Autoreferenza e diagonalizzazione

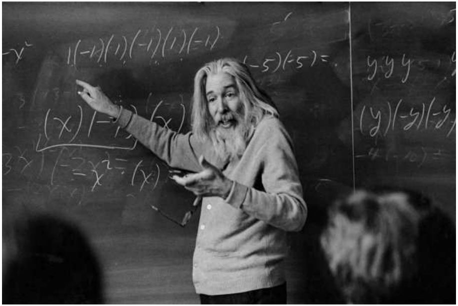


Figure: "C'è qualcosa che vorrei dirvi prima di cominciare a parlare", Raymond Smullyan

Autoreferenza e diagonalizzazione

- Alla base di alcuni tra i risultati, matematicamente e filosoficamente forse più importanti della logica contemporanea, che creano un ponte con altre aree della conoscenza come la filosofia del linguaggio o la filosofia della mente e le neuroscienze, vi è un medesimo meccanismo di autoriferimento.
- Secondo Douglas Hofstadter (*Gödel, Escher, Bach*), lo stesso fenomeno dell'ermergere della coscienza individuale e la nozione dell'io nascono da uno "strano anello" all'interno del nostro cervello come il risultato di un processo autoreferenziale.

Autoreferenza e diagonalizzazione

Figure: "A mirror mirroring a mirror." - Douglas R. Hofstadter, *I Am a Strange Loop*. Il cervello rispecchia il mondo esterno e gradualmente inizia a rispecchiare il suo stesso rispecchiamento del mondo e il suo rispecchiamento di quel rispecchiamento, ecc. Questa recursione infinita, afferma Hofstadter, è ciò che porta alla formazione della coscienza.

Autoreferenza e diagonalizzazione

> *Mi è stato detto a volte: "questa storia dell'autoreferenza eccetera è molto divertente e godibile, ma pensi realmente che vi sia qualcosa di serio in essa? ". Io credo veramente di sì. Penso che alla fine risulterà che essa è il cuore dell'IA e che costituisce il punto focale su cui dovranno concentrarsi tutti i tentativi di chiarire in che modo funziona la mente umana.*

...una figura che a livello artistico l'autore ritrova nelle modulazioni del *Canon circularis per tonos*, dall'*Offerta musicale* di Johann Sebastian Bach, che alla fine riconducono alla identica tonalità di partenza, senza che l'ascoltatore se ne sia reso conto, o nella figura "impossibile" della *Cascata* di Maurits Cornelis Escher:

Autoreferenza e diagonalizzazione

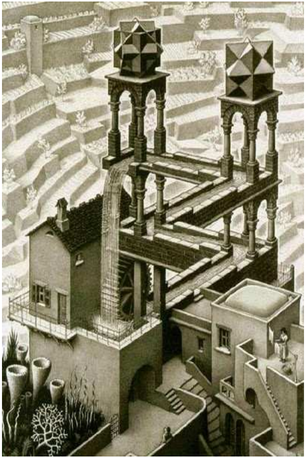


Figure: La cascata di Escher

Autoreferenza e diagonalizzazione

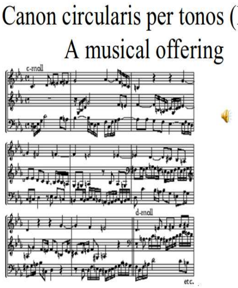


Figure: *La voce più alta espone il tema in otto battute, ripetendolo sei volte modulando in modo tale da terminare ogni volta un tono sopra, fino a tornare alla tonalità di partenza, secondo la successione do min, re min, mi min, fa diesis min, sol diesis min, si bemolle min, do min.*(vedi https://www.youtube.com/watch?v=A41CITk85jk)

Autoreferenza e diagonalizzazione

Sono detti *autoreferenti* gli enunciati che in qualche modo si riferiscono a sé stessi. Gli enunciati autoreferenti sono tipicamente generati da una procedura detta 'diagonalizzazione'. Verranno qui di seguito considerate alcune conseguenze ed applicazioni particolarmente significative dal punto di vista filosofico:

1. la nozione di infinito,
2. l'analisi dei paradossi,
3. il concetto di verità nel suo rapporto con il concetto di dimostrabilità,
4. il dibattito sul computazionalismo, sul meccanicismo e sui limiti dei computers.

Autoreferenza e diagonalizzazione

Uno dei motivi di interesse verso questo genere di enunciati è legato ai paradossi. Il più celebre paradosso basato sull'autoriferimento è senz'altro quello del mentitore. Ci chiediamo se la seguente frase è vera o no:

io mento

Se ne conclude tuttavia che se io mento, allora non mento e viceversa. In questa forma, il paradosso risale ad Eubulide di Mileto (IV a. C.), ma il primo a scoprirlo pare sia stato il cretese Epimenide di Cnosso (VI a.C.). Un'altra versione è:

Questa affermazione è falsa

Chiaramente, se la frase è vera, allora è falsa, e viceversa.

Autoreferenza e diagonalizzazione

San Paolo cita Epimendide nell'epistola a Tito:

> *Uno di loro, proprio un loro profeta, disse che tutti i Cretesi sono sempre mentitori*

Si noti che tuttavia, così formulata, l'affermazione non conduce a nessuna contraddizione (essa non può essere vera, ma può essere falsa). Altre versioni sono più complicate, come:

> *La frase di sotto è falsa.*
> *La frase di sopra è vera.*

ed altre innumerevoli si trovano nella letteratura, da Cervantes a Lewis Carroll.

Autoreferenza e diagonalizzazione

Willard van Orman Quine formulò il seguente paradosso, *indirettamente* autoreferenziale, sulla falsariga del "mentitore" (in una frase indirettamente autoreferenziale si rimpiazza ad esempio "questa frase" con un'espressione che si riferisce ad essa, ma non usa indicali, come il pronome "questo"):

> *"è falsa se preceduta dalla sua citazione" è falsa se preceduta dalla sua citazione.*

Questa versione del mentitore ci sarà utile per quello che diremo in seguito circa i programmi che replicano il proprio codice e le analogie con la riproduzione del DNA.

Autoreferenza e diagonalizzazione

Un meccanismo per generare frasi autoreferenti è quello della *diagonalizzazione*: con ciò, secondo la definizione di Smullyan, si intende il rimpiazzamento di tutte le occorrenze di una data variabile in una espressione, con citazioni della medesima espressione. In taluni casi questo processo produce frasi autoreferenti. Ad esempio, la diagonalizzazione di:

(A) John is reading the diagonalization of x

secondo la definizione di Smullyan è la seguente:

(B) John is reading the diagonalization of "John is reading the diagonalization of x"

La (B) afferma che John sta leggendo la diagonalizzazione di (A), cioè (B) stessa!Questa frase è pertanto autoreferente.

