Dal Test di Turing alla disputa su IA debole e IA forte

Edmondo Trentin (DIISM)

La macchina di Turing (MdT)

Alan Turing (1912-1954)

MdT (A. Turing, 1936), descrizione informale

Una MdT è composta da:
(1) un **nastro** monodimensionale infinito suddiviso in **celle** atomiche ognuna delle quali, in un dato istante, contiene un singolo simbolo di un dato alfabeto A oppure è vuota (simbolo speciale '#');
(2) un **orologio** (*clock*) tempo-discreto che scandisce il funzionamento della macchina a istanti temporali successivi $t_0, t_1, t_2, \ldots$;
(3) una **testina** di lettura/scrittura che in un dato istante si trova in corrispondenza di una ben precisa cella, ne legge il contenuto e ad esso sovrascrive un simbolo di A, per poi compiere un passo a sinistra (L) o a destra (R) sul nastro;
(4) un certo numero di **stati interni** in cui la MdT può essere. In un dato istante la MdT è in uno e uno solo di questi stati;
(5) uno specifico insieme di regole (dette **regole di transizione**) che ne determinano il comportamento in funzione dello stato interno e dei simboli letti sul nastro.

MdT - rappresentazioni grafiche

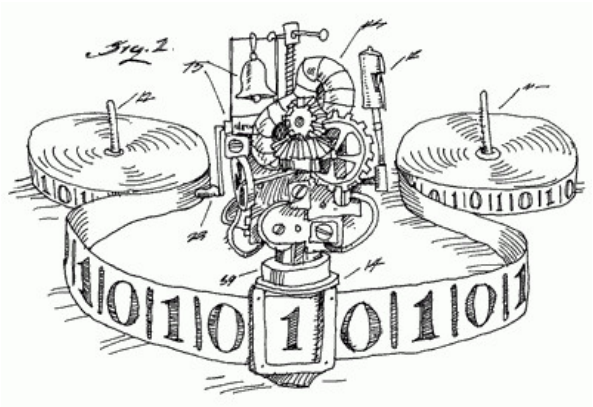


Figura: MdT: una fantasticheria

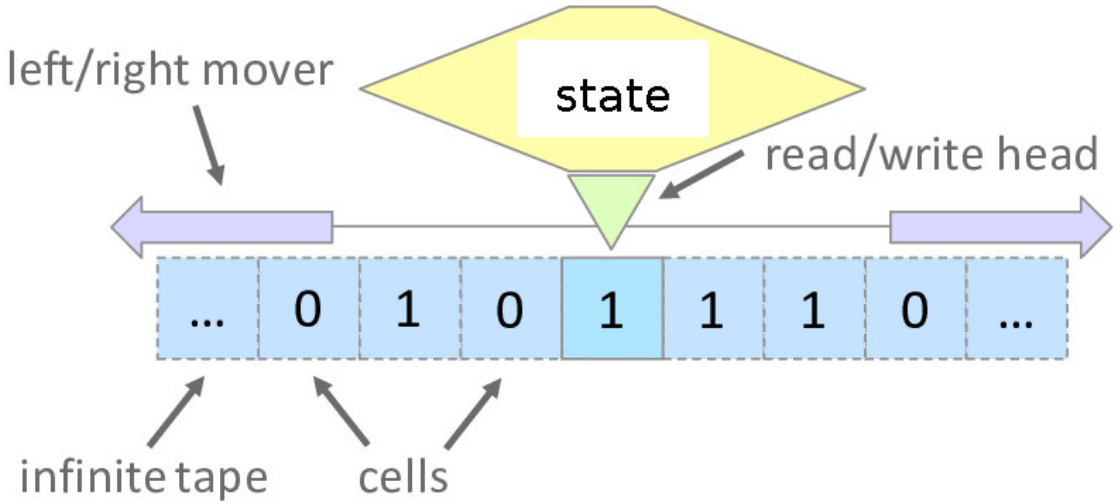


Figura: MdT: una rappresentazione più formale

MdT - Definizione "formale"

