli cinesi non interpretati. È ancora lo stesso caso tranne per il fatto che non agisce più volontariamente.

Osservazioni marginali. Mi è parso che alcuni dei commentatori non hanno colto il vero problema o si sono concentrati su argomenti periferici, per cui le mie annotazioni su di essi saranno anche più brevi. Penso che Bridgeman abbia mancato il vero punto della mia tesi quando dichiara che sebbene l’homunculus nel mio esempio potrebbe non sapere ciò che sta accadendo, potrebbe subito impararlo, e che ciò richiederebbe semplicemente più informazioni, specificamente “informazioni con una relazione sconosciuta e nota al mondo esterno”. Sono d’accordo. Nella misura in cui l’homunculus ha tale informazione esso è più di una semplice instanziazione di un programma di computer, ed è così irrilevante ai fini della mia disputa nei confronti dell’IA forte. Secondo quest’ultima, se l’homunculus ha il giusto programma deve già avere l’informazione. Ma dissento dalla dichiarazione di Bridgeman per cui le sole proprietà del cervello sono le proprietà che esso ha a livello dei neuroni.

Io penso che tutti i partecipanti alla disputa sarebbero d’accordo sul fatto che il cervello ha ogni genere di proprietà che non sono attribuibili al livello dei neuroni individuali — per esempio, proprietà causali (come il controllo del *cervello* sul respiro).

Simili fraintendimenti si applicano alle osservazioni di Marshall. Egli denuncia pesantemente l’idea che c’è qualcosa di debole nelle grandi realizzazioni dell’IA debole, e conclude: “Chiaramente ci deve essere qualche fraintendimento radicale”. Il solo fraintendimento era però nel suo supporre che nel porre a fronte l’IA debole con quella forte, io in qualche modo mettevo in svantaggio la prima.

Marshall trova strano che qualcuno debba pensare che un programma possa essere una teoria. Ma la parola “programma” è usata ambiguamente. Qualche volta “programma” si riferisce al mucchio di schede perforate, qualche volta a una serie di affermazioni. È nel secondo senso che si suppone che qualche

volta i programmi siano teorie. Se Marshall ha obiezioni a quel senso, la disputa è ancora puramente verbale e può essere risolta col dire non che il programma è una teoria, ma che il programma è una parte di una teoria. E l'idea che i programmi possano essere teorie non è qualcosa che ho inventato io. Considerate il seguente: "Occasionalmente, dopo aver visto quello che il programma può fare, qualcuno chiederà una precisazione della teoria che sta dietro esso. Spesso la risposta corretta è che il programma è la teoria" (Winston 1977, p. 259).

Anche Ringle non ha capito il mio argomento principale. Dice che mi rifugio nel misticismo sostenendo che "le proprietà fisiche dei sistemi neuronali sono tali che non possono per principio essere simulate da un computer nonprotoplasmico". Ma questo non è nemmeno remotamente il senso della mia tesi. Io penso che di qualunque cosa possa essere data una simulazione formale, ed è una questione empirica in ogni caso se la simulazione sia stata la copia dei caratteri causali. La questione è se la simulazione formale di per sé, senza ulteriori elementi causali, è sufficiente a riprodurre il mentale. E la risposta a quella domanda è no, a motivo degli argomenti che ho affermato ripetutamente, e a cui Ringle non risponde. È proprio illogico supporre che, poiché il cervello ha un programma e poiché il computer potrebbe avere lo stesso programma, quello che fa il cervello non è niente di più di quello che fa il computer. È in ogni caso una questione empirica se un sistema in competizione duplica poteri causali del cervello, ma è una questione del tutto diversa da quella se instanziare un programma formale è di per sé costitutivo del mentale.

Ho anche la sensazione, forse basata su un malinteso, che la discussione di Menzel sia basata su una confusione tra *come* si sa che qualche sistema ha stati mentali e *che cosa* è avere uno stato mentale. Egli ritiene che io cerchi un criterio per il mentale, e non riesce a vedere il nocciolo della questione quando dico cose così vaghe sul cervello. Ma io non cerco affatto un criterio per il mentale. So che cosa sono gli stati mentali, almeno in parte, essendo io stesso un sistema di stati mentali. La mia obiezione all'IA forte non è, come Menzel proclama, che potrebbe fallire in un singolo possibile caso, ma piuttosto che, nel caso in cui esso fallisce, non possiede più risorse; per cui, se fallisce in quel caso fallisce in ogni caso. Nell'articolo di Walter non mi riesce di scoprire alcun argomento, solo alcune deboli analogie. Lamenta il mio fallimento nel rendere più esplicite le mie idee sull'intenzionalità. Esse sono tali nei tre articoli citati da Natsoulas (Searle 1979a; 1979b; 1979c).

