Libertà e macchine

Il credente tra neuroscienze e **intelligenza artificiale -** TAVOLA ROTONDA

Angelo Montanari

Dipartimento di Matematica e Informatica

Università degli Studi di Udine

Sacile, 6 giugno, 2014

Sommario

INTRODUZIONE

- Libertà e macchine: la questione dell'intenzionalità
- Riduzionismo e intelligibilità delle macchine

TRE FIGURE DI RIFERIMENTO

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

LA BIONICA

CONCLUSIONI

Libertà delle macchine?

La questione: “si può parlare di libertà delle macchine?” può essere declinata in vari modi sostanzialmente equivalenti (il riferimento privilegiato, ma non esclusivo, è ai sistemi di intelligenza artificiale):

- macchine e coscienza
- macchine e **intenzionalità**
- responsabilità delle macchine

Intenzionalità e macchine

La **questione**: "si può parlare di libertà delle macchine?" può essere riformulata come: "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**

Riduzionismo e macchine

L'affermazione circa la possibilità di riprodurre artificialmente caratteristiche e comportamento dell'essere umano (sistemi di Intelligenza Artificiale) si presta a due letture apparentemente speculari:

(i) se il comportamento di una macchina sarà assimilabile a quello dell'uomo, sarà possibile parlare di **libertà**, e responsabilità, **delle macchine**

(ii) se il comportamento dell'uomo sarà assimilabile a quello di una macchina, non sarà più possibile parlare (e, retrospettivamente, si è parlato inappropriatamente) di libertà, e responsabilità, dell'uomo (**riduzionismo**)

Alcune figure paradigmatiche

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

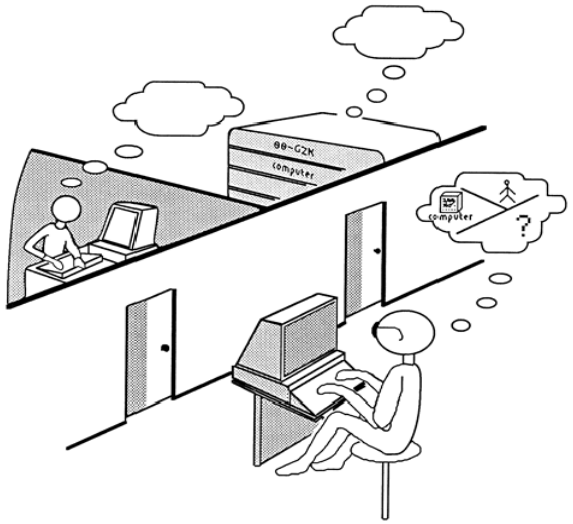
Alan M. **Turing** (Computing Machinery and Intelligence, in «Mind», volume 59, 1950, pp. 433-460)

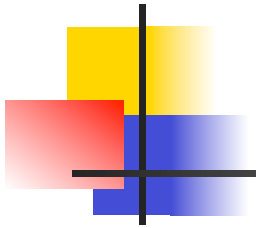
Marvin **Minsky** ("The society of mind", Simon and Schuster, 1986)

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 (l'influsso del **comportamentismo**)





Il sistema ELIZA (DOCTOR)

- Senza disporre di informazioni significative su emozioni e pensieri delle persone, **DOCTOR** realizza delle interazioni con gli utenti/pazienti del tutto simili a quelle umane
- Quando il "paziente" esce dai confini della piccola base di conoscenze del sistema, DOCTOR fornisce delle risposte generiche.
- Ad esempio, all'affermazione: "mi fa male la testa", DOCTOR potrebbe rispondere con la frase: "Perché dici che ti fa male la testa?"

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

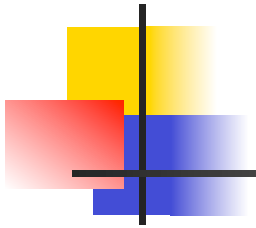
Posizione anti-riduzionista *(Searle)*: sistemi di IA visti come "macchine sintattiche"; impossibilità per tali sistemi di possedere un'intenzionalità, caratteristica distintiva degli esseri umani (e animali)

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"

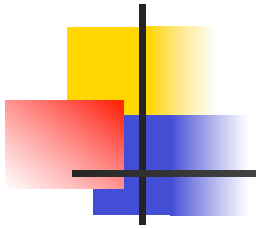


Rapporto mente-cervello

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

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

Cervello come **società organizzata**, composta da una molteplicità di componenti organizzate in modo gerarchico, alcune 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 (**agenti della mente)**. Insiemi di agenti collegati fra loro da una rete di interconnessioni: **agenzia**

Le **teorie degli agenti**, che occupano la scena dell' Intelligenza Artificiale contemporanea, muovono dal modello di Minsky

L'antiriduzionismo di Searle

L'**esperimento (mentale) di Searle:** Searle immagina di sostituire un agente umano al calcolatore nel ruolo di esecutore di una specifica istanza di un programma e mostra come tale esecuzione possa avvenire senza forme significative di intenzionalità