Una specifica MdT $\mathcal{M}$ è una quintupla $\mathcal{M} = (A, S, q_0, q_k, T)$ dove

- A è un **alfabeto** finito e discreto, $A = \{a_1, \ldots, a_n\}$ (ad es. $A = \{0, 1\}$, oppure $A = \{a, b, c, \ldots, z\}$, ecc.)
- S è un insieme finito e discreto di **stati** interni $S = \{q_0, q_1, \ldots, q_k\}$
- q_0 è lo **stato iniziale**, $q_0 \in S$
- q_k è lo **stato finale**, $q_k \in S$
- T è un isnieme di **regole di transizione**: $T = \{a_i q_j \rightarrow a_\ell q_m L,\ \ a_u q_v \rightarrow a_w q_z R,\ \ \ldots\}$ (dove a_i, a_ℓ, a_u, a_w sono in A e q_j, q_m, q_v, q_z sono in S).

Ad esempio, $a_i q_j \rightarrow a_\ell q_m L$ significa "se $\mathcal{M}$ è nello stato interno q_j e la testina legge il simbolo a_i nella cella su cui è posizionata, allora la testina scrive nella stessa cella il simbolo a_ℓ, la MdT entra nello stato q_m, e la testina fa un passo a sinistra sul nastro ('R' avrebbe denotato invece un passo a destra).

MdT: funzionamento

1. Una stringa di simboli di A (una **frase**) è posta sul nastro (*input*)
2. la testina è posizionata sulla cella corrispondente all'ultimo carattere a destra dell'input
3. la MdT è posta nello stato iniziale q_0
4. l'orologio della MdT viene fatto partire
5. a ogni rintocco dell'orologio (*tic*, *tac*, ...):
   - 5.1 la testina legge il simbolo a_i contenuto nella cella su cui essa è posizionata
   - 5.2 la MdT cerca se c'è una regola di transizione corrispondente al simbolo a_i e allo stato interno attuale q_j
   - 5.3 se tale regola c'è, ad es. $a_i q_j \rightarrow a_\ell q_m L$, essa viene applicata
   - 5.4 se tale regola non c'è, la MdT si ferma (restando in q_j)

MdT: funzionamento

La MdT può procedere all'infinito senza mai fermarsi (**loop infinito**) oppure si ferma non trovando più regole da applicare.

- Se si ferma **e** lo stato in cui si trova è lo stato finale q_k allora la stringa di caratteri lasciata sul nastro prende il nome di **output** della MdT a fronte dell'input inizialmente fornitole.
- Se si ferma **e** lo stato in cui si trova non è lo stato finale q_k, oppure se va in loop infinito, si dice che l'output della MdT a fronte dell'input inizialmente fornitole è **indefinito**.

Ogni MdT realizza dunque una specifica **trasformazione** da frasi di input a frasi di output sull'alfabeto A.

La danza della testina sul nastro, scandita dai rintocchi dell'orologio, è un **processo** dinamico (come un software in esecuzione è un *processo*, e come la mente nell'essere umano è il processo realizzato dal cervello) che prende il nome di **computazione** (o **calcolo**).

Simulatore di MdT

Clicca qui per lanciare il simulatore

(serve una connessione internt attiva)

Esempi di MdT per i connettivi della logica classica (calcolo proposizionale) *NOT* (negazione), *AND* (congiunzione) e *OR* (disgiunzione):