Ulteriori implicazioni

Posso solo esprimere il mio apprezzamento per i contributi di Danto, Libet,

Maxwell. In vari modi, ognuno di loro aggiunge argomenti di sostegno e commenti alla tesi principale. Sia Natsoulas che Maxwell mi stimolano a fornire risposte a domande sull'importanza della discussione delle problematiche ontologiche tradizionali e della questione mente-corpo. Io cerco di evitare per quanto possibile il vocabolario e le categorie tradizionali, e questo è il mio quadro molto approssimativo.

Gli stati mentali sono reali quanto ogni altro fenomeno biologico. Essi sono determinati e realizzati nel cervello. Ciò non è più misterioso del fatto che proprietà come l'elasticità e la resistenza alla puntura di una gomma gonfiata di macchina sono entrambe causate e realizzate nella sua microstruttura. Naturalmente, questo non implica che gli stati mentali siano attribuibili ai neuroni individuali, più di quanto le proprietà al livello della gomma siano ascrivibili agli elettroni individuali. Proseguendo l'analogia: il cervello opera causalmente sia al livello dei neuroni che al livello degli stati mentali, nello stesso senso che la gomma opera causalmente sia al livello delle particelle che al livello delle sue proprietà globali. Gli stati mentali sono non più epifenomenici di quanto lo siano l'elasticità e la resistenza alla puntura di una gomma gonfiata, e si possono descrivere interazioni sia ai livelli più alti che ai livelli più bassi, proprio come nell'analogo caso della gomma.

Alcuni stati mentali, ma non tutti, sono consapevoli, e la distinzione fra intenzionale e non-intenzionale tronca di netto la distinzione fra consapevole e inconsapevole. A ogni livello di fenomeni sono causali. Suppongo che questo sia "interazionismo", e valuto che esso sia, in qualche senso, "monismo", ma preferirei non usare comunque questo tipo di vocabolario.

Conclusione

Concludo che la stanza cinese è sopravvissuta agli assalti dei suoi critici. L'enigma che resta per me è questo: perché tanti studiosi di Intelligenza Artificiale vogliono ancora aderire all'IA forte? Sicuramente l'IA debole è più stimolante, interessante e difficile.

Bibliografia

Anderson, J. (1980) Cognitive units. Paper presented at the Society for Philosophy and Psychology, Ann Arbor, Mich.

Block, N. J. (1978) Troubles with functionalism. In: *Minnesota studies in the philosophy of science, vol. 9,* ed. C.W. Savage, Minneapolis University of Minnesota Press.

(di prossima pubblicazione) Psychologism and behaviorism. *Philosophical Review*.

Bower, G.H.; Black, J.B., & Turner, T.J. (1979) Scripts in text comprehension and memory. *Cognitive Psychology* 11: 177-220.

Carroll, C.W. (1975) *The great chess automation.* New York: Dover.

Cummins, R. (1977) Programs in the explanation of behavior. *Philosophy of Science* 44: 269-87.

Dennett D.C. (1969) *Content and consciousness.* London: Routledge & Kegan Paul.

(1971) Intentional systems. *Journal of Philosophy* 68: 87-106.

(1972) Reply to Arbib and Gunderson. Paper presented at the Eastern Division meeting of the American Philosophical Association. Boston, Mass.

(1975) Why the law of effect won't go away. *Journal for the Theory of Social Behavior* 5: 169-87.

(1978) *Brainstorms.* Montgomery, Vt.: Bradford Books.

Eccles, J.C. (1978) A critical appraisal of brain-mind theories. In: *Cerebral correlates of conscious experiences,* ed. P.A. Buser and A. Rougeul-Buser, pp. 347-55. Amsterdam: North Holland.

(1979) *The human mystery,* Heidelberg: Springer Verlag.

Fodor, J.A. (1968) The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65: 627-40. (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences* 3:1.

