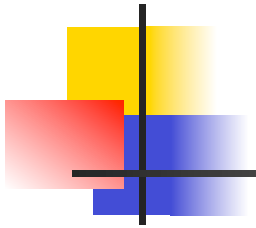
La coscienza: aspetti informatici

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

SCUOLA SEFIR - Perugia, 5 aprile, 2013



Sommario

Alcune **questioni preliminari**:

- Della neutralità della scienza (informatica)
- Dell'uso di un vocabolario antropomorfico
- (Della coscienza) dell'uomo e della macchina

La **coscienza** e l'informatica:

- Il test di Turing e il sistema ELIZA
- La coscienza nella "società della mente" di Minsky
- Le teorie degli agenti in intelligenza artificiale

Coda:

- Alcune considerazioni
- Piccola digressione: le interfacce uomo-computer

Della neutralità della scienza

A dispetto della presunta **neutralità** della scienza, le posizioni ideologiche/culturali/filosofiche dei ricercatori (le loro metafisiche) influenzano la loro ricerca scientifica

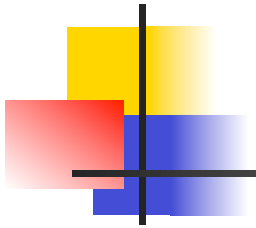
Ciò è vero, ad esempio, nell'ambito della matematica

- platonici vs. formalisti
- l'infinito attuale di Cantor
- la matematica costruttivista (intuizionista) di Brouwer

Della neutralità dell'informatica

Ciò è ancor più vero nell'ambito dell'informatica:

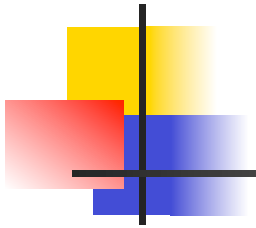
- influenza del **comportamentismo** di Watson (l'introspezione non può fornire alcun dato affidabile; alternativa: studio esclusivo delle misurazioni delle percezioni/stimoli forniti ad un animale e delle azioni/risposte risultanti) sul famoso **test di Turing** (test proposto da Turing per stabilire se una macchina può essere definita intelligente)
- legame tra **l'approccio riduzionista** alla base della "società della mente" di Minsky (cervello come società organizzata, composta da una molteplicità di componenti fortemente diverse fra loro, i cosiddetti agenti della mente) e le **teorie degli agenti** che tanto spazio hanno nella ricerca contemporanea in intelligenza artificiale (IA)



Un vocabolario antropomorfico

L'uso di un **vocabolario antropomorfico** nella descrizione delle caratteristiche e del funzionamento dei sistemi informatici

- è particolarmente evidente nel caso dei sistemi di IA (intelligenza, conoscenza, apprendimento, ragionamento),
- ma si è verificato in misura più o meno rilevante in molti altri casi (memoria, comunicazione, interrogazione)



Potenza e limiti

Ragioni dell'uso di un vocabolario antropomorfico:

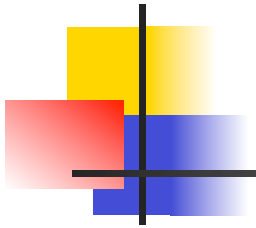
- uomo/animale come modello (fonte di ispirazione) in cibernetica (Cybernetis or Control and Communication in the Animal and the Machine, N. Wiener) e successivamente in molti ambiti dell'informatica (IA, robotica, bionica, ..)

Conseguenze circa il rapporto tra l'uomo e il calcolatore:

- simulazione vs. emulazione (IA forte e debole/cauta)

La coscienza: le macchine e noi

- Osservazione: ogni **discorso** sulle proprietà "antropomorfiche" delle macchine/calcolatori non riguarda tanto la macchina (il calcolatore) in sé, ma il modo in cui noi vediamo la macchina, e indirettamente noi stessi (Minsky rivendica la legittimità/utilità dell'uso di termini antropomorfici)
- Ciò vale anche per la questione relativa alla **coscienza nelle/delle macchine**: parlare delle macchine è un modo (indiretto) per parlare di noi stessi