Contesto: simulazione della capacità umana di **comprendere un testo narrativo**

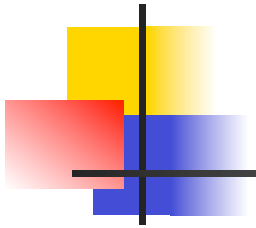
Caratteristica distintiva di tale abilità: la capacità di rispondere a domande che coinvolgono informazioni non fornite in modo esplicito dalla narrazione, ma desumibili da essa sfruttando conoscenze di natura generale

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

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

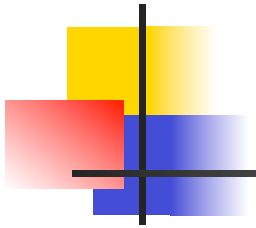
Osservazione: confutazione della validità del cosiddetto test di Turing

Alcune questioni (teoria)

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) – INTENZIONALITA' DERIVATA

Le ragioni dell'**inadeguatezza** dei **sistemi artificiali / formali**: impossibilità di sintetizzare un sistema 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



Alcune questioni (pratica)

Tesi: risultati fondamentali in informatica illuminano aspetti critici del funzionamento dei sistemi artificiali informatici

Questioni:

- possiamo sempre controllare/predire il comportamento di un sistema?
- possiamo sempre garantire la presenza di esseri umani nei cicli di controllo (control loop)?
- possiamo sempre confidare nelle credenze di un sistema?
- può un sistema portare a termine qualunque compito ad esso assegnato?

Nuovi scenari in IA

Gli sviluppi più recenti della ricerca in IA nell'ambito della **bionica** introducono nella discussione elementi del tutto nuovi

Non si dà intelligenza nell'uomo senza **corporeità** (la nostra interazione col mondo è mediata dagli organi di senso). La **robotica**: insieme delle teorie e delle tecniche per la costruzione e l'utilizzo dei robot

La **bionica**: uomo e macchina come sistema integrato (gli organismi cibernetici, i cosiddetti cyborg). Un nuovo tipo di **protesi**: dal recupero (artificiale) di funzionalità perdute all'introduzione di nuove funzionalità (potenziamento)

Un'idea diversa di **vita buona**

Interazioni brain-actuated

Le interazioni brain-actuated usano i risultati delle neuroscienze relativi alla caratterizzazione dell'attività svolta dalle diverse aree di tessuto nervoso

Diverse modalità di indagine: dalla registrazione dei segnali elettrici generati a livello di singolo neurone (impianto chirurgico di elettrodi) all'analisi delle variazioni dell'attività metabolica relativa a specifiche aree del sistema nervoso (dispositivi di superficie)

Corrispondentemente, **interazioni brain-actuated** realizzate in modo invasivo (impianto chirurgico di elettrodi) o non invasivo (elettrodi di superficie che registrano segnali di elettroencefalogrammi)

Stati mentali e stati cerebrali

L'esistenza di uno **stretto legame** tra l'attività mentale e l'attività fisico-chimica del cervello è testimoniata da una grande massa di dati sperimentali

I recenti risultati ottenuti nell'ambito della neurofisiologia, utilizzati per la realizzazione di interazioni brain-actuated, vengono letti da molti attraverso uno schema riduzionista che **assimila il mentale al cerebrale**, posizione già presente sia in Minsky sia in Searle

Stati mentali e cerebrali possono essere **identificati**?

- (im)possibilità di circoscrivere le aree cerebrali coinvolte in una determinata attività mentale (logica della localizzazione)
- (ir)riducibilità delle attività mentali ad antecedenti fisico-chimici del cervello

Il mondo in prima persona

Come spiegare, (i) la costruzione di una rappresentazione mentale degli eventi fisici del mondo (ad esempio, nella percezione visiva cosciente) e (ii) il passaggio dall'evento mentale della decisione di compiere una determinata azione alle attività cerebrale e muscolare che la realizzano?

In generale, come cogliere l'intrinseca soggettività di ogni evento mentale (esempio, il dolore fisico)? La conoscenza degli stati elettrochimici del cervello, direttamente indagabili in terza persona, non riesce a descrivere in modo esauriente **il mondo in prima persona**

F. Tempia, Attività cerebrale e rappresentazioni mentali, in Scienze Informatiche e Biologiche. Epistemologia e Ontologia, Città Nuova, 2011, pp. 185-208.

Conclusioni

Si può parlare di "libertà" delle macchine (e degli esseri umani)?

Due forme di **riduzionismo**:

- assimilazione della mente ad un calcolatore (critica di Searle)
- assimilazione della mente al cervello (questione aperta)

(In)adeguatezza della nozione di **proprietà emergente**

il ruolo delle **relazioni** fra gli elementi presenti ad un certo livello di descrizione di un sistema e l'influenza che esse esercitano sui livelli superiori di descrizione del sistema, a loro volta caratterizzati da un insieme di elementi (in generale diversi) fra loro in relazione: Qual è la natura delle relazioni? Qual è il loro "substrato materiale"?