Autoreferenza e diagonalizzazione

- Meccanismi simili stanno alla base, non solo di molti paradossi, ma anche dei più importanti risultati della logica contemporanea, scoperti da logici come Gödel, Tarski, Turing e Kleene.
- Sottilineiamo però, che in alcuni importanti risultati, a fianco del procedimento di *auto-applicazione*, opera opera un meccanismo di *alterazione* del simbolo o della formula ottenuta con esso.
- La diagonalizzazione, come metodo di dimostrazione per *reductio absurdum* è legata in primo luogo al nome di Georg Cantor; negli esempi che vedremo i risultati scaturiscono assieme dall'azione congiunta di *auto-applicazione* e *negazione* (o alterazione).

La prova diagonale di Cantor

Autoreferenza e diagonalizzazione

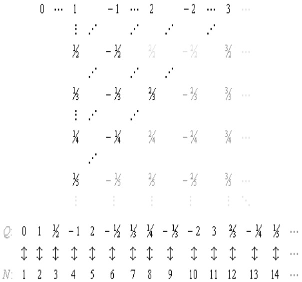


Figure: Cantor: i numeri razionali sono tanti quanti i numeri naturali.

Autoreferenza e diagonalizzazione

- Georg Cantor (1845-1918) è stato, assieme a Dedekind, il fondatore della teoria degli insiemi, il cui primo contributo è stato il chiarimento della nozione di infinito matematico.
- Per secoli, sia in ambito filosofico che in ambito matematico si era diffidato del concetto di infinito "attuale" o "completo", accettando l'infinito solo in senso "potenziale", cioè come inesauribile processo di accrescimento.
- Viceversa, nella teoria di Cantor e Dedekind, non solo l'infinito è ammesso in senso attuale, ma esistono addirittura molti infiniti, di diversa "grandezza", il più piccolo dei quali, che chiameremo *infinito contabile*, è l'insieme degli interi non negativi $\mathbb{N}$.

Autoreferenza e diagonalizzazione

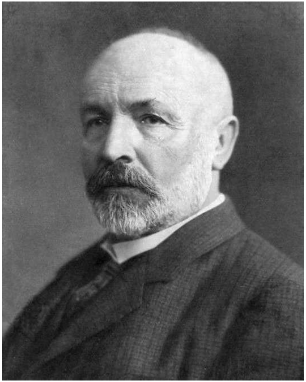


Figure: Georg Cantor (1845-1918)

Autoreferenza e diagonalizzazione

In passato veniva evidenziata la paradossalità della nozione di infinito, che ne sconsigliava l'uso. Ad esempio un insieme infinito può essere posto in corrispondenza biunivoca con una sua parte propria. Ecco come ad esempio si esprime Galileo:

> *Salviati...non vi è numero alcuno che non sia radice di qualche quadrato; e stante questo, converrà dire che i numeri quadrati siano quanti tutti i numeri.*

Questo aspetto paradossale diviene con Richard Dedekind (1831-1916) la definizione stessa di "insieme infinito":

> *si dice che un insieme è Dedekind-infinito se esiste una corrispondenza biunivoca tra esso ed un suo sottoinsieme proprio.*

Autoreferenza e diagonalizzazione

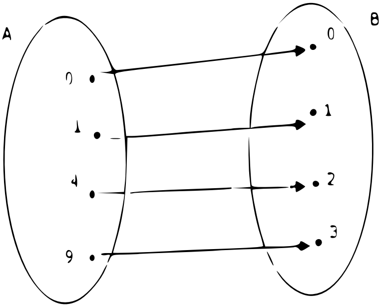


Figure: Corrispondenza biunivoca fra numeri e i loro quadrati

Corrispondenza biunivoca o biiezione: corrispondenza tra due insiemi (o classi) tale che a ogni elemento del primo è associato uno e un solo elemento del secondo e viceversa. È quindi una corrispondenza univoca, iniettiva e suriettiva.

Autoreferenza e diagonalizzazione

- In realtà si dimostra facimente, sempre usando la nozione di corrispondenza biunivoca, che anche l'insieme dei numeri interi positivi e negativi $\mathbb{Z}$ e l'insieme dei numeri razionali $\mathbb{Q}$ sono contabilmente infiniti, cioè possono essere posti in corrispondenza biunivoca con $\mathbb{N}$.
- Era dunque legittimo, prima di Cantor, sospettare che anche l'insieme dei numeri reali (razionali ed irrazionali) $\mathbb{R}$ fosse contabile. Che in realtà l'infinito fosse unico era del resto la visione tradizionale.
- Ebbene la *prova diagonale* di Cantor confuta invece questa visione, dimostrando che i numeri reali sono più dei numeri interi positivi, cioè a dire, questo insieme non è contabile. Oggi sappiamo che la sua cardinalità è $2^{\aleph_0}$.

Autoreferenza e diagonalizzazione

$$
\begin{array}{cccccc}
 & 0, & 1, & 2, & \ldots, & n, \quad \ldots \\
X_0 & \boxed{X_0(0)} & X_0(1) & X_0(2) & & X_0(n) \\
X_1 & X_1(0) & \boxed{X_1(1)} & X_1(2) & & X_1(n) \\
X_2 & X_2(0) & X_2(1) & \boxed{X_2(2)} & & X_2(n) \\
\vdots & & & & \ddots & \\
X_n & X_n(0) & X_n(1) & X_n(2) & & \boxed{X_n(n)} \\
\vdots & & & & & \quad \ddots
\end{array}
$$

Figure: Nella matrice è evidenziata la diagonale

Supponi per assurdo che vi sia una corrispondenza biunivoca fra reali nell'intervallo $[0, 1]$ e naturali. Ad esempio, nella matrice in figura sia $X_i = 0, X_i(0), X_i(1), X_i(2)...$ l'i-esimo reale, dove con $X_i(n)$ indichiamo il suo n-esimo decimale.

Autoreferenza e diagonalizzazione

Adesso definisci un nuovo reale:

$$X^* = 0, X_0^*, X_1^*, X_2^* \dots$$

dove $X_i^* = X_i(i) + 1$. È evidente che questo reale non sta nella lista precedente, giacché differisce da tutti quelli che ci stanno: per ogni i, differisce dall'i-esimo reale, proprio per il decimale $X_i(i)$. Tenendo presente la matrice della figura e la sua diagonale, cercheremo di far vedere come molti importanti risultati della logica che chiamano in causa l'autoriferimento si conformino a questo schema.

Autoreferenza e diagonalizzazione

Possiamo innanzitutto far discendere da questo risultato generale altri paradossi, segnatamente:

- Il paradosso di Cantor,
- Il paradosso (o antinomia) di Russell

Osservando la precedente matrice, possiamo dimostrare anche il seguente:

Teorema. Non esiste una corrispondenza biunivoca fra l'insieme $\mathbb{N}$ dei numeri naturali e l'insieme $\mathcal{P}(\mathbb{N})$ di tutti i suoi sottinsiemi.