Innalzare vs. ridurre

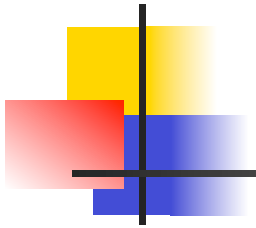
Se ogni discorso sulle caratteristiche delle macchine (calcolatori) è, in realtà, un discorso sull'uomo,

innalzare la macchina/calcolatore al livello dell'uomo

e

ridurre l'uomo al livello della macchina/calcolatore

sono due movimenti solo apparentemente opposti



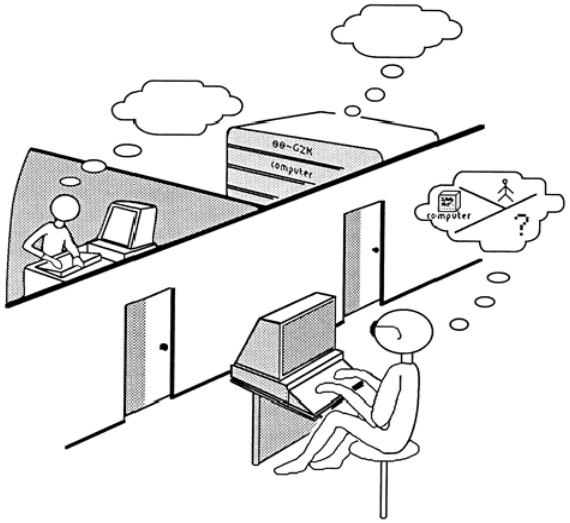
L'approccio comportamentista

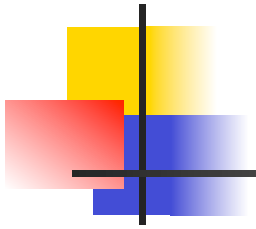
Il movimento dell'**innalzare**

- Il **test** di **Turing**: come stabilire se una macchina artificiale è intelligente
- Un curioso esperimento: **ELIZA**. ELIZA è un semplice programma per l'elaborazione del linguaggio naturale che gestisce le risposte fornite dagli utenti ad alcuni script predefiniti (il più famoso è lo script DOCTOR, che simula una seduta psicoanalitica)
- La confutazione del test di Turing: la **stanza cinese** di **Searle**

Il test di Turing

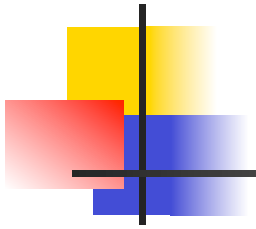
Il **test di Turing** (o gioco dell'imitazione): una macchina può essere definita intelligente se riesce a convincere una persona che il suo comportamento, dal punto di vista intellettuale, non è diverso da quello di un essere umano medio





Il test di Turing: dettagli - 1

- Il test si svolge in 3 stanze separate. Nella prima si trova l'esaminatore umano (A); nelle altre 2 vi sono 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, 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

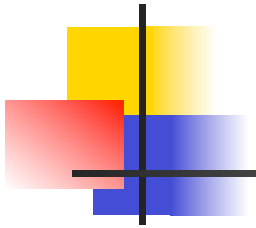


Il test di Turing: dettagli - 2

- 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

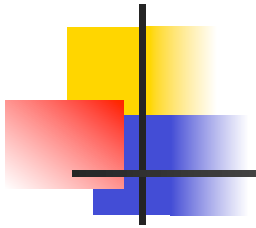
Alcune considerazioni

- Una **nozione astratta** (disincarnata) di **intelligenza**: il test ha quale prerequisito un'astrazione da tutti gli elementi "di contorno", in particolare, dalla conformazione / caratteristiche fisiche dei soggetti coinvolti
- Un'**interpretazione operativa/comportamentale** dell'intelligenza: un soggetto può essere definito intelligente o meno esclusivamente sulla base del suo comportamento
- Il ruolo del **linguaggio**: il test stabilisce uno stretto legame tra intelligenza e capacità linguistiche (l'intelligenza si manifesta attraverso la comunicazione linguistica)



Il sistema ELIZA: lo script 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?"

