dalla rete alla URL

https://digilander.libero.it/massimo_miranda/
Searle Menti 1980.htm

Precedente

Menti, cervelli e programmi

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

(1980)

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 possono 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 ([6]), 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 intendo esaminare. Ma nulla di ciò che segue dipende nei dettagli dai programmi di Schank. Gli stessi argomenti si applicherebbero a SHRDLU di Winograd ([10]), ELIZA di Weizenbaum ([8]) 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.

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, dìversamente 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 constatare 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* ed *output*) dei nativi di madrelingua e, in aggiunta, assumiamo che i parlanti nativi hanno un certo livello di descrizione dove essi stessi sono istanze 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 che cosa 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*[1]: in molte di queste questioni si trovano numerose considerazioni assai semplici, ma 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 clacolatrice 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[2] 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 ([3]) 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.

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 potrebbe 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 manipolazioni di simboli'' *comprende* realmente il cinese. Il sottosistema dell'uomo che è il sistema di manipolazioni 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 manipolazioni 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 la storia 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 il test 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. [5]). 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 di 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 dell'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 princìpi 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'' ([2]).

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 al 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 soddisfazioni: 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 gli permeterebbero di ``agire'', e tutto questo sarebbe controllato dal suo cervello computerizzato. Un tale robot, diversamente dal computer di Schank, avrebbe una genuina capacità di comprendere 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. [1]). 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 gamba 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 del 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 della 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 o output è una condizione sufficiente, anzi costitutiva, di intenzionalità. Come Newell ([4]) 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.

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 sul 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 materiale 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 compiuta 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 attività 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 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 ([9]), 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 serie 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 insegna 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. [7]). 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 produca effettivamente 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 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 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 trattamnento 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''. M 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, trasfomarla e produrre infromazioni 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 soprendente 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 saparabile 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 della mente è specificamente mentale 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.

References

[1]

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences 3*:1.

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

[3] Newell, A. e Simon, H. A. (1963). GPS, a Program that Simulates Human Thought. In: *Computers and Thought*, ed. A. Feigenbaum e V. Feldman, pp. 279-93. New York: McGraw Hill.

[4] Newell, A. (1979). Physical Symbol Systems. Lectures at the La Jolla Conference on Cognitive Science.

[5] Pylyshyn, Z. W. (1980). Computation and Cognition: Issues in the Foundations of Cognitive Science. *The Behavioral and Brain Sciences 3*.

[6] Schank, R. C. e Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*, Hillsdale, N.J.: Lawrence Erlbaum Press.

[7] Searle, J. R. (1979). What is an intentional state? *Mind* 88: 74-92.

[8] 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.

[9] Weizenbaum, J. (1976). *Compuer Power and Human Reason*. San Francisco: W. H. Freeman.

[10] Winograd, T. (1973). A Procedural Model of Language Understanding. In: *Computer Models of Thought and Language*, ed. R. Schank e K. Colby, San Francisco: W. H. Freeman.

Precedente

Footnotes:

[1] "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.

[2] 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 [7].