Autoreferenza e diagonalizzazione

Infatti,

- supponiamo per ipotesi che una tale corrispondenza vi sia e sia dunque $X_0, X_1, X_2 \ldots$ una enumerazione di tali sottinsiemi e interpretiamo ora la scrittura $X_n(a)$ come $a \in X_n$.
- Definiamo X^* come l'insieme dei numeri n tali che $n \notin X_n$. Ma questo insieme comparirà nella enumerazione, ad esempio sarà un dato X_e.
- È facile vedere che allora $e \in X_e$ se e solo se $e \notin X_e$.

Con un pizzico di astrazione in più questo risultato può essere esteso ad ogni insieme: la "cardinalità" di un insieme è strettamente minore di quella dell'insieme delle sue parti.

Autoreferenza e diagonalizzazione

Teorema di Cantor. Non esiste una corrispondenza biunivoca F tra un insieme X e l'insieme dei suoi sottinsiemi, denotato da $\mathcal{P}(X)$: supponi per assurdo che vi sia e chiamiamo stavolta $X_a = F(a)$. Sia dunque:

$$X^* = \{a \in X | a \notin X_a\} = \{a \in X | a \notin F(a)\}$$

Siccome abbiamo ipotizzato che F sia una corrispondenza biunivoca fra X e l'insieme delle sue parti $\mathcal{P}(X)$, vi sarà un $a \in X$ tale che $F(a) = X^*$. Ora, questo elemento a, o appartiene ad X^*, oppure non gli appartiene:

- Se $a \in X^*$, allora per definizione $a \notin F(a)$; ma $X^* = F(a)$, dunque $a \notin X^*$ (assurdo).
- Se $a \notin X^*$, allora per definizione $a \in F(a)$ e dunque $a \in X^*$ (assurdo).

Ne segue che una fale F non può esistere.

Autoreferenza e diagonalizzazione

Teorema di Cantor.

- Consideriamo un qualsiasi insieme X e l'insieme di tutti i suoi sottinsiemi, che denotiamo con $\mathcal{P}(X)$. Siccome ad ogni elemento $a \in X$ posso associare l'insieme $\{a\}$, ne segue che $\mathcal{P}(X)$ è più "almeno grosso" quanto X stesso: in termini più matematici, esiste una funzione *iniettiva*.
- In realtà è "più grosso" di X: dal teorema di Cantor segue infatti che non esiste una funzione *suriettiva* da un insieme all'insieme delle sue parti.

Autoreferenza e diagonalizzazione

Paradosso di Cantor.

- Supponi però che X sia "l'insieme di tutti gli insiemi": allora la collezione di tutti i suoi sottinsiemi *deve* essere un suo sottinsieme, e dunque meno o ugualmente "grossa" (cioè di cardinalità minore od uguale), e questo contraddice il risultato precedente.
- Se ne conclude che la collezione di tutti gli insiemi non è un insieme e si dice "classe propria": Cantor distinse pertanto fra collezioni che sono insiemi e collezioni che non lo sono.

Autoreferenza e diagonalizzazione

Queste sono applicazioni di un metodo più generale, condensato nella seguente proposizione:

Lemma diagonale di Cantor. *Consideriamo una relazione binaria X su un insieme U. Per conformarci alla notazione precedente scriviamo, per ogni $a \in U$:*

$$X_a = \{b \in U | X_a(b)\}$$

intendendo l'insieme dei b appartenenti ad U che stanno in relazione con a, e

$$X^* = \{a \in U | non - X_a(a)\}$$

cioè l'insieme degli elementi a che non stanno in relazione con sé stessi. Allora $X^* \neq X_a$, per ogni $a \in U$.

Autoreferenza e diagonalizzazione

Paradosso (o antinomia) di Russell.

- Il paradosso di Cantor giunse all'orecchio di Bertrand Russell nel 1901 ed è alla base anche del suo celeberrimo "paradosso" (o meglio, una *antinomia*, già identificata anche da Ernst Zermelo).
- Essa investe il modo in cui formiamo gli insiemi:
  - elencare semplicemente i membri di un insieme $\{a, b, c, d\}$,
  - oppure si può specificare una condizione ben definita $P(x)$ che deve essere soddisfatta dai membri (che possono essere, e nelle usali teorie in effetti sono, a loro volta insiemi) dell'insieme. Così $R = \{x|P(x)\}$ è l'insieme degli oggetti che godono di quella proprietà.
- la teoria ingenua degli insiemi assumeva il cosiddetto *Principio di Comprensione* senza restrizioni, secondo cui *per ogni* proprietà $P(x)$, esiste *sempre* l'insieme i cui membri sono esattamente quegli oggetti che soddisfano quella proprietà.

Autoreferenza e diagonalizzazione

Il "Principio di Comprensione".

- L'antinomia emerge quando prendiamo per $P(x)$ la proprietà di non appartenersi $x \notin x$ (ogni insieme ha la caratteristica o di contenere sé stesso come elemento, oppure di non contenere sé stesso come elemento: l'insieme delle tazze da caffè non è una tazza da caffè). Se ci chiediamo se l'insieme degli insiemi che non si appartengono $R = \{x|x \notin x\}$ si appartiene o no, vediamo chiaramente che $R \in R$ se e solo se $R \notin R$.
- L'antinomia colpiva il sistema logico di Frege e dunque il suo progetto di riduzione dell'aritmetica alla logica, poiché, in effetti, mostrava che gli assiomi che stava usando per formalizzare la logica erano incoerenti. In particolare, l'Assioma V di Frege implicava che, corrispondente a qualsiasi predicato, esistesse la classe di tutte e solo quelle cose che lo soddisfano

Autoreferenza e diagonalizzazione

L'antinomia di Russell può essere ricavata dal "Lemma diagonale" di Cantor. Basta infatti porre:

$$X_a(y) = y \in a$$

Si ha allora che per ogni insieme a:

$$X_a = \{x | X_a(x)\} = \{x | x \in a\} = a$$

Pertanto:

$$X^* = \{y | non - X_y(y)\} = \{y | y \notin y\} \neq X_a$$

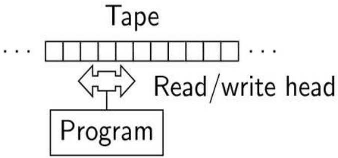
per ogni insieme a. Ma $X_a = a$, dunque X^* (la collezione degli insiemi che non si appartengono) è diverso da ogni insieme: nota infatti che se $X^* = X_a$, per qualche a, avremmo che $a \in a$ se e solo se $a \notin a$.

Limiti dei computers

Il teorema di Turing e le limitazioni dei computers

- Con lo stesso metodo si può dimostrare un altro importante risultato, sui cui intrecci con il primo teorema di Gödel torneremo più avanti, ossia il teorema di Turing.
- Useremo "macchine Turing", "algoritmo" e "programma" in modo intercambiabile. Le macchine di Turing, introdotte da Alan Turing nel 1936, sono computer ideali molto potenti che dispongono di una memoria illimitata, capaci di calcolare tutto ciò che un computer reale calcola.
- La macchina di Turing (benché come modello sia meno realistico ad es. delle macchine RAM) costituisce il modello di riferimento sia in Teoria della Calcolabilità che in Teoria della Complessità, per la sua semplicità, abbinata ad un potere computazionale almeno pari a quello di qualunque altro modello.