A	$\neg A$
0	1
1	0

NOT

A	B	$A \wedge B$
0	0	0
1	0	0
0	1	0
1	1	1

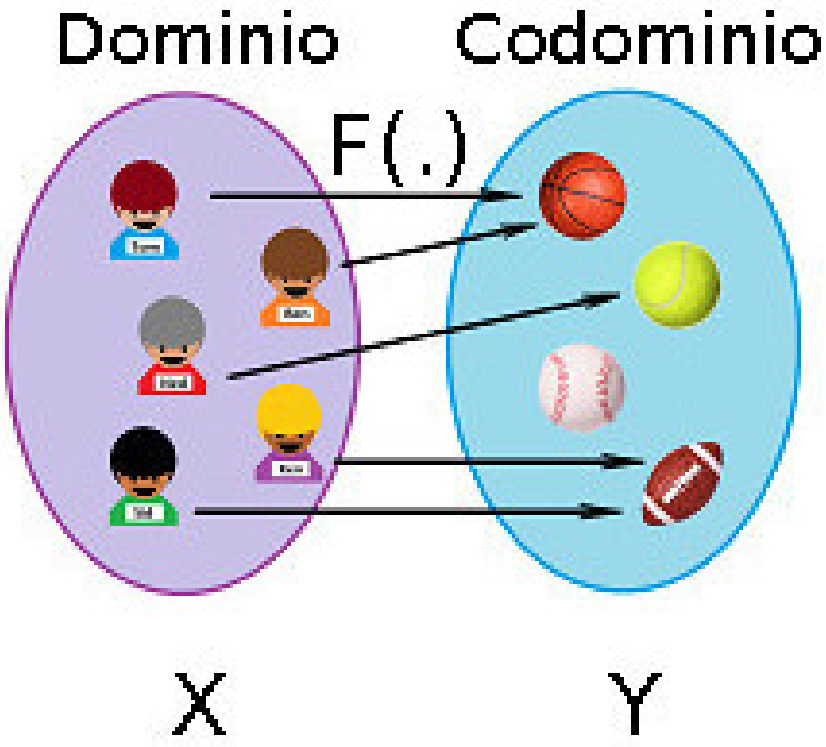
AND

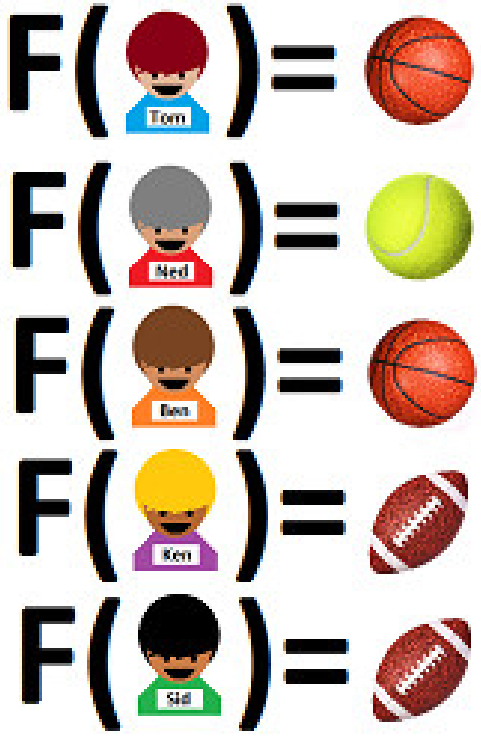
A	B	$A \vee B$
0	0	0
1	0	1
0	1	1
1	1	1

OR

Esercizio: divertiti a scrivere una MdT che realizzi il connettivo OR.

Le funzioni, queste sconosciute





MdT: funzione calcolata

- Consideriamo una **generica MdT** $\mathcal{M} = (A, S, q_0, q_k, T)$
- Scriviamo A^* per rappresentare l'**insieme di tutte le possibili stringhe** (frasi) sull'alfabeto A
- Sia $x_1x_2x_3 \ldots x_L$ un'**arbitraria stringa** di caratteri di A, ovvero $x_1 \in A$, $x_2 \in A$, ..., $x_L \in A$ (e dunque $x_1x_2x_3 \ldots x_L \in A^*$).