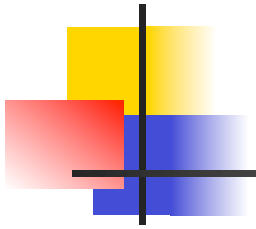


Il funzionamento di ELIZA

- ELIZA venne implementato mediante semplici tecniche di **pattern matching**: parsing delle frasi del paziente e sostituzione di parole chiave in frasi preconfezionate
- Nonostante la sua semplicità, ELIZA venne preso sul serio da molti dei suoi utenti, anche dopo la spiegazione del suo funzionamento da parte dello sviluppatore (Weizenbaum)
- La **confutazione** di **Weizenbuam**: nel suo libro "Computer Power and Human Reason: From Judgment to Calculation", Weizenbaum discute i limiti dei calcolatori affermando che le visioni antropomorfiche dei computer sono delle ingiustificate riduzioni degli esseri umani

La confutazione di Searle

- **Intenzionalità** (dimensione essenziale della **coscienza**): caratteristica che contraddistingue certi stati mentali, quali le credenze, i desideri e le intenzioni, diretti verso oggetti e situazioni del mondo
- Assunto fondamentale: impossibilità per una macchina computazionale 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



Intenzionalità e computer

Per Searle, l'esecuzione di un programma su un dato input non è mai di per se stessa una condizione sufficiente per l'intenzionalità

Dimostrazione (via esperimento mentale): sostituire un agente umano al computer nel ruolo di esecutore di una specifica istanza di un programma e mostrare come tale esecuzione possa avvenire senza forme significative di intenzionalità

Searle prende in esame i lavori sulla simulazione della capacità umana di **comprendere narrazioni**, che richiede l'abilità 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 - 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.
- Assume che tale persona riceva un **insieme di regole**, espresse nella propria lingua (**inglese**), che permettano 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 inglese, che consentano di collegare i simboli formali che compaiono nel terzo gruppo di testi a quelli degli altri due e di produrre 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).

La stanza cinese - 2

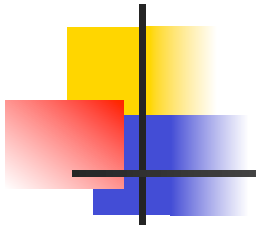
- Immaginiamo ora uno scenario in cui la persona riceva il testo narrativo e le domande ad esso relative nella propria lingua (**inglese**) e fornisca le risposte in tale lingua, sfruttando la propria conoscenza di senso comune.
- Tali risposte saranno indistinguibili da quelle di un qualunque altro 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 ("carattere non intenzionale, e, quindi, semanticamente vuoto, dei simboli elaborati da un sistema artificiale", Marconi)

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

Confutazione della validità del test di Turing



IA forte e debole

- **Prima conseguenza**: impossibilità di spiegare le modalità con le quali il cervello produce l'intenzionalità attraverso il meccanismo dell'esecuzione 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

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?

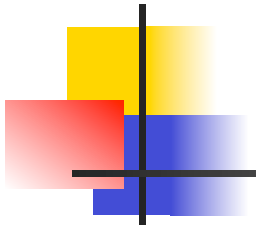
Seconda conseguenza: ogni meccanismo in grado di produrre intenzionalità deve avere abilità di tipo causale pari a quelle del cervello

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

L'approccio riduzionista

Il movimento del **ridurre**

- Nel testo in cui riassume in modo organico il proprio punto di vista sul rapporto tra mente/cervello e calcolatore ("La società della mente"), Minsky intende mostrare come la mente possa essere spiegata in termini di una combinazione di cose più semplici prive di mente
- "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"
- "quando sapremo quali **macchine meravigliose** siamo, quali **complicati circuiti** costituiscono la macchina della mente, avremo maggior rispetto per noi stessi"