Limiti dei computers



- La *tesi di Church-Turing* afferma che ogni problema che può essere risolto da un algoritmo, può essere risolto da una macchina di Turing.
- La *macchina Turing universale* (un computer programmabile) può eseguire qualsiasi operazione che può essere eseguita dai più potenti computer di oggi. In effetti, tutti i moderni computer digitali sono essenzialmente macchine Turing universali.

Limiti dei computers

- Una macchina di Turing su un dato input può fare tre cose:
  1. può *accettare* l' input, ovvero arrestarsi dopo un numero finito di passi di calcolo in un particolare stato detto di accettazione, cioè dicendo "SI",
  2. può *rifiutare* l'input e arrestarsi dunque in uno stato di rifiuto, cioè dicendo "NO",
  3. oppure *entrare in un loop*.

  Un input viene *accettato* da una macchina di Turing che lo riceve, se dopo un tempo finito questa si arresta in uno stato detto *di accettazione*.

Limiti dei computers

Il teorema di Turing.

- Il teorema di Turing evidenzia che ci sono comunque alcuni *problemi che i computer non possono risolvere*, anche se hanno spazio e tempo illimitati per eseguire i loro calcoli.
- Consideriamo macchine standard il cui alfabeto degli input sia 0,1. Le macchine di Turing, intese come programmi, ovvero insiemi di istruzioni, possono esse stesse essere codificate in base due e come tali ordinate lessicograficamente: nella matrice iniziale, stavolta interpretiamo dunque $X_0, X_1, X_2 \ldots$ come una enumerazione delle macchine di Turing. Con la notazione $\mathbf{n}$ indichiamo una rappresentazione in base due della macchina X_n.
- Ora $X_n(\mathbf{n})$ sarà quindi l'n-esimo programma applicato alla propria descrizione.

Limiti dei computers

Il teorema di Turing.

- Possiamo usare lo stesso ragionamento del lemma di diagonalizzazione, per ricavare questa versione del teorema di insolubilità dello *halting problem*, dimostrato da Alan Turing:

  Non esiste un programma C che, quando riceve in input una coppia $(\mathbf{k}, \mathbf{n})$ *dopo un numero finito di passi dice "SI", ovvero accetta, se* $X_k(\mathbf{n})$ *a sua volta dopo un numero finito di passi accetta, e dice "NO", ovvero rifiuta, se* $X_k(\mathbf{n})$ *non accetta (ossia rifiuta, oppure la computazione va avanti all'infinito).*

Autoreferenza e diagonalizzazione

$$
\begin{array}{lllll}
 & 0 & 1 & 2 & \dots \\
X_0 & X_0(\mathbf{0}) & X_0(\mathbf{1}) & X_0(\mathbf{2}) & \dots \\
X_1 & X_1(\mathbf{0}) & X_1(\mathbf{1}) & X_1(\mathbf{2}) & \dots \\
X_2 & X_2(\mathbf{0}) & X_2(\mathbf{1}) & X_2(\mathbf{2}) & \dots \\
\vdots & \dots & \dots & \dots &
\end{array}
$$

- Supponi infatti che un tale programma C esista e sia X^* la macchina che, dato n, fa girare C su $(\mathbf{n}, \mathbf{n})$ e, se C accetta, allora X^* rifiuta e viceversa.
- Questa X^* sarà a sua volta una qualche X_e della lista; ma allora $X_e(\mathbf{e})$ accetta se e solo se rifiuta (assurdo).

Autoreferenza e diagonalizzazione

Riassumendo.

1. Negli esempi precedenti abbiamo definito una funzione f per trasformare la diagonale di una matrice in una sequenza che non apparteneva alla matrice,
2. nei risultati che vedremo adesso definiremo una funzione f per trasformare la diagonale in una riga che al contrario *appartiene* alla matrice.
3. in questo secondo caso la funzione avrà un punto fisso, cioè un elemento x del suo domino tale che $f(x) = x$.

Logica e sistemi formali assiomatici del prim'ordine

Formalizzazione

Forma grammaticale vs. forma logica.

- La logica moderna si occupa di ragionamenti formulati in linguaggi formali, che a differenza di quelli naturali, non contengono ambiguità.
- Frege: il linguaggio naturale è paragonabile all'occhio nudo e il linguaggio formale al microscopio, che ci consente una maggiore precisione.
- Russell: la forma grammaticale di un enunciato, ossia la sua struttura di superficie, è spesso fuorviante e può differire dalla forma logica sottostante.
- Carnap: il linguaggio naturale è uno strumento inadatto al dibattito scientifico e filosofico.

Autoreferenza e diagonalizzazione

Linguaggi logici. **Logica del prim'ordine (o calcolo dei predicati del prim'ordine).**

- Per "logica" qui intenderemo quella cosiddetta "del prim'ordine". La logica del primo ordine riveste un ruolo centrale nella maggior parte delle applicazioni, tanto da essere considerata la logica standard.
- La messa a fuoco delle peculiarità, e dunque l'affermazione della logica *del prim'ordine* sono legate principalmente a logici come Thoralf Skolem, David Hilbert e Kurt Gödel.
- Si deve principalmente a Gottlob Frege l'introduzione dei linguaggi logici moderni, segnatamente il chiarimento del meccanismo della quantificazione (sebbene il simbolismo da lui adottato divergesse notevolmente) con la pubblicazione del volume *Begriffsschrift* (Ideografia) nel 1879.
- In realtà Frege (come poi Russell e Whitehead nei *Principia Mathematica*) lavorasse in sistemi di ordine superiore.

Autoreferenza e diagonalizzazione

Table: Linguaggio proposizionale

Calcolo proposizionale	significato
$p_0, p_1, p_2 \dots$	infinite variabili proposizionali
$\neg$	non
$\wedge$	e
$\vee$	oppure
$\rightarrow$	implica
$\leftrightarrow$	se e solo se
$($, $)$	parentesi

Autoreferenza e diagonalizzazione

Table: Linguaggi del prim'ordine: aggiungono a quelli della tabella precedente i seguenti simboli:

Linguaggi del prim'ordine	significato
$x_0, x_1, x_2 \ldots$	infinite variabili individuali
$P(x_0, \ldots, x_n)$	simboli relazionali di ogni arietà $n \geq 0$
$f(x_0, \ldots, x_n)$	simboli funzionali di ogni arietà $n \geq 0$
$c_0, c_1, c_2 \ldots$	costanti individuali
$\forall$	quantificatore "per ogni"
$\exists$	quantificatore "esiste"

Talvolta le variabili proposizionali del calcolo proposizionali vengono identificate con realzioni zero-arie e le costanti vengono identificate con funzioni zero-arie. Un caso particolare, che può ricevere un diverso trattamento (teorie con o senza identità), è quello della relazione di identità $=$.