Ora diamo $x_1x_2x_3 \ldots x_L$ in input a $\mathcal{M}$ e avviamo la computazioe. Se a un certo istante $\mathcal{M}$ si ferma trovandosi nello stato finale q_k, sul nastro ci sarà una corrispondente stringa di output $y_1y_2 \ldots y_N$. La funzione $f_{\mathcal{M}} : A^* \to A^* \cup \{\perp\}$ definita come

$$f_{\mathcal{M}}(x_1x_2x_3 \ldots x_L) = \begin{cases} y_1y_2 \ldots y_N & \text{se M si ferma in stato } q_k \\ \perp & \text{altrimenti } (\textit{indefinito}) \end{cases}$$

si dice **funzione calcolata** dalla MdT $\mathcal{M}$. Se (A, S, q_0, q_k, T) è la **sintassi** di $\mathcal{M}$, $f_{\mathcal{M}}$ ne è la **semantica**.

Funzioni calcolabili

- Ogni MdT calcola una e una ben precisa funzione
- Se una certa funzione $g(.)$ è la funzione calcolata da una MdT $\mathcal{M}$, ovvero se $g(.) = f_{\mathcal{M}}(.)$, allora esistono infinite MdT sintatticamente diverse tra di loro ma semanticamente **equivalenti** che calcolano $g(.)$ [*Esercizio: dimostralo per via costruttiva!*]
- La classe delle **funzioni calcolabili** è composta da tutte le funzioni per le quali esista una MdT che le calcoli
- *Teorema*: esistono infinite funzioni **non calcolabili**!
- La calcolabilità (e la non calcolabilità) sono proprietà intrinseche delle funzioni, immutabili da sempre e per sempre!
- La calcolabilità *alla Turing* è una proprietà **universale**, non dipende cioè dal paradigma di calcolo (MdT, calcolatore digitale, ecc.).

Proprietà universale: MdT vs. calcolatore digitale

- Abbiamo visto che i connettivi logici *NOT*, *AND* e *OR* sono calcolabili (alla Turing);
- è facile intuire che se la MdT $\mathcal{M}_h$ calcola la funzione $h(\cdot)$ e la MdT $\mathcal{M}_g$ calcola la funzione $g(\cdot)$ allora la funzione composta $f(\cdot) = g(h(\cdot))$ è calcolabile e per calcolarla "basta" prima far girare $\mathcal{M}_h$ e, sull'output di questa, eseguire poi la computazione di $\mathcal{M}_g$ (in pratica si può definire una nuova MdT $\mathcal{M}_f$ che racchiuda in sé tanto $\mathcal{M}_h$ quanto $\mathcal{M}_g$, usando lo stato finale di $\mathcal{M}_h$ come stato iniziale di $\mathcal{M}_g$);
- infine, qualsiasi funzione logica può essere scritta come composizione di *NOT*, *AND* e *OR* (cosiddette *forme canoniche*) e può dunque essere calcolata da una MdT.

Tutto questo forma la base per dimostrare che qualsiasi funzione realizzata da un computer digitale può essere calcolata da una MdT (visto che le *reti logiche* che costituiscono la circuiteria dei microprocessori altro non fanno che calcolare funzioni logiche in cascata).

La madre di tutte le domande:

L'INTELLIGENZA È CALCOLABILE?

Turing capisce subito che per rispondere formalmente (tramite MdT) sarebbe necessario

1. definire formalmente l'intelligenza come una funzione $\mathcal{I}(.)$ su un certo alfabeto $A_{\mathcal{I}}$, cioè $\mathcal{I} : A^*_{\mathcal{I}} \rightarrow A^*_{\mathcal{I}}$
2. verificare formalmente se sia o meno calcolabile, ovvero
   - 2.1 scrivere una MdT $\mathcal{M}_{\mathcal{I}}$ che la calcoli, per cui cioè $f_{\mathcal{M}_{\mathcal{I}}}(.) = \mathcal{I}(.)$, oppure dimostrare che almeno in via di principio la si potrebbe scrivere; oppure
   - 2.2 dimostrare formalmente che non è possibile scriverla;

e che tutto questo non è realisticamente fattibile. Turing trova però un modo alternativo per cercare una risposta significativa alla madre di tutte le domande.