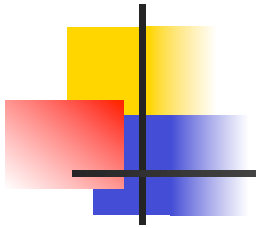
Rapporto mente-cervello

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

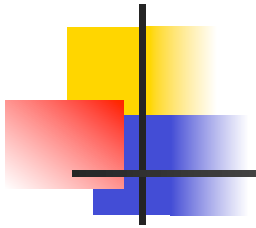
- **Cervello** come **società organizzata**, composta da una molteplicità di componenti organizzate in modo gerarchico, alcune delle quali operano in modo del tutto autonomo, la maggior parte in un rapporto alle volte di collaborazione, più spesso di competizione, con altre componenti

- 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

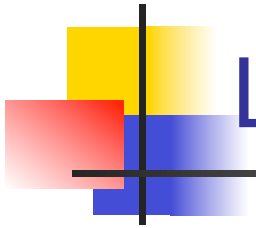
- **Questione**: come può l'opera combinata di un insieme di agenti produrre un comportamento che ogni singolo agente, considerato separatamente, non è in grado di fornire?
- La **nozione di agenzia** come superamento di posizioni di riduzionismo ingenuo difficilmente sostenibili (Minsky contesta chi 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**



Un esempio: il mondo dei blocchi

Il **mondo dei blocchetti** delle costruzioni per bambini

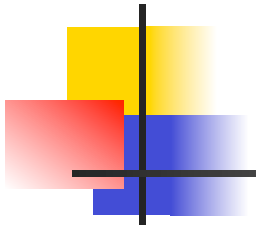
- Necessità di programmi (di livello superiore) che consentano al sistema/robot di pianificare le cose da fare (agente COSTRUTTORE)
- Necessità di programmi (di livello più basso) che consentano al robot di eseguire quanto pianificato (agente AGGIUNGERE, che, a sua volta, sfrutta gli agenti TROVARE, PRENDERE e METTERE)
- Necessità di programmi che consentano al robot di accertarsi che i piani siano stati effettivamente eseguiti



La coscienza nella società della mente

- In generale, osserva Minsky, siamo meno consapevoli (coscienti) di ciò che la nostra mente fa meglio
- E' soprattutto quando gli altri sistemi cominciano a fallire che facciamo intervenire quelle particolari agenzie che hanno a che fare con ciò che chiamiamo **coscienza**
- Di conseguenza siamo più consapevoli (coscienti) dei processi semplici che non funzionano bene che dei processi complicati che funzionano impeccabilmente

Per Minsky la coscienza ha più a che fare con le cose semplici/banali che con le cose complesse/fondamentali



Coscienza e memoria

- La coscienza è collegata con i nostri ricordi più recenti (ci sono dei limiti a quanto essa può dirci su se stessa)
- Normalmente si ritiene che la coscienza sia sapere ciò che accade nella nostra mente nell'istante attuale, ma

  la coscienza riguarda il passato, non il presente

- Ci sono agenzie che imparano a riconoscere gli eventi interni al cervello (esempio: agenzie che gestiscono i ricordi)

Alla **radice della coscienza** stanno gli agenti preposti all'**uso** e alla **modifica dei ricordi** più recenti

La dissoluzione del sé

- L'**identità personale**. Perché accettiamo questa immagine paradossale di un sé centrale dentro di sé? Perché in molti campi della vita pratica risulta **comodo**
- Ci sono molteplici ragioni per cui è utile considerarci come individui singoli (esempio: il senso di responsabilità)
- La **dissoluzione del sé**: dobbiamo renderci conto che ogni persona ha la sua identità, ma può avere nello stesso tempo convinzioni, disposizioni e piani diversi (l'immagine dell'agente unico è un grave impedimento per la scoperta di buone idee in ambito psicologico)

La non persistenza del sé

Che cosa intendiamo con parole come me, me stesso e io?

