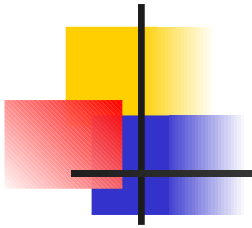
Intelligenza e macchine: la questione dell'intenzionalità

Angelo Montanari
Dipartimento di Matematica e Informatica
Università degli Studi di Udine

Udine, 24 maggio, 2012



Sommario

- La questione dell'intenzionalità
- Riduzionismo e intelligibilità delle macchine
- Le macchine informazionali

- Un approccio comportamentista: il test di Turing
- La società della mente di Minsky
- Menti, cervelli e programmi: la stanza cinese di Searle

- Conclusioni

La nozione di intenzionalità

Nella riflessione filosofica (Brentano, Husserl, Carnap), l'intenzionalità viene riconosciuta quale elemento distintivo della coscienza (in generale, di ogni fenomeno psichico)

Per Brentano, l'intenzionalità è il carattere costitutivo di ogni **fenomeno psichico**

Per Husserl, l'intenzionalità è il carattere costitutivo della **coscienza** e del rapporto soggetto (umano) - oggetto.

Compito della filosofia è descrivere la struttura immanente con cui l'oggetto è intenzionato dalla coscienza

Una definizione

In generale, possiamo definire l'**intenzionalità** come il riferimento interno di un atto o di uno stato mentale a un determinato oggetto, ossia la connessione che l'atto o lo stato hanno, in virtù della loro identità, con un certo oggetto, indipendentemente dalla sussistenza di questo eventuale oggetto nella realtà esterna.

Esempio. Dell'identità di uno stato emotivo di speranza fa parte ciò che è sperato, indipendentemente dal fatto che si realizzi oppure no.

Intenzionalità e macchine

Questione: "si può dare intenzionalità nelle macchine?"

Questioni collegate/sottese: qual è il rapporto tra **menti** (**persone umane**) e **macchine**? Si può instaurare una corrispondenza tra stati mentali/cerebrali e stati di una macchina? Possiamo parlare di (auto)coscienza delle macchine (ad esempio, rispetto al problema della responsabilità delle macchine)?

L'**approccio riduzionista**

Angelo Montanari, "Riduzionismo e non in Intelligenza Artificiale", Anthropologica, Annuario di Studi Filosofici, 2009, pp. 113-128.

L'approccio riduzionista

Riduzionismo: posizione di chi riconduce le proprietà di un'entità complessa (oggetto, sistema o organismo) alla "somma" delle caratteristiche delle sue singole componenti

Questione fondamentale: cosa vuole dire **somma**?

Per i riduzionisti, il modo in cui le caratteristiche delle componenti elementari concorrono alla determinazione delle caratteristiche del composto può essere definito in modo **semplice** e **chiaro**

Per chi si oppone al riduzionismo, la debolezza della posizione riduzionista si manifesta nella **complessità** delle interazioni fra le componenti, non riducibili alle proprietà delle singole componenti

Le forme del riduzionismo

Il riduzionismo è presente in forme diverse in discipline diverse, ma vi sono forti **contaminazioni** fra i vari ambiti

Esempi.

- Posizioni riduzioniste sviluppate a livello di riflessione filosofica e di studi di psicologia, quali il funzionalismo e il comportamentismo, hanno pesantemente influenzato le linee di sviluppo della ricerca in cibernetica prima e IA poi
- Influenza della ricerca in neurofisiologia, in particolare delle tecniche di imaging funzionale, sui più recenti sviluppi dell'IA nell'ambito della bionica

Riduzionismo filosofico e **riduzionismo scientifico**

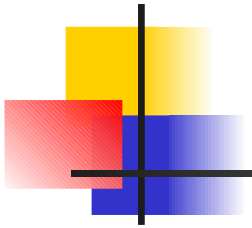
Il riduzionismo scientifico

Efficacia pratica dell'applicazione dello schema riduzionistico in ambito scientifico e ingegneristico

La possibilità di definire il comportamento di un sistema complesso in termini di proprietà ed interazioni delle sue componenti elementari è stato uno dei fattori chiave nello sviluppo di numerose discipline scientifiche

Esempio: la **fisica**

Una grandissima varietà di fenomeni viene ricondotta all'interazione di un insieme ridotto di particelle e campi di forze