The Imitation Game

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

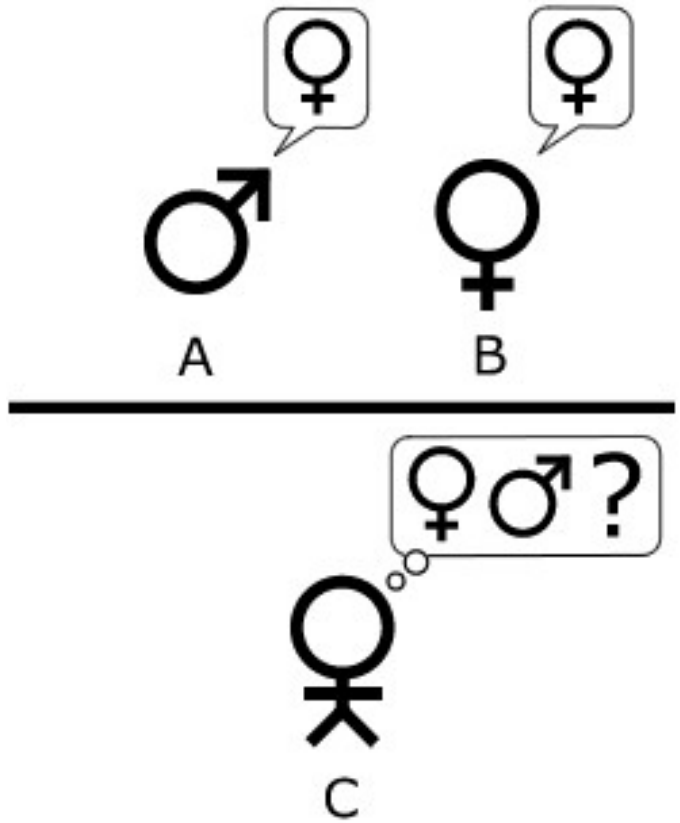
BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning

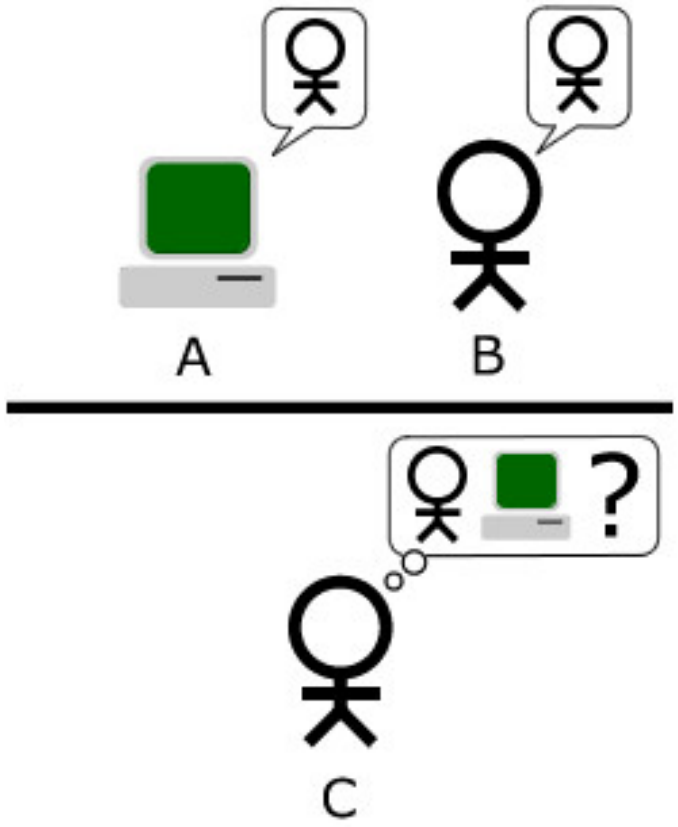
The Imitation Game

C comunica via telescrivente con A e B e deve indovinare chi è la femmina e chi il maschio. A e B cercano di ingannarlo con le loro risposte:



The Imitation Game

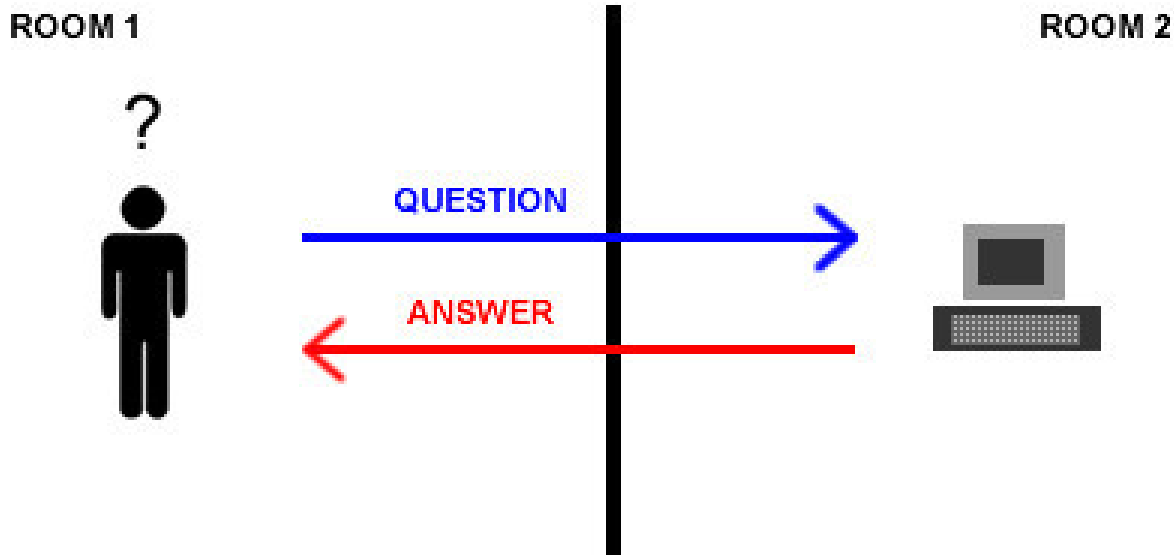
Cosa succede se, all'insaputa di C, una MdT opportunamente programmata viene poi messa al posto di A o B?



Test di Turing: la MdT si può dire **intelligente** se riesce a ingannare C (che la scambia per un essere umano).

Test di Turing

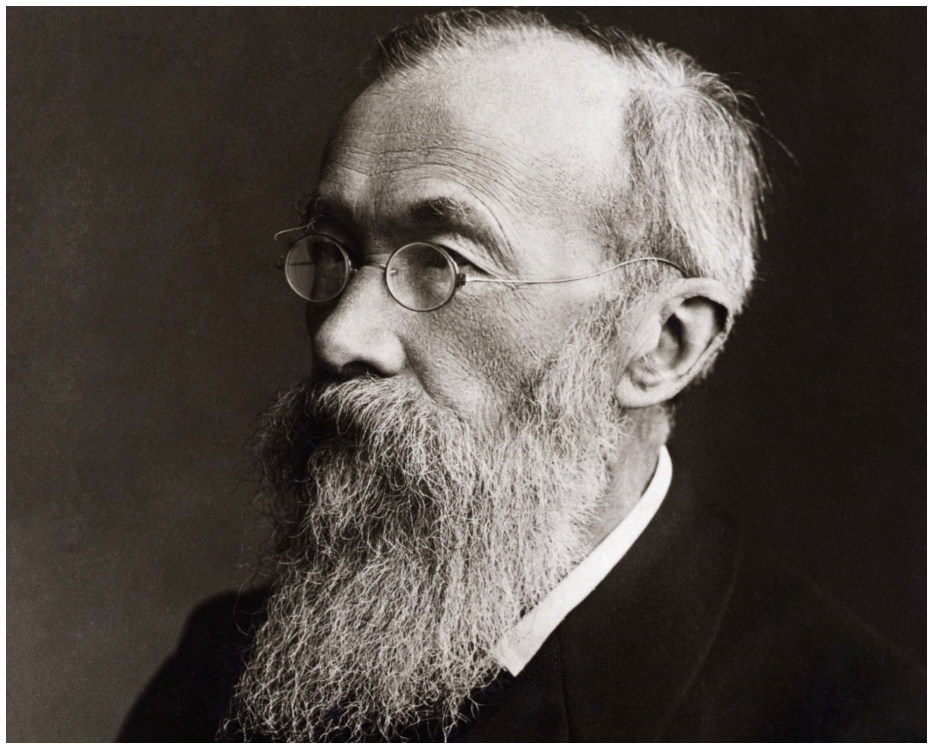
Nella versione odierna basta capire se si ha di fronte una macchina o no:



Più precisamente, **ripetendo l'esperimento** più volte (variando anche lo sperimentatore) si conta la frequenza con cui la macchina riesce a ingannare lo sperimentatore: se essa è **statisticamente significativa** il test è superato e la macchina si può dire **intelligente**.

Test di Turing

È un criterio di tipo **sperimentalista/behaviorista**.



Clicca qui per parlare con Eugene Goostman

(il tredicenne artificiale che ha "passato il test di Turing")

Nel 1990 il test di Turing si trova al centro della disputa tra *supporter* dell'**IA debole** (John Searle) e dell'**IA forte** (P. Churchland & P. Churchland)

← supporter dell'IA forte supporter dell'IA debole →

John Searle, paladino dell'IA debole

J. Searle, *Is the Brain's Mind a Computer Program?* (1990): **più una macchina pensare?**
Naturalmente sì, si consideri il cervello. Ma **limitiamoci alle macchine "alla Turing"** (calcolatori digitali). Sostiene Searle:

- **L'IA forte ritiene di sì**, cioè che le macchine manipolatrici di simboli possano pensare
- Per verificare se una macchina sia pensante, **l'IA forte fa leva su un test: il test di Turing**
- **Ma la manipolazione di simboli è solo sintassi**, senza semantica: anche simulando alla perfezione il cervello su una MdT, la MdT continuerebbe a non pensare.

Searle: l'esperimento mentale della "stanza cinese"

- Il tizio nella stanza non sa nulla di cinese, riceve domande in cinese dall'esterno e risponde in cinese mettendo in fila gli ideogrammi presenti nei cesti in base alle regole di associazione domanda-risposta elencate nel librone
- Il tizio non ha alcuna comprensione (semantica) del cinese, né ce l'hanno la stanza o gli oggetti in essa contenuti.
- La stanza supera però il test di Turing per il cinese.

Searle: pseudo-sillogismo contra IA forte

- **Premessa 1**: tutti i programmi per computer (MdT) sono sintattici
- **Premessa 2**: tutte le menti hanno contenuti mentali (semantica)
- **Premessa 3**: la sintassi di per sé non è necessaria né sufficiente per la semantica

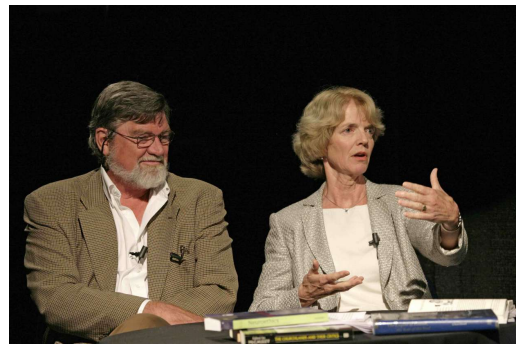
---

Conclusione: i programmi per computer non sono necessari né sufficienti per le menti

Note:

- le premesse 1 e 2 sono "autoevidenti"
- la premessa 3 consegue dalla *stanza cinese*

Paul e Patricia Chruchland, paladini dell'IA forte



P. Chruchland e P. Chruchland, *Could a Machine Think?* (1990): **può una macchina pensare?** Risposta: potenzialmente sì.