Freud, S. (1895) Project for a scientific psychology. In: *The standard edition of the complete psychological works of Sigmund Freud,* vol. 1, ed. J.Strachey. London: Hogarth Press, 1966.

Frey, P.W. (1977) An introduction to computer chess. In: *Chess skill in man and machine,* ed. P.W. Frey. New York, Heidelberg, Berlin: Springer-Verlag.

Fryer, D.M. & Marshall, J.C. (1979) The motives of Jacques de Vaucanson. *Technology and Culture* 20: 257-69.

Gibson, J.J. (1966) *The senses considered as perceptual systems.* Boston: Houghton Mifflin.
(1967) New reasons for realism. *Synthese* 17: 162-72.
(1972) A theory of direct visual perception. In: *The psychology of knowing,* ed. S.R. Royce & W.W. Rozeboom. New York: Gordon & Breach.

Graesser, A.C.; Gordon S.E., & Sawyer, J.D. (1979) Recognition memory for typical and atypical actions in scripted activities: tests for a script pointer and tag hypotheses. *Journal of Verbal Learning and Verbal Behavior* 1: 319-32.

Gruendel, J. (1980) Scripts and stories: a study of children's event narratives. Ph.D. dissertation, Yale University.

Hanson, N.R. (1969) *Perception and discovery.* San Francisco: Freeman, Cooper.

Hayes, P.J. (1977) In defence of logic. In: *Proceedings of the 5th international joint conference on artificial intelligence,* ed. R. Reddy. Cambridge, Mass.: The MIT Press.

Hobbes, T. (1651) *Leviathan,* London: Willis.

Hofstadter, D.R. (1979) *Gödel, Escher, Bach,* New York: Basic Books.

Householder, F.W. (1962) On the uniqueness of semantic mapping *Word* 1: 173-85.

Huxley, T.H. (1874) On the hypothesis than animals are automata and its history. In: *Collected Essays,* vol. 1, London: Macmillan, 1893.

Kolers, P.A. & Smythe, W.E. (1979) Images, symbols, and skills. *Canadian Journal of Psychology* 33: 158-84.

Kosslyn, S.M. & Shwartz, S.P. (1977) A simulation of visual imagery. *Cognitive Science* 1: 265-95.

Lenneberg, E.H. (1975) A neuropsychological comparison between man, chimpanzee and monkey. *Neuropsychologia* 13: 125.

Libet, B. (1973) Electrical stimulation of cortex in human subjects and conscious sensory aspects. In: *Handbook of sensory physiology,* vol. II, ed. A. Iggo, pp. 743-90. New York: Springer-Verlag.

Libet, B., Wright, E.W., Jr., Feinstein, B. and Pearl, D.K. (1979) Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man. *Brain* 102: 191-222.

Longuet-Higgins, H.C. (1979) The perception of music. *Proceedings of the Royal Society of London* B 205: 307-22.

Lucas, J.R. (1961) Minds, machines, and Gödel. *Philosophy* 36: 112-127.

Lycan, W.G. (di prossima pubblicazione) Form, function, and feel. *Journal of Philosophy.*

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence,* ed. M. Ringle. Atlantic Highlands, N.J.: Humanities Press.

Marr, D. & Poggio, T. (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London* B 204: 301-28.

Marshall, J.C. (1971) Can humans talk? In: *Biological and social factors in psycholinguistics,* ed. J. Morton. London: Logos Press.

(1977) Minds, machines and metaphors. *Social Studies of Science* 7: 475-88.

Maxwell, G. (1976) Scientific results and the mind-brain issue. In: *Consciousness and the brain,* ed. G.G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press.

(1978) Rigid designators and mind-brain identity. In: *Perception and cognition: Issues in the foundations of psychology,* Minnesota Studies in the Philosophy of Science, vol. 9, ed. C.W. Savage. Minneapolis: University of Minnesota Press.

Mersenne, M. (1636) *Harmonie universelle.* Paris: Le Gras.

Moor, J.H. (1978) Three myths of computer Science. *British Journal of the Philosophy of Science* 29: 213-22.

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83: 435-50.

Natsoulas, T. (1974) The subjective, experiential element in perception. *Psychological Bulletin* 81: 611-31.

(1977) On perceptual aboutness. *Behaviorism* 5: 75-97.