Autoreferenza e diagonalizzazione

Nei linguaggi *del prim'ordine* i quantificatori vengono applicati solo alle variabili individuali: $\forall x_i$, $\exists x_i$, ma non a simboli per relazioni o funzioni. Ad esempio, le frasi sulla sinistra possono essere formalizzate da espressioni della forma di quelle sulla loro destra (per comodità usiamo qui x, y, z... come metavariabili, simboli cioè che stanno per variabili individuali). Le prime due sono *atomiche*, cioè consistono nell'attribuzione di un predicato ad uno o più termini:

Socrate è mortale	$P(a)$
Bob ama Alice	$A(a, b)$
Sono tutti pazzi	$\forall x P(x)$
Tutti gli uomini sono mortali	$\forall x(U(x) \rightarrow M(x))$
Esiste almeno un non pazzo	$\exists x \neg P(x)$
C'è qualcuno amato da tutti	$\exists x \forall y A(y, x)$
Ogni artista dona almeno un quadro a Mario	$\forall y(A(y) \rightarrow \exists x(Q(x) \wedge M(y, x, m)))$
La Commutatività dell'addizione	$\forall x \forall y(x + y = y + x)$
Ordine denso	$\forall x \forall y(x < y \rightarrow \exists z(x < z \wedge z < y))$

Autoreferenza e diagonalizzazione

Il linguaggio dell'aritmetica del prim'ordine. A noi qui interesserà un particolare linguaggio del prim'ordine, ossia quello dell'aritmetica di Peano **PA**, che oltre ai simboli logici, all'uguaglianza $=$ e ad infinite variabili individuali contiene:

Costanti	$\bar{0}$
Simboli predicativi	$<$
Simboli funzionali	$+$, $\cdot$, S(x)

dove $S(x) = x + 1$ è detto successore. Si usa per comodità la notazione infissa, ad esempio scrivendo $(x + y)$ e non $+(x, y)$. I *termini* t di questo linguaggio sono le variabili, la costante per lo zero e quelli ottenuti applicando i simboli funzionali ad altri termini.

Autoreferenza e diagonalizzazione

- I numeri vengono denotati da termini detti *numerali*:

$$\overline{n} = \overbrace{SSS...S\overline{0}}^{n-volte}$$

- Il dominio d'interpretazione privilegiato saranno i numeri naturali e l'interpretazione in esso di questi simboli sarà quella ovvia.
- Si ricordi però che una peculiarità delle teorie del prim'ordine è quella di poter avere infinite interpretazioni "non-standard" che verificano gli assiomi della teoria, cioè *non isomorfe a quella intuitiva che le motiva*.

Autoreferenza e diagonalizzazione

Il concetto di sistema formale assiomatico e di dimostrazione formale nella logica moderna.

- Anche questo, inteso come sistema per dedurre teoremi a partire da assiomi, mediante regole d'inferenza che conservano la verità, è un contributo originariamente dovuto a Frege.
- Una dimostrazione, secondo Frege, è una "catena senza lacune" in cui ogni passo è giustificato da un assioma o da una regola completamente esplicitati. Frege, in questo richiamo al rigore euclideo per fondare la logica come scienza, intende andare addirittura oltre il rigore euclideo:

  ...vado oltre Euclide, pretendo che tutti i modi di inferire che vengono applicati vengano specificati all'inizio (Frege, 1883)

Autoreferenza e diagonalizzazione

Questa è una assiomatizzazione della logica del prim'ordine classica "alla Hilbert-Frege":

Schemi d'assioma

1. $\alpha \to (\beta \to \alpha)$
2. $(\alpha \to \beta) \to ((\alpha \to (\beta \to \gamma)) \to (\alpha \to \gamma))$
3. $\alpha \to (\beta \to (\alpha \wedge \beta))$
4. $(\alpha \wedge \beta) \to \alpha$
5. $(\alpha \wedge \beta) \to \beta$
6. $\alpha \to (\alpha \vee \beta)$
7. $\beta \to (\alpha \vee \beta)$
8. $(\alpha \to \gamma) \to ((\beta \to \gamma) \to ((\alpha \vee \beta) \to \gamma))$
9. $(\alpha \to \beta) \to ((\alpha \to \neg\beta) \to \neg\alpha)$
10. $\neg\neg\alpha \to \alpha$
11. $\alpha(t) \to \exists x \alpha$
12. $\forall x \alpha \to \alpha(t)$

Autoreferenza e diagonalizzazione

cui si aggiungono le seguenti regole per derivare teoremi da assiomi:
Regole d'inferenza. Una regola proposizionale ("modus ponens"):

$$\frac{\alpha \quad \alpha \to \beta}{\beta}$$

e due regole sui quantificatori:

$$\frac{\beta \to \alpha}{\beta \to \forall x \alpha} \qquad \frac{\alpha \to \beta}{\exists x \alpha \to \beta}$$

(dove in entrambi i casi x non deve essere libera in β).

- Una *derivazione* di un *teorema* è pertanto una sequenza finita di formule dove ciascuna formula o è un assioma, oppure è derivata da qualcuna delle formule precedenti nella sequenza mediante le suddette regole.

Autoreferenza e diagonalizzazione

Una specifica teoria matematica, formulata in questa logica, aggiunge ai precedenti *assiomi logici* degli *assiomi specifici* di quella teoria. Ad esempio l'aritmetica di Peano **PA** aggiunge i seguenti assiomi:

1. $\forall x(\neg(S(x) = 0)$
2. $\forall x \forall y(S(x) = S(y) \to x = y)$
3. $\forall x(x + \overline{0} = x)$
4. $\forall x \forall y(x + S(y) = S(x + y))$
5. $\forall x(x \cdot \overline{0} = \overline{0})$
6. $\forall x \forall y(x \cdot S(y) = (x \cdot y) + x)$

Ci saranno poi assiomi per $<$ (se è stato messo fra i simboli del linguaggio), cui si aggiunge (per ogni formula α) lo schema d'assiomi d'induzione matematica:

$$(\alpha(\overline{0}) \wedge \forall x(\alpha(x) \to \alpha(S(x))) \to \forall x \alpha(x)$$

Autoreferenza e diagonalizzazione

Completezza e incompletezza.

- In logica si hanno due nozioni di validità:
  1. *validità sintattica*, o dimostrabilità dagli assiomi, cioè essere un teorema, sia una nozione di
  2. *validità semantica*, o verità in tutte le interpretazioni del linguaggio del sistema.

  Una importante proprietà è che nella logica e le teorie del prim'ordine, per un noto teorema di *completezza semantica* dovuto a Gödel (1930), 1. e 2. *coincidono*. L'implicazione 1. $\Rightarrow$ 2. è detta *correttezza*, mentre 2. $\Rightarrow$ 1. è la *completezza* in senso proprio.
- si ricordi che una peculiarità delle teorie del prim'ordine è quella di possedere *infinite* interpretazioni non isomorfe che ne verificano gli assiomi.