Riduzionismo e macchine - 1

L'**intelligibilità delle macchine**, ossia la possibilità di descriverne in modo comprensibile le caratteristiche strutturali e funzionali e le tecniche di costruzione, è condizione essenziale per il loro sviluppo e il loro utilizzo

Solo l'esistenza di una **spiegazione adeguata** (razionale) del funzionamento di una macchina complessa consente, infatti, di predirne, nei limiti del possibile, il comportamento e di diagnosticarne gli eventuali guasti e malfunzionamenti

Riduzionismo e macchine - 2

La **spiegazione** mediante il **paradigma riduzionista**: l'analisi del sistema nel suo complesso viene ridotta all'analisi separata delle sue componenti elementari e delle loro interazioni

Efficace nel caso di macchine relativamente semplici, tale approccio diventa problematico in presenza di **meccanismi di controllo** (meccanismi di anticipazione e meccanismi di retroazione). Tali meccanismi possono essere visti come il tentativo di introdurre nella macchina un'opportuna rappresentazione dell'obiettivo (causa finale) per il quale la macchina è stata costruita

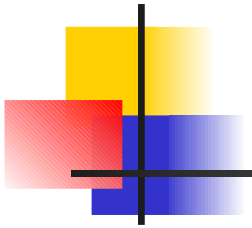
Le macchine informazionali

Macchine cibernetiche (o informazionali): macchine che incorporano sofisticati meccanismi di controllo

Esempio. I **sistemi di intelligenza artificiale**

E' problematico ricorrere ad uno schema riduzionista di tipo tradizionale per spiegare il funzionamento di tali macchine

Ciò nonostante, non viene meno il tentativo di assimilazione dell'uomo ad una particolare classe di macchine: uomo come macchina di natura meccanica; successivamente, uomo come macchina termodinamica prima e come macchina chimica poi; nel secolo scorso **uomo come macchina informazionale**



Un'osservazione

- Ragioni e problematiche relative all'uso di un **linguaggio antropomorfo** nella descrizione delle caratteristiche e del funzionamento di una macchina
- Ciò è particolarmente evidente nel caso dei sistemi di **intelligenza artificiale** (memoria, comprensione, apprendimento, ragionamento, ..), ma si è verificato in misura più o meno rilevante in molti altri casi
- Una possibile ragione: l'uomo come modello (**cibernetica**)
- Due approcci: simulazione vs. emulazione

Alcune figure paradigmatiche

Illustreremo i passaggi fondamentali della riflessione sul rapporto tra macchine informazionali (sistemi di intelligenza artificiale) e uomo attraverso la descrizione del contributo di alcune figure paradigmatiche:

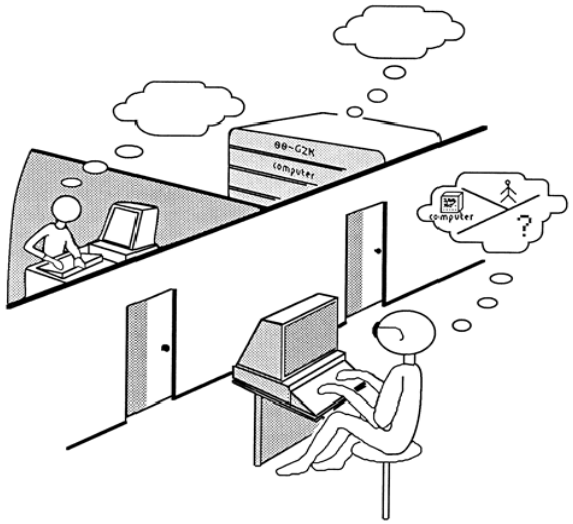
Alan M. **Turing** (Computing Machinery and Intelligence, in «Mind», volume 59, 1950, pp. 433-460)

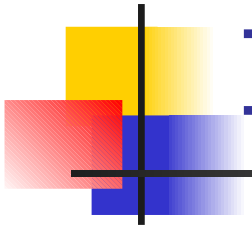
Marvin **Minsky** ("The society of mind", Simon and Schuster, 1986)

John R. **Searle** ("Minds, brains, and programs", Behavioral and Brain Sciences, volume 3, 1980, pp. 417-424)

Il test di Turing

Il **test di Turing** (o gioco dell'imitazione): una macchina può essere definita intelligente se riesce a convincere una persona che il suo comportamento, dal punto di vista intellettuale, non è diverso da quello di un essere umano medio





Il test di Turing: dettagli - 1

Il test si svolge in tre stanze separate. Nella prima si trova l'esaminatore umano (A); nelle altre due vi sono rispettivamente un'altra persona e il computer che si sottopone al test. Dei due A conosce i nomi (B e C), ma ignora chi sia la persona e chi il computer.