- Per quanto riguarda la coscienza, ci è quasi impossibile separare l'aspetto delle cose dal **significato** che hanno assunto per noi
- Non riusciamo a ricordare come ci apparivano le cose prima che imparassimo ad annettere loro significati nuovi (ad esempio, le parole prima e dopo aver imparato a leggere), come possiamo credere di poter ricordare come noi un tempo apparivamo a noi stessi?
- **La non persistenza del sé**: io corrisponde a quell'insieme di ricordi il cui significato muta solo lentamente

La non consapevolezza del sé

- Guidiamo il nostro corpo e la nostra mente senza sapere come funziona il nostro sé. Possiamo pensare senza sapere che cosa significhi pensare, possiamo avere idee senza essere in grado di spiegare che cosa siano le idee
- Nella mente di ogni persona normale sembrano esservi dei processi che chiamiamo coscienza. Spesso riteniamo che essi ci consentano di sapere che cosa accade nella nostra mente (**autoconsapevolezza**); in realtà, i nostri pensieri coscienti ci rivelano pochissimo di ciò che li genera. Essi inviano segni/segnali per pilotare il motore della nostra mente, controllando innumerevoli processi di cui non siamo mai molto consapevoli

La svalutazione della coscienza - 1

- I termini coscienza e autoconsapevolezza si riferiscono alla percezione della propria mente al lavoro
- Per rispondere a domande quali: “avevi coscienza di ..?” dobbiamo avere una registrazione dell'attività recente di alcuni agenti. In ogni cosa che diciamo e facciamo sono, però, implicate numerose altre attività. Se fossimo veramente autoconsapevoli, dovremmo sapere anche di tutte queste altre attività

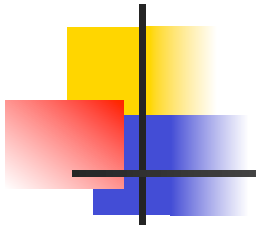
La svalutazione della coscienza - 2

- In ogni caso, il nostro **interesse primario** non è imparare a descrivere i nostri stati mentali quanto **realizzare cose pratiche** come fare progetti e portarli a compimento

- E' soprattutto quando i nostri sistemi falliscono che entra in gioco la coscienza: coscienza come capacità di riconoscere, affrontare e, se possibile, risolvere problemi

La scomparsa della coscienza

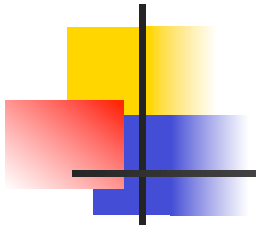
- Molti affermano che nessun calcolatore potrà mai essere senziente, cosciente, dotato di volontà propria o in qualche altro modo "consapevole" di se stesso. Ma che cosa ci rende tutti così sicuri di possedere noi queste meravigliose qualità?
- Se autoconsapevolezza significa sapere che cosa accade nella propria mente, nessuno può veramente sostenere che le persone siano in grado di "vedere dentro”
- **La scomparsa della coscienza**: è ingiustificata la convinzione che il modo in cui noi apprendiamo le cose che riguardano le persone, compresi noi stessi, sia fondamentalmente diverso dal modo in cui apprendiamo tutto il resto



Coscienza di sé e comunicazione

- Come una mente può **comunicare** con un'altra?
- La situazione è la stessa dentro la nostra mente: neppure noi stessi possiamo mai sapere con precisione che cosa noi intendiamo dire

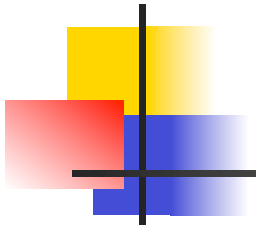
- Ciò che una cosa significa per noi dipende dal modo in cui l'abbiamo collegata alle altre cose che conosciamo (la nozione di **rete semantica**)



Agenti e sistemi multi-agente in IA

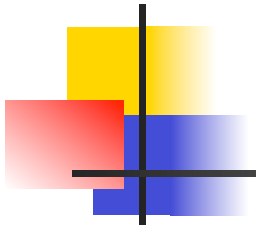
La teoria degli **agenti** / i **sistemi multi-agente** in IA possono essere visti come la controparte "applicativa" della società della mente di Minsky