- L'argomento di J. Searle è un abbaglio
- Se anche esso fosse corretto per un solo tizio chiuso nella stanza cinese, potrebbe smettere di valere per una messe di operatori in un sistema grande come l'universo e sufficientemente veloce
- se anche esso fosse corretto per una MdT (macchina manipolatrice di simboli), potrebbe smettere di valere per una macchina **parallela e distribuita** architettata in modo analogo al cervello.

Churchland: l'esperimento mentale della "stanza luminosa"

- A.D. 1865: James Clerk Maxwell ha da poco formulato una teoria sulla natura della luce come radiazione elettromagnetica
- un tizio è chiuso in una stanza completamente buia
- il tizio alza e abbassa ostinatamente e il più velocemente possibile un grosso magnete che tiene in mano
- Se la teoria di Maxwell fosse giusta, dovrebbe svilupparsi luce; ma nella stanza resta un buio pesto.

Churchland: pseudo-sillogismo searliano contra Maxwell

- **Premessa 1**:L'elettricità e il magnetismo sono forze
- **Premessa 2**: La proprietà essenziale della luce è la luminanza
- **Premessa 3**: le forze di per sé non sono necessarie né sufficienti per la luminanza

---

Conclusione: elettricità e magnetismo non sono necessari né sufficienti per la luce

Note:

- le premesse 1 e 2 sono autoevidenti
- la premessa 3 consegue dalla *stanza luminosa*

Secondo i Churchland **lo schema di ragionamento pseudo-sillogistico di Searle non sarebbe dunque altro che un paralogismo.**

Punti deboli dei Churchland

- L'esperimento mentale della stanza luminosa **non è affatto equivalente** a quello della stanza cinese. Esso infatti consiste in uno sterile esercizio fine a se stesso che **non soddisfa alcun criterio minimale di "successo" o "insuccesso"** (non è insomma scientifico in quanto non falsificabile), mentre il punto chiave della stanza cinese è proprio che essa soddisfa un siffatto criterio, cioè l'avere superato il test di Turing.
- L'idea che aumentando velocità di esecuzione, parallelismo, e dimensioni della macchina si possano di per sè superare i limiti evidenziati da una versione estesa, parallela e velocizzata della stanza cinese (la **palestra cinese** proposta da Searle in replica alle critiche dei Churchland) è ingenuo e destituito di fondamento scientifico.

Punti debolissimi di Searle

- Searle confonde platealmente i **programmi** (le "regole di transizione" della MdT) con i **processi** (che sono programmi **in esecuzione** da parte della macchina)
- assume (arbitrariamente) che l'IA forte richieda come **condizione necessaria e sufficiente il superamento del test di Turing**, cosa non vera in generale
- Il *design* della stanza cinese presenta forti criticità:
  - anche solo in via teorica, come è possibile scrivere un **librone esaustivo e univoco di regole** di associazione linguistica domanda-risposta che risulti sempre credibile agli osservatori esterni?
  - Come tenere conto dell'evolversi cronologico del dialogo con l'esterno? (Alla stanza cinese mancano **memoria** e **adattività**)

Infine, tanto Searle quanto i Churchland si concentrano interamente sull'intelligenza intesa come **"mente"** (che, non avendo una definizione, ammette ogni argomento e il suo contrario) e non su quella emergente dai **comportamenti** (ottica behaviorista).

Concludendo ...

(Recondite armonie/d'intelligenze diverse)

Esercizio: guardare il film *The Imitation Game* di Morten Tyldum (USA, 2014)