Sia B che C si relazionano separatamente con A attraverso un computer. Via computer A può porre domande a B e C e leggere le loro risposte. Compito di A è scoprire l'identità di B e C (**chi è la persona, chi è la macchina?**) entro un limite di tempo prefissato.

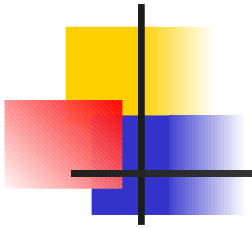
Il test di Turing: dettagli - 2

A può effettuare qualunque tipo di domanda; il computer ovviamente cercherà di rispondere in modo tale da celare la propria identità. La macchina supera il test se A non riesce a identificarla nel tempo prefissato. Il test verrà ripetuto più volte, coinvolgendo anche esaminatori diversi, in modo da ridurre i margini di soggettività.

Osservazioni.

1. **astrazione** da tutti gli elementi di contorno (in particolare, dalla conformazione dei soggetti, dalle loro caratteristiche fisiche): un pensiero disincarnato;

2. interpretazione operativa/comportamentale dell'intelligenza che si manifesta attraverso la **comunicazione linguistica** (stretto legame tra intelligenza e capacità linguistiche).



Alcuni anni dopo.. Minsky e Searle

Questione: possibilità/impossibilità di assimilare le capacità cognitive dell'uomo (la sua **mente** / il suo **cervello**) ad un sistema artificiale (una **macchina**)

Posizione riduzionista *(Minsky)*: mente e cervello, descritti come una comunità di agenti interagenti, raggruppati in agenzie; possibilità di assimilare il cervello ad una macchina

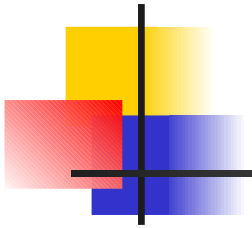
Posizione anti-riduzionista *(Searle)*: sistemi di IA visti come "macchine sintattiche"; impossibilità per tali sistemi di possedere un'intenzionalità, caratteristica distintiva degli esseri umani (e animali)

Il riduzionismo di Minsky

Obiettivo: spiegare l'intelligenza come una combinazione di cose più semplici

il cervello come macchina

Citando Minsky, "non vi alcun motivo per credere che il **cervello** sia qualcosa di diverso da una **macchina** con un numero enorme di componenti che funzionano in perfetto accordo con le leggi della fisica"



Rapporto mente-cervello

Rapporto tra mente e cervello: la mente è semplicemente ciò che fa il cervello (la **mente come processo**). Analogia con la distinzione tra programma e processo (programma in esecuzione) in informatica

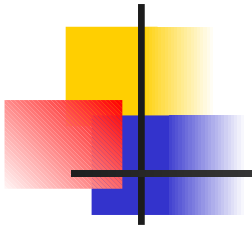
Problema: per spiegare la mente evitando la circolarità occorre descrivere il modo in cui le menti sono costruite a partire da materia priva di mente, parti molto più piccole e più semplici di tutto ciò che può essere considerato intelligente

Questione: una mente può essere associata solo ad un cervello o, invece, qualità tipiche della mente possono appartenere, in grado diverso, a tutte le cose?

La società della mente

Per Minsky, **cervello** come **società organizzata**, composta da una molteplicità di componenti organizzate in modo gerarchico, alcune delle quali operano in modo del tutto autonomo, la maggior parte in un rapporto alle volte di collaborazione, più spesso di competizione, con altre componenti

Intelligenza umana frutto dell'interazione di un numero enorme di componenti fortemente diverse fra loro, i cosiddetti **agenti della mente**, componenti elementari ("particelle") di una (teoria della) mente



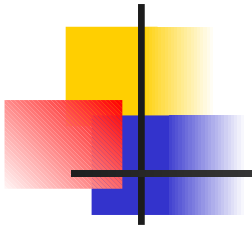
La nozione di agenzia

Questione: come può l'opera combinata di un insieme di agenti produrre un comportamento che ogni singolo agente, considerato separatamente, non è in grado di fornire?