- **Agente** (**artificiale**) **intelligente**: un agente intelligente è un sistema in grado di decidere cosa deve fare e di intraprendere le azioni necessarie a realizzare quanto deciso
- **Sistemi multi-agente**: sistemi costituiti da un insieme di agenti interagenti (interazione = scambio di messaggi tra agenti artificiali)



Agenti e agenzie

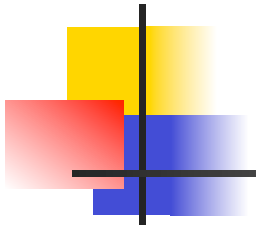
- Gli agenti sono **entità computazionali** in grado di interagire fra loro e con gli esseri umani
- Sono definiti **intelligenti** perché sono in grado di eseguire compiti e attività che richiedono dei processi di ragionamento, mostrando in tal modo un comportamento intelligente
- Un' **agenzia** è un insieme (una società) di agenti interagenti



La nozione di agente

Un agente è un sistema computazionale (programma software più eventuale supporto hardware) che:

- interagisce con l'ambiente circostante / reagisce agli stimoli di tale ambiente (**reattivo**)
- è in grado di prendere decisioni, e di conseguenza di agire, in modo autonomo, per raggiungere un obiettivo predefinito o negoziato (**proattivo**)
- è in grado di comunicare (coordinarsi, cooperare, negoziare) con altri agenti e/o con esseri umani (capacità di **interazione sociale**)



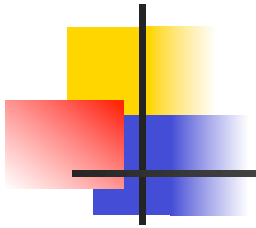
Tipologie di agente e loro proprietà

Tipi di agente:

- robot (ambiente fisico)
- agente software (ambiente computazionale)
- agente artificiale (ambiente virtuale)

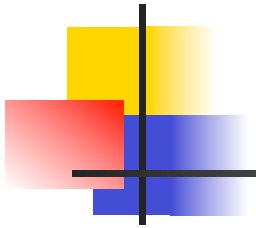
Proprietà avanzate:

- mobilità (capacità di muoversi nell'ambiente - mondo fisico o rete)
- apprendimento (capacità di acquisire nuove conoscenze o migliorare le proprie prestazioni sulla base dell'esperienza)



I diversi approcci

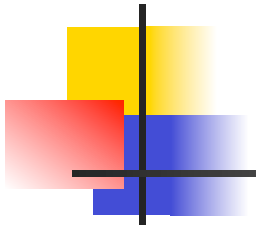
- Fino agli anni ’80, approcci di natura simbolico-dichiarativa caratterizzati da una **rappresentazione esplicita** e **simbolica della conoscenza** (di tipo logico – esempio: logiche epistemiche) che l’agente ha del mondo, che consente di sfruttare tecniche formali di ragionamento
- A partire dagli anni ‘90, comportamento intelligente di tipo reattivo: esso emerge dall’**interazione di comportamenti semplici** e non è supportato da una rappresentazione della conoscenza esplicita e simbolica
- Sistemi ibridi: combinazione di ragionamento formale e comportamento reattivo



Approcci simbolico-dichiarativi

Basati sul paradigma "ragionamento e pianificazione", richiedono che all'agente venga fornita una **descrizione formale**:

- dell'ambiente circostante (in termini di stati), che rappresenta la conoscenza che l'agente ha del mondo
- dell'obiettivo (uno stato), che rappresenta l'intenzione dell'agente
- delle azioni che è possibile effettuare nell'ambiente per raggiungere l'obiettivo, ciascuna con i relativi effetti e precondizioni (entrambi forniti in termini di stati dell'ambiente)

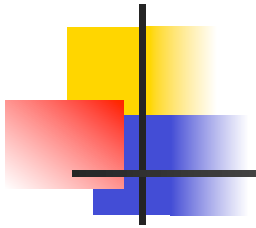


Critiche

1. L'approccio simbolico-deliberativo funziona bene solo nel caso di domini applicativi semplificati (**problemi giocattolo**)
2. Presta poca attenzione all'**ambiente** circostante, mentre