Autoreferenza e diagonalizzazione

Completezza e incompletezza.

- Questa nozione non va confusa con quella che occorre nel celebre risultato di *incompletezza* di Gödel (1931) che concerne le teorie matematiche come ad esempio **PA**, che ha un'interpretazione privilegiata, ossia i numeri naturali.
- In questo caso parliamo di *correttezza aritmetica*, se la teoria non dimostra proposizioni false riguardo ai numeri naturali, e di *completezza aritmetica*, se ogni proposizione vera riguardo ad essi è dimostrabile nella teoria.
- (**Gödel-Rosser**) in ogni sistema formale assiomatico consistente (cioè non contraddittorio) in cui una certa quantità di aritmetica elementare può essere eseguita, esistono enunciati né dimostrabili, né refutabili, ma *veri* nella interpretazione privilegiata costituita dai numeri naturali con addizione e moltiplicazione, o "modello standard".

Autoreferenza e diagonalizzazione

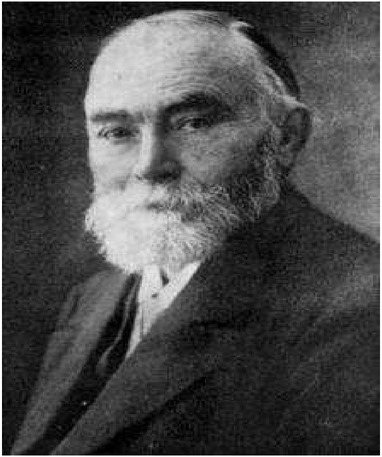


Figure: Gottlob Frege (1848-1925)

Autoreferenza e diagonalizzazione

Indecidibilità.

- I primi logici del '900 ambivano ad ottenere il risultato vagheggiato da Leibniz con il suo avveniristico progetto di un linguaggio universale e di un calcolo logico:

  > *Questa scrittura sarà una specie di algebra generale e offrirà il modo di ragionare calcolando, di modo che , invece di disputare, si potrà dire: calcoliamo! (G. W. Leibniz, Phil. VII, 26, 1679)*

- Oggi sappiamo che questo ideale (ripreso da David Hilbert nella formulazione dello *Entscheidungsproblem*, o problema della decisione) è realizzabile solo al livello del calcolo proposizionale, ma non in generale del calcolo dei predicati del prim'ordine e delle sue estensioni (A. Church e A. Turing, 1936)

Autoreferenza e diagonalizzazione

Decidibilità e semi-decidibilità. Quando parleremo di sistemi formali assiomatici, assumeremo che godano della proprietà basilare che l'insieme dei loro teoremi è *computabilmente enumerabile*, proprietà soddisfatta da **PA**. Ricorda i seguenti fondamentali concetti:

- Un insieme è detto *computabilmente enumerabile* (o semi-decidibile), se l'appartenenza ad esso può essere verificata in modo algoritmico, ma la *non* appartenenza *potrebbe non esserlo*.
- È invece detto *decidibile*, se è possibile programmare un computer in modo che, se riceve in input una stringa di simboli, dopo un numero finito di passi dice si o no, a seconda che la stringa stia o no nell'insieme (abbiamo un test positivo ed un test negativo).

Autoreferenza e diagonalizzazione

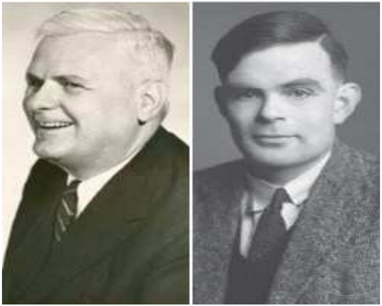


Figure: Teorema di indecidibilità (Church e Turing 1936): i programmi non riescono neppure a dominare le leggi logiche

Autoreferenza e diagonalizzazione

- In generale, non abbiamo un algoritmo che quando riceve in input una formula, dopo un numero di passi dice "si" se essa è un teorema, e dice "no" se essa non lo è.
- Nei sistemi formali assiomatici che citeremo abbiamo viceversa solo un algoritmo che dice "si", se la formula è un teorema, ma può girare all'infinito se non lo è.
- In questi casi diciamo che l'insieme dei teoremi è *computabilmente enumerabile* (o semi-decidibile): per sapere se una formula è un teorema, possiamo ad esempio scorrere tutte le stringhe possibili di simboli e se qualcuna costituisce una prova di essa prima o poi la troviamo.

Autoreferenza e diagonalizzazione

Teorie capaci di parlare di sé stesse: riduzione della metamatematica alla teoria dei numeri

- Per far sì che **PA** possa "parlare di se stessa", si codifica con dei numeri la sua sintassi.
- Ma il sistema formale in questione contiene *numerali*, ovvero i termini utilizzati per denotare questi numeri e fungeranno pertanto da "nomi" delle formule codificate;
- ad ogni formula α è dunque possibile assegnare un termine, solitamente un numero denotato da un termine che indichiamo con $\ulcorner\alpha\urcorner$ e che costituisce il suo "nome".
- Formule, prove (cioè sequenze di formule), proprietà metateoriche in genere vengono in tal modo *riflesse nel sistema formale* o teoria, così da poter parlare della teoria, all'interno della teoria stessa: in particolare la nozione di *dimostrabilità* è in tal modo internalizzabile.

Autoreferenza e diagonalizzazione

- Questo risultato può essere ottenuto grazie ad un procedimento detto "aritmetizzazione", che consente di rappresentare stringhe finte di simboli di un linguaggio qualsiasi *mediante numeri*, assegnando loro un codice (un "nome") in modo non ambiguo.
- Se ad esempio assegniamo ai simboli questi numeri:

$\neg$	$\wedge$	$\vee$	$\rightarrow$	$\leftrightarrow$	$\forall$	$\exists$	$=$	(	)	0	S	$+$	$\times$	x	y	z	...
1	3	5	7	9	11	13	15	17	19	21	23	25	27	2	4	6	...

Possiamo sfruttare l'unicità della scomposizione in fattori primi per codificare la formula $\exists y(S0 + y) = SS0$ come:

$$2^{13} \times 3^{4} \times 5^{17} \times 7^{23} \times 11^{21} \times 13^{25} \times 17^{4} \times 19^{19} \times 23^{15} \times 29^{23} \times 31^{23} \times 37^{21}$$

Autoreferenza e diagonalizzazione

In vista del teorema di Gödel ci sarà utile questa osservazione:

- Una dimostrazione è una sequenza finita di formule. *Codificare numericamente* una dimostrazione alla fine si riduce a *codificare una sequenza di codici di formule*, con il metodo che abbiamo prima accennato.
- Ma relazioni (decidibili) come: "la sequenza codificata da n è una dimostrazione in **PA** della formula codificata da m", sono, in un senso tecnico preciso, *rappresentabili numericamente* da una formula del linguaggio di **PA**, che nella fattispecie denotiamo $Prf_{PA}(\overline{n}, \overline{m})$;
- il predicato che definisce l'insieme (computabilmente enumerabile) dei teoremi: "m è dimostrabile" è pertanto formalizzato con una formula relativamente semplice, della forma $\exists y Prf_{PA}(y, \overline{m})$.