(1978a) Haugeland's first hurdle. *Behavioral and Brain Sciences* 1: 243.

(1978b) Residual subjectivity. *American Psychologist* 33: 269-83.

(1980) Dimensions of perceptual awareness. Psychology Department. University of California, Davis. Manoscritto non pubblicato.

Nelson, K. & Gruendel, J. (1978) From person episode to social script: two dimensions in the development of event knowledge. Paper presented at the biennial meeting of the Society for Research in Child Development, San Francisco.

Newell, A. (1973) Production systems: models of control structures. In: *Visual information processing,* ed. W.C. Chase. New York: Academic Press.

(1979) Physical symbol systems. Lecture at the La Jolla Conference on Cognitive Science.

(1980) Harpy, production systems, and human cognition. In: *Perception and production of fluent speech,* ed. R. Cole. Hillsdale. N.J.: Erlbaum Press.

Newell, A. & Simon, H.A. (1963) GPS, a program that simulates human thought. In: *Computers and thought,* ed. A. Feigenbaum & V. Feldman, pp. 279-93. New York: McGraw-Hill.

Panofsky, E. (1954) *Galileo as a critic of the arts.* The Hague: Martinus Nijhoff.

Popper, K.R. & Eccles, J.C. (1977) *The self and its brain.* Heidelberg: Springer-Verlag.

Putnam, H. (1960) Minds and machines. In: *Dimensions of mind,* ed. S.

Hook. pp. 138-64 New York: Collier.
(1975a) The meaning of “meaning”. In: *Mind, language and reality.* Cambridge University Press.
(1975b) The nature of mental states. In: *Mind, language and reality.* Cambridge: Cambridge University Press.
(1975c) Philosophy and our mental life. In: *Mind, language and reality.* Cambridge: Cambridge University Press.

Pylyshyn, Z.W. (1980a) Computation and cognition: issues in the foundations of cognitive Science. *Behavioral and Brain Sciences* 3.
(1980b) Cognitive representation and the process-architecture distinction. *Behavioral and Brain Sciences.*

Russell, B. (1948) *Human knowledge: its scope and limits.* New York: Simon and Schuster.

Schank, R.C. & Abelson, R.P. (1977) *Scripts, plans, goals, and understanding,* Hillsdale, N.J.: Lawrence Erlbaum Press.

Searle, J.R. (1979a) Intentionality and the use of language. In: *Meaning and use,* ed. A. Margalit. Dordrecht: Reidel.
(1979b) The intentionality of intention and action. *Inquiry* 22: 253-80.
(1979c) What is an intentional state? *Mind* 88: 74-92.

Sherrington, C.S. (1950) Introductory. In: *The physical basis of mind,* ed. P. Laslett, Oxford: Basil Blackwell.

Slate, J.S. & Atkin, L.R. (1977) CHESS 4.5 — the Northwestern University chess program. In: *Chess skill in man and machine,* ed. P.W. Frey. New York, Heidelberg, Berlin: Springer-Verlag.

Sloman, A. (1978) *The computer revolution in philosophy.* Harvester Press and Humanities Press.
(1979) The primacy of non-communicative language. In: *The analysis of meaning (informatics 5),* ed. M. McCafferty & K. Gray. London: ASLIB and British Computer Society.

Smith, E.E.; Adams, N., & Schorr, D. (1978) Fact retrieval and the paradox of interference. *Cognitive Psychology* 10: 438-64.

Smythe, W.E. (1979) *The analogical/propositional debate about mental representation: a Goodmanian analysis.* Paper presented at the 5th annual meeting of the Society for Philosophy and Psychology, New York City.

Sperry, R.W. (1969) A modified concept of consciousness. *Psychological Review* 76: 532-36.
(1970) An objective approach to subjective experience: further explanation of a hypothesis. *Psychological Review* 77: 585-90.
(1976) Mental phenomena as causal determinants in brain function. In: *Consciousness and the brain,* ed. G.G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press.

Stick, S.P. (in preparazione) On the ascription of content. In: *Entertaining*

thoughts, ed. A. Woodfield.

Thorne, J.P. (1968) A computer model for the perception of syntactic structure. *Proceedings of the Royal Society of London* B 171: 377-86.

Turing, A.M. (1964) Computing machinery and intelligence. In: *Minds and machines,* ed. A.R. Anderson, pp. 4-30. Englewood Cliffs, N.J.: Prentice-Hall.