- il comportamento intelligente è indissolubilmente legato all'interazione (fisica) con esso
- il comportamento intelligente emerge non dalla manipolazione di rappresentazioni simboliche, ma dall'interazione di comportamenti più semplici

Soluzione: approcci comportamentismi (situati o reattivi)

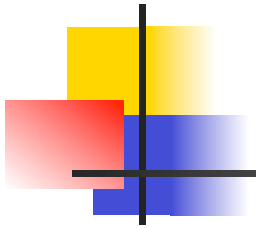


Gli agenti reattivi

L'agente reattivo sceglie le proprie azioni sulla base della percezione corrente. Tali azioni possono essere definite utilizzando regole oppure funzioni matematiche e probabilistiche che sfruttano conoscenza sub-simbolica

Esempio. La *Subsumption Architecture* di Brooks

Essa è costituita da un insieme di comportamenti strutturati su livelli gerarchici (**agenzie**): ogni comportamento è una funzione che riceve input percettivi e li mette in corrispondenza con delle azioni. Ogni funzione è specializzata nell'esecuzione di un particolare **compito elementare**

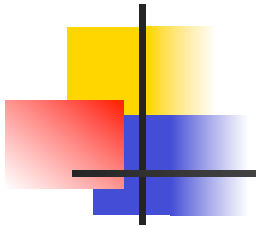


Architetture degli agenti

Agenti organizzati in architetture a livelli. Sulla base del tipo di interazione fra i livelli si possono distinguere

- **architetture orizzontali**, in cui i livelli operano in parallelo e concorrono a generare l'output (più semplici, ma richiedono spesso un coordinatore che garantisca la coerenza dell'elaborazione)
- **architetture verticali**, in cui ogni livello esegue un'elaborazione dell'input e poi passa il controllo al livello successivo, fino al livello che genera l'output

Un tipo particolare di approccio comportamentista/reattivo è rappresentato dai **modelli connessionisti**



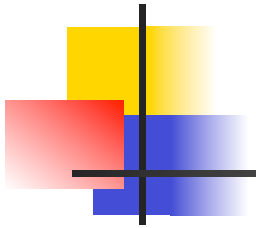
I modelli multi-agente

Un **modello multi-agente** può sia definire la struttura interna di un sistema (agente) sia modellare una società di agenti che svolgono funzioni di alto livello e ricoprono ruoli diversi

Ogni agente interagisce con l'ambiente circostante in modo autonomo sulla base delle proprie preferenze relativamente agli stati del mondo.

Problema: come realizzare la cooperazione tra gli agenti?

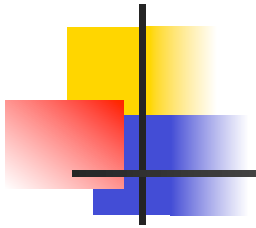
Risulta essenziale la **comunicazione** tra gli agenti del sistema. Soluzione: i piani dei singoli agenti, che modellano le possibili azioni, includono anche **azioni comunicative**



Alcune considerazioni

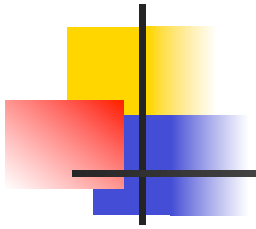
- Legame tra **intenzionalità** e capacità di creare degli **artefatti**: l'intenzionalità si manifesta nella sintesi dei programmi, ma non si trasferisce al programma sintetizzato (al programma in sé)

- **Intelligibilità** dei sistemi artificiali computazionali: non appena i compiti che il sistema deve eseguire diventano sufficientemente generali, il modello computazionale sottostante deve essere sufficientemente potente e, conseguentemente, non c'è modo di garantire che il sistema soddisfi le proprietà attese (teorema di Rice)