La **nozione di agenzia** come superamento di posizioni di riduzionismo ingenuo difficilmente sostenibili (Minski contesta chi considera la fisica e la chimica modelli ideali di come dovrebbe essere la psicologia)

Un'agenzia è un insieme di agenti collegati fra loro da un'opportuna rete di interconnessioni

La **gerarchia delle agenzie**



L'antiriduzionismo di Searle

Assunto fondamentale: impossibilità per una macchina di manifestare l'**intenzionalità** che caratterizza gli esseri umani e, sia pure in forme diverse, gli animali

Per Searle, l'intenzionalità è un **dato di fatto empirico** circa le effettive relazioni causali tra mente e cervello, che consente (unicamente) di affermare che certi processi cerebrali sono sufficienti per l'intenzionalità

L'esecuzione di un programma su un dato input (**istanziazione di un programma** nella terminologia di Searle, **processo** nel linguaggio informatico comune) non è mai di per se stessa una condizione sufficiente per l'intenzionalità

Un esperimento mentale

La "dimostrazione": sostituire un agente umano al calcolatore nel ruolo di esecutore di una specifica istanza di un programma e mostrare come tale esecuzione possa avvenire senza forme significative di intenzionalità

La struttura dell'**esperimento mentale**: una teoria della mente può essere confermata/falsificata immaginando che la propria mente operi secondo i principi di tale teoria e verificando la validità o meno delle affermazioni/previsioni della teoria

L'**esperimento di Searle**: Searle prende in esame i lavori sulla simulazione della capacità umana di **comprendere narrazioni**, che richiede la capacità di rispondere a domande che coinvolgono informazioni non fornite in modo esplicito dalla narrazione, ma desumibili da essa sfruttando conoscenze di natura generale

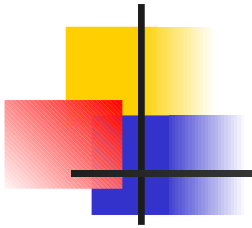
La stanza cinese

L'esperimento in dettaglio - 1

Searle immagina che una persona venga chiusa in una stanza e riceva **3 gruppi di testi** scritti in una lingua a lei sconosciuta (**cinese**), interpretabili (da chi fornisce i testi) rispettivamente come il testo di una narrazione, un insieme di conoscenze di senso comune sul domino della narrazione, e un insieme di domande relative alla narrazione.

Immagina, inoltre, che tale persona riceva un **insieme di regole**, espresse nella propria lingua (**inglese**), che consentano di collegare in modo preciso i simboli formali che compaiono nel primo gruppo di testi a quelli che compaiono nel secondo e **un altro insieme di regole**, anch'esse scritte in una lingua a lei nota, che permettano di collegare i simboli formali che compaiono nel terzo gruppo di testi a quelli degli altri due e che rendano possibile la produzione di opportuni simboli formali in corrispondenza di certi simboli presenti nel terzo gruppo di testi.

Le **regole** vengono interpretate (da chi le fornisce) come un **programma** e i **simboli prodotti** come **risposte** alle domande poste attraverso il terzo gruppo di testi. Quanto più il programma è ben scritto e l'esecuzione delle regole spedita, tanto più il comportamento della persona sarà assimilabile a quello di un parlante nativo (un cinese).



L'esperimento in dettaglio - 2

Immaginiamo ora uno scenario in cui la persona riceva il testo narrativo e le domande ad esso relative nella propria lingua (**inglese**) e fornisca le risposte in tale lingua, sfruttando la propria conoscenza di senso comune.

Tali risposte saranno indistinguibili da quelle di un qualunque altro parlante nativo, in quanto la persona è un parlante nativo. Dal punto di vista esterno, le risposte fornite in lingua cinese e quelle fornite in lingua inglese saranno egualmente buone; il modo in cui vengono prodotte è, però, radicalmente diverso.