Punti fissi

Autoreferenza e diagonalizzazione

Una applicazione significativa, ai nostri fini, di questa capacità di "parlare di sé stesse" delle teorie, è che per esse vale il teorema, detto "di punto fisso":

> *Per ogni formula $\alpha(x)$ esiste una formula β tale che è dimostrabile l'equivalenza:*

$$\beta \leftrightarrow \alpha(\ulcorner \beta \urcorner)$$

Vedremo due applicazioni di esso:

1. quando $\alpha(x)$ è la formalizzazione di "x non è vero" (Tarski)
2. quando $\alpha(x)$ è la formalizzazione di "x non è dimostrabile nella teoria" (Gödel)

Autoreferenza e diagonalizzazione

- Invece di dimostrare il teorema di punto fisso, cerchiamo qui di evidenziare come esso in qualche modo emerga dalla matrice che abbiamo utilizzato all'inizio
- Va comunque sottolineato che secondo alcuni logici, questo *meccanismo puramente sintattico* non autorizza immediatamente a parlare di "autoriferimento", se non in maniera naïve: l'esistenza di punti fissi è condizione necessaria, ma non sufficiente per avere l'autoriferimento in un senso più raffinato.
- È importante notare che nelle espressioni autoreferenti che vedremo, uno stesso numero riveste un duplice ruolo: come *codice* della formula e come *argomento* della formula. Così le formule si riferiscono indirettamente a sé stesse.

Autoreferenza e diagonalizzazione

- Le formule del linguaggio di queste teorie possono essere enumerate e dunque usiamo per maggior chiarezza la notazione precedente per tale enumerazione:

  $$X_0, X_1, X_2 \ldots$$

  Supponiamo senza perdita di generalità che i pedici denotino proprio i codici delle relative formule e che ogni numero sia il codice di una e una sola formula, cioè che $\ulcorner X_e \urcorner = \overline{e}$, cosicché, se $X_e(x)$ rappresenta un predicato ad un posto, $X_e(\overline{e})$ sarà proprio $X_e(\ulcorner X_e \urcorner)$ (la *diagonalizzazione* di $X_e(x)$).

Autoreferenza e diagonalizzazione

- La notazione $X_n(\overline{m})$ verrà interpretata come $X_n(\ulcorner X_m \urcorner)$, ovvero, la formula ad una variabile di codice n applicata alla formula ad una formula di codice m (e quindi $X_n(\ulcorner X_n \urcorner)$ sarà la diagonalizzazione di X_n).
- Tornando alla matrice iniziale, adesso considera una funzione f ("switching function") che agisce sulla diagonale di questa matrice:
$$D = X_0(\overline{0}), X_1(\overline{1}), X_2(\overline{2})...$$
trasformandola in:
$$f(D) = f(X_0(\overline{0})), f(X_1(\overline{1})), f(X_2(\overline{2}))...$$

Autoreferenza e diagonalizzazione

Si danno due possibilità:

1. se f è stata scelta in modo che $f(D)$ è diversa da ogni riga della matrice, allora abbiamo il caso di Cantor.
2. Se viceversa:

$$f(X_a(\overline{a})) = \text{“}X_a(\overline{a}) \text{ ha la proprietà } P\text{”}$$

   dove la frase fra virgolette può essere formalizzata da una formula $X_n(\overline{a})$ della precedente lista, per qualche n, allora $f(D)$ è uguale a una riga della matrice e precisamente alla riga n,

Autoreferenza e diagonalizzazione

cioè a dire:

$f(X_0(\overline{0})) = X_n(\overline{0})$,
$f(X_1(\overline{1})) = X_n(\overline{1})$,
...
$\mathbf{f(X_n(\overline{n})) = X_n(\overline{n})}$
...
$f(X_k(\overline{k})) = X_n(\overline{k})$

Si noti che la f in questo caso ha un punto fisso (evidenziato in grassetto). Più precisamente, c'è una formula α dimostrabilmente equivalente ad $f(\alpha)$.

Indefinibilità della verità

Autoreferenza e diagonalizzazione: il concetto di "verità"

Il teorema di Tarski.

- Tra i più importanti concetti semantici (cioè quei concetti che hanno a che fare col *significato* delle espressioni linguistiche), vi è in primo luogo quello di *verità*, che tuttavia, come rivela il paradosso del mentitore, possiede degli aspetti problematici.
- Il concetto di verità fu introdotto nella logica moderna in modo rigoroso dal logico polacco-americano Alfred Tarski (1901-1983) negli anni '30 del secolo scorso.
- Tarski pubblicò due classici lavori in cui dette una definizione della verità per i linguaggi formali, *The Concept of Truth in Formalized Languages* (1933) e *On the Concept of Logical Consequence* (1936). Egli riteneva che una simile trattazione rigorosa non fosse possibile per le lingue naturali.

Autoreferenza e diagonalizzazione

- È abbastanza comune considerare la cosiddetta *teoria semantica della verità* di Tarski come una evoluzione della teoria aristotelica e scolastica della verità come *corrispondenza* (*veritas est adequatio rei et intellectus*), presentata in forma moderna. Il legame della teoria tarskiana con la teoria aristotelico - tomistica della verità come corrispondenza è però oggetto di controversia.
- Secondo Tarski i "portatori di verità" (le entità di cui si può dire che sono vere o false) sono gli *enunciati*: ogni portatore di verità è correlato uno stato di cose, e se quello stato di cose effettivamente sussiste, il portatore della verità è vero.

Autoreferenza e diagonalizzazione

Consider the sentence 'snow is white.' We ask the question under what conditions this sentence is true or false. It seems clear that if we base ourselves on the classical conception of truth, we shall say that the sentence is true if snow is white, and that it is false if snow is not white. Thus, if the definition of truth is to conform to our conception, it must imply the following equivalence:

The sentence 'snow is white' is true if, and only if, snow is white

(A. Tarski).

Autoreferenza e diagonalizzazione

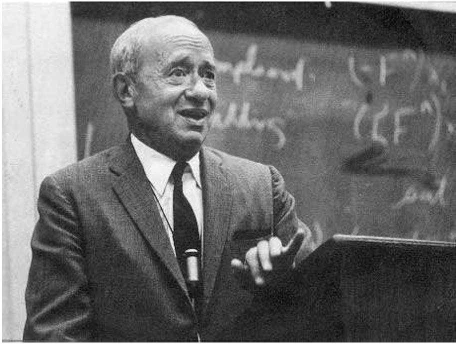


Figure: Alfred Tarski (1901-1983)

Autoreferenza e diagonalizzazione

- Tuttavia, secondo Tarski, nelle lingue naturali emerge il *paradosso del mentitore*, e questo perché sono *semanticamente chiuse*, ovvero:
  1. contengono il proprio predicato di verità, cioè l'espressione "è vero",
  2. attraverso il meccanismo della citazione contengono gli strumenti per formare nomi dei propri enunciati.
- Per questo Tarski considera nella sua analisi solo linguaggi formali, che sono aperti e "gerarchizzabili", cioè possiamo distinguere il linguaggio che svolge il ruolo di linguaggio oggetto L da quello, gerarchicamente superiore, che contiene L, ma non è del tutto traducibile in esso, che svolge il ruolo di metalinguaggio ML.