Weizenbaum, J. (1965) Eliza — a computer program for the study of natural language communication between man and machine. *Communication of the Association for Computing Machinery* 9: 36-45.

(1976) *Computer power and human reason.* San Francisco: W.H. Freeman.

Winograd. T. (1973) A procedural model of language understanding. In: *Computer models of thought and language,* ed. R. Schank & K. Colby, San Francisco: W.H. Freeman.

Winston, P.H. (1977) *Artificial intelligence.* Reading, Mass. Addison-Wesley.

Woodruff, G. & Premack, D. (1979) Intentional communication in the chimpanzee: the development of deception. *Cognition* 7: 333-62.

Note

[←1]
Vedi Michie D., *On Machine Intelligence*, New York, Wiley, 1974, p. 1.

[←2]

Vedi Newell A., *Artificial Intelligence and the Concept of Mind*, in Schank, R.C. e Colby, M.K., *Computer Models of Thought and Language*, San Francisco, Freeman, 1973, p. 1.

[←3]
Pamela McCorduck ricorda questa frase di Cobanis nel suo *Machines who Think, A Personal Inquiry into the History and Prospects of Artificial Intelligence*, San Francisco, Freeman, 1979, p. 36.

[←4]
Vedi Schank, R.C., *Language and Memory*, in *Cognitive Science*, 4, 1980, p. 244.

[←5]
Vedi Schank, R.C., *Conceptual Information Processing*, Amsterdam, North-Holland, 1975.

[←6]
Non sto, naturalmente, dicendo che Schank stesso sia impegnato in queste affermazioni.

[← 7]
“Comprendere” implica sia il possesso degli stati mentali (intenzionali) che la verità (validità, successo) di questi stati. Agli scopi di questa discussione ci interessa solo il possesso degli stati.

[←8]

L’intenzionalità è per definizione quella caratteristica di certi stati mentali per la quale essi sono diretti verso o riguardano oggetti e modi di essere della realtà nel mondo. Così opinioni, desideri e intenzioni sono stati intenzionali, forme di ansietà e depressione non lo sono. Per ulteriori discussioni vedi Searle (1979c).

[←9]

Mentre la versione rozza del comportamentismo è provata falsa da argomenti ben noti, c'è una versione più sofisticata che li evita: comunque, può essere provata falsa usando un esempio simile a quello che Searle usa contro Schank. Un tale esempio è delineato in Block 1978, p. 294, ed elaborato in Block di prossima pubblicazione.

[←10]
Assumo, per semplicità, che il cervello instanzi un solo programma (cosa, naturalmente, non vera). Notate, in proposito, che anche superare il test di Turing richiede di più che il semplice manipolare dei simboli. Un congegno che non aziona una macchina da scrivere, non può eseguire tale gioco.

[←11]

Per esempio, potrebbe essere che, dal punto di vista fisico, solo le cose che hanno gli stessi valori simultanei di peso e densità che il cervello ha possono fare le cose che il cervello può fare. Ciò sarebbe sorprendente, ma è difficile vedere perché uno psicologo dovrebbe occuparsene più di tanto. Nemmeno se risultasse — ancora come un dato di fatto reale — che i cervelli sono le sole cose che possono avere quel peso, densità e colore: se questo è dualismo, bene, immagino che possiamo vivere con esso.

[←12]
Questa caratterizzazione è necessariamente grezza e vaga. Per un'utile rassegna delle diverse versioni del funzionalismo e dei rispettivi punti deboli, cfr. Block (1978); io ho sviluppato e difeso quello che penso sia la più promettente versione del funzionalismo in Lycan (in corso di stampa).

[←13]
Per un’ulteriore discussione dei casi di questo genere, cfr. Block (in corso di stampa).

[←14]
Una versione molto esauriente di questa risposta appare nella sezione 4 di Lycan (in corso di stampa).

[←15]

Non capisco il suggerimento positivo di Searle per quanto riguarda la sorgente dell'intenzionalità nei nostri cervelli. Quali sono le "proprietà causali neurobiologiche"?

[←16]
Come Fodor (in corso di stampa) nota, SHRDLU come lo interpretiamo è la vittima di un cattivo demone cartesiano: i “blocchi” che egli manipola in realtà non esistono.

Indice