Potere computazionale e controllo

- C'è un **trade-off** tra potere espressivo/computazionale di un sistema e livello di controllo del suo comportamento: ogni aumento del primo si traduce in una riduzione del secondo
- Il problema della **responsabilità**: l'effettivo bilanciamento tra potere computazionale e controllabilità deve essere preso in considerazione ogni volta che dobbiamo stabilire la responsabilità delle azioni, e dei relativi effetti, intraprese da una macchina



Esseri umani e cicli di controllo

- L'interazione tra un sistema aperto e il suo ambiente (comunicazione asincrona) è spesso soggetta a vincoli temporali che non consentono di assicurare la presenza di un essere umano nei **cicli di controllo**
- I sistemi devono essere in grado di prendere delle **decisioni autonome** (decisioni che non prevedono alcuna esplicita autorizzazione da parte di un essere umano), per prevenire un incidente/guasto o per circoscriverne gli effetti

Un interessante appello

The scientists' call to ban autonomous lethal robots

- As computer scientists, engineers, AI experts, roboticists and professionals from related disciplines, we call for a ban on the development and deployment of weapon systems in which the decision to apply violent force is made **autonomously**
- We are concerned about the potential of robots to undermine **human responsibility** in decisions to use force, and to obscure accountability for the consequences
- We hold that fully autonomous robots that can trigger or direct weapons fire without a human effectively in the decision loop are similarly unacceptable

Interfacce cervello-computer (BCI)

Lo scenario cambia radicalmente se ci spostiamo nel campo delle **interfacce cervello-computer**

- Possibilità di influenzare gli stati mentali/di coscienza di una persona: sfruttare le interfacce cervello-computer come un canale per trasmettere direttamente al cervello dei segnali elettrici esterni
- L'effetto di tale stimolazione sul controllo della propria attività cerebrale da parte del soggetto è oggetto di studio e sperimentazione
- Questioni di natura etica: quali alterazioni della continuità degli stati cerebrali/mentali di un soggetto possono essere accettabili da un punto di vista etico?

Breve bibliografia - 1

- A. Goy, I. Torre, Agenti artificiali e agenti intelligenti: paradigmi, applicazioni e prospettive, Lexia, 03/04, Aracne, 2009, 297-313
- M. L. Minsky, The Society of Mind, Simon and Schuster, 1986
- A. Montanari, Alcune questioni di tecnoetica dal punto di vista di un informatico, Teoria XXVII/2, 2007, pp. 57-72
- A. Montanari, Riduzionismo e non in intelligenza artificiale, in: La differenza umana. Riduzionismo e antiumanesimo, a cura di L. Grion, Anthropologica, Annuario di Studi Filosofici 2009, pp. 113-128
- A. Montanari, Scienza e immortalità terrena, in: La sfida postumanista. Colloqui sul significato della tecnica, a cura di L. Grion, Il Mulino, 2012, pp. 101-125
- S. Russel e P. Norvig, Artificial intelligence: a modern approach, Vol 1 e 2, Prentice Hall (trad. it. Intelligenza Artificiale: un approccio moderno, Pearson, 2005)

Breve bibliografia - 2

- J. R. Searle in Minds, Brains, and Programs, Behavioral and Brain Sciences, volume 3, Cambridge University Press, Cambridge 1980
- A. M. Turing, Computing Machinery and Intelligence, Mind 59, 1950, pp. 433-460; tr. it. Macchine calcolatrici e intelligenza, in V. Somenzi - R. Cordeschi (edd.), La filosofia degli automi. Origini della intelligenza artificiale, Bollati Boringhieri, Torino 1994; oppure in A. M. Turing, Intelligenza meccanica, Bollati Boringhieri, Torino 1994, Collected Works of A.M. Turing: Mechanical Intelligence (1992); tr. it. di G. Lolli e N. Dazzi
- N. Wiener, Cybernetics or Control and Communication in the Animal and the Machine (second edition), The M.I.T. Press, Cambridge, MA 1962
- M. Wooldridge, An Introduction to Multi Agent Systems, John Wiley & Sons, 2002