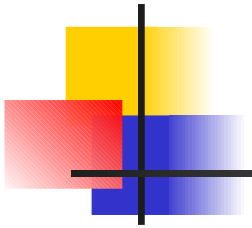
A differenza del secondo caso, nel primo caso le risposte vengono ottenute attraverso un'opportuna manipolazione algoritmica di simboli formali ai quali la persona non associa alcun significato (simboli non interpretati). Il **comportamento della persona** è, in questo caso, del tutto **assimilabile** all'**esecuzione di un programma** su una specifica istanza (processo) da parte di un sistema artificiale.

Esito dell'esperimento

Risultato: la capacità (di un uomo/una macchina) di manipolare le informazioni ricevute secondo regole formali ben definite non è sufficiente a spiegare il processo di comprensione (non vi è nemmeno alcuna evidenza che essa debba essere una condizione necessaria) – "carattere non intenzionale, e, quindi, semanticamente vuoto, dei simboli elaborati da un sistema artificiale" (Diego Marconi)

Conclusione: i processi mentali non possano essere ridotti a processi di natura computazionale che operano su elementi formalmente definiti

Osservazione: confutazione della validità del cosiddetto test di Turing



Conseguenze

L'affermazione dell'**irriducibilità** dell'intenzionalità all'esecuzione di programmi su particolari input ha alcune importanti conseguenze:

- dall'IA forte all'IA debole (cauta/prudente)
- condizioni per un'intenzionalità artificiale

IA forte e debole

Prima conseguenza: impossibilità di spiegare le modalità con le quali il cervello produce l'intenzionalità attraverso il meccanismo dell'istanziazione di programmi

Contro un'**interpretazione forte** dell'IA: non vi è alcuna distinzione sostanziale tra mente umana e un computer opportunamente programmato

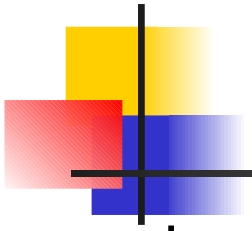
Per un'**interpretazione debole** dell'IA: strumento per lo studio delle capacità cognitive dell'uomo (formulazione precisa e verifica rigorosa di ipotesi su determinati aspetti di specifiche abilità cognitive attraverso sviluppo e validazione di opportuni modelli)

Un'intenzionalità artificiale?

Problemi (**irrisolti**): cosa differenzia il caso in cui la persona comprende il testo (inglese) da quello in cui non vi è alcuna comprensione (cinese)? Questo qualcosa può (se sì, come) essere trasferito ad un macchina?

Seconda conseguenza: ogni meccanismo in grado di produrre intenzionalità deve avere abilità di tipo causale pari a quelle del cervello

Per Searle, ogni eventuale tentativo di creare un'intenzionalità artificiale non può ridursi allo sviluppo di un qualsivoglia programma, ma richiede la capacità di replicare le **abilità causali** tipiche della mente umana



Due approfondimenti

Legame tra **intenzionalità** e capacità di creare degli **artefatti**: l'intenzionalità si manifesta nella sintesi dei programmi, ma non si trasferisce al programma sintetizzato (al programma in sé)

Le ragioni dell'**inadeguatezza** dei **sistemi artificiali / formali**: impossibilità di sintetizzare un sistema (formale) corretto e completo in grado di catturare il processo di comprensione (lo stesso per le altre capacità cognitive)

Angelo Montanari, Alcune questioni di tecnoetica dal punto di vista di un informatico, Teoria XXVII/2, 2007, pp. 57-72

Conclusioni

Si può parlare di "intenzionalità" delle macchine? **Critica di Searle** all'assimilazione della mente/cervello ad un calcolatore

Nonostante le molte osservazioni critiche di cui è stato oggetto (ad esempio, plausibilità/legittimità degli esperimenti mentali), il lavoro di Searle rimane un punto di riferimento fondamentale

Applicazioni recenti: i babbuini e la lettura – un esperimento presso l'Università di Aix-Marsiglia (si veda la riflessione di Giuseppe Longo, Università di Trieste, su Avvenire 19-04-2012)

Rilevanza dell'argomento di Searle rispetto alle problematiche legate al rapporto **mente/cervello** in neurofisiologia (le coppie sofware/hardware o processo/programma non sono un buon modello) - plasticità composizionale e computazionale degli stati mentali (Putnam)