Autoreferenza e diagonalizzazione

- Uno dei criteri fondamentali individuati da Tarski per una corretta definizione della verità, è dunque la condizione, detta di *adeguatezza materiale*, che rispecchia le considerazioni informali del brano che abbiamo letto, però con alcune precisazioni:

  ogni definizione della verità per un linguaggio L in un metalinguaggio ML deve implicare tutti gli enunciati della forma:

  $$X \text{ è vero se e solo se } p$$

  dove X è il nome di un enunciato di L e p è una descrizione (o traduzione) di esso nel metalinguaggio ML.

Autoreferenza e diagonalizzazione

- Le lingue naturali però sono troppo "universali", confondono spesso X con p consentendo quindi di riprodurre il "mentitore".
- Lo sviluppo rigoroso della semantica, secondo Tarski, prevede che linguaggio e metalinguaggio siano sempre distinti e che il metalinguaggio non possa essere completamente tradotto nel linguaggio oggetto: non possa esserlo in particolare il predicato di verità 'è vero', pena il riprodursi del Mentitore.
- lo schema precedente va dunque precisato in questo senso:
  1. la definizione di "è vero" per L avviene in un metalinguaggio ML,
  2. X è il nome in ML di un enunciato σ di L e p è una copia in ML di σ.

Autoreferenza e diagonalizzazione

Tarski mostra come lo schema di adeguatezza conduca al paradosso del mentitore, se ammettiamo che un linguaggio contenga il proprio predicato di verità:

> *We have implicitly assumed that the language in which the antinomy is constructed contains, in addition to its expressions, also the names of these expressions, as well as semantic terms such as the term "true" referring to sentences of this language; we have also assumed that all sentences which determine the adequate usage of this term can be asserted in the language. A language with these properties will be called "semantically closed (A. Tarski, 1944).*

Autoreferenza e diagonalizzazione

La difficoltà evidenziata da Tarski emerge più chiaramente se formalizziamo il ragionamento in linguaggi formali e in sistemi formali abbastanza espressivi. Il Teorema di indefinibilità di Tarski afferma:

> *Nella cornice di una logica bivalente e con alcune risorse espressive di base, nessun linguaggio con capacità di autoreferenzialità può contenere il proprio predicato di verità.*

Siccome facciamo ricorso a codifiche numeriche, è ragionevole prendere allora un sistema formale che includa anche un minimo di teoria dei numeri e considereremo a tal fine l'Aritmetica di Peano **PA**, fermo restando che esso ha un significato più generale e si estende a qualsiasi sistema formale che possegga un sufficente grado di auto-referenzialità.

Autoreferenza e diagonalizzazione

Sulla base di ciò che abbiamo detto questo teorema segue in maniera pressoché immediata, supponendo, per assurdo, che un tale predicato di verità *True*(x) esista:

1. In quanto predicato di verità, è dimostrabile, per ogni formula α, l'equivalenza $\alpha \leftrightarrow \mathit{True}(\lceil \alpha \rceil)$,
2. ma per il teorema di punto fisso abbiamo anche, per qualche β, che $\beta \leftrightarrow \neg \mathit{True}(\lceil \beta \rceil)$,
3. prendendo dunque $\alpha = \beta$ abbiamo una contraddizione, simile al "mentitore".

Importante: Kurt Gödel osservò che certi paradossi non sarebbero sorti necessariamente se il concetto di *verità* fosse stato sostituito dal concetto di *dimostrabilità*. Ciò implica che questi due concetti *non coincidono*.

Verità e dimostrabilità

Autoreferenza e diagonalizzazione

Figure: Kurt Goedel (1906-1978)

Autoreferenza e diagonalizzazione

Il Teorema di Gödel. Che succede se invece di prendere una formula equivalente alla sua falsità, prendiamo una formula equivalente alla sua indimostrabilità? L'idea alla base del primo teorema di incompletezza di Gödel fu quella di prendere il paradosso del mentitore, ma sostituire il predicato "è vero" con la sua *approssimazione sintattica* "è dimostrabile", che al contrario, è definibile nel linguaggio della teoria, se esso è abbastanza ricco. Vediamo perciò cosa succede se invece di prendere la frase:

questa frase è falsa

prendiamo la frase che dice di sé stessa di non essere *dimostrabile*:

questa frase non è dimostrabile.

Autoreferenza e diagonalizzazione

- La *dimostrabilità* in un sistema formale è definibile e la *verità* non lo è, e quindi, se assumiamo che le proposizioni dimostrabili in un sistema formale assiomatico siano vere, ne traiamo una indiretta versione del primo teorema di Gödel, ovvero che esistono proposizioni vere non dimostrabili in tale sistema, da cui il famoso primo teorema di incompletezza.
- Più precisamente, se la teoria non dimostra cose false riguardo all'interpretazione privilegiata costituita dai numeri naturali con le operazioni (è cioè *aritmeticamente corretta*) e se α è il punto fisso di Gödel, allora né α, né $\neg\alpha$ risultano dimostrabili.

Autoreferenza e diagonalizzazione

Infatti (lasciando al lettore di verificare che la correttezza aritmetica implica la non-contraddittorità o "consistenza"), consideriamo il punto fisso di Gödel:

$$\alpha \leftrightarrow \neg \exists y Prf_{PA}(y, \ulcorner \alpha \urcorner)$$

1. Se α è dimostrabile, allora esiste un numero che codifica una sua dimostrazione e dunque è dimostrabile la negazione della parte destra della suddetta equivalenza, e di conseguenza è dimostrabile $\neg \alpha$ (assurdo)
2. Se $\neg \alpha$ è dimostrabile, allora lo è la negazione della parte destra dell' equivalenza, e dunque (correttezza) è anche vera: cioè a dire qualche numero codifica una dimostrazione di α (assurdo).

Autoreferenza e diagonalizzazione

Ma per prevenire le incomprensioni dell'ambiente filosofico e matematico di allora, caratterizzato da una generale diffidenza verso i concetti semantici coinvolti nella nozione di correttezza, Gödel preferì cercare un'altra strada, quasi completamente sintattica, basata sulla nozione di "coerenza" e di un suo rafforzamento ("$\omega -$ *coerenza*") ed immune da possibili critiche di tale sorta.

Menti, corpi e macchine

La mente è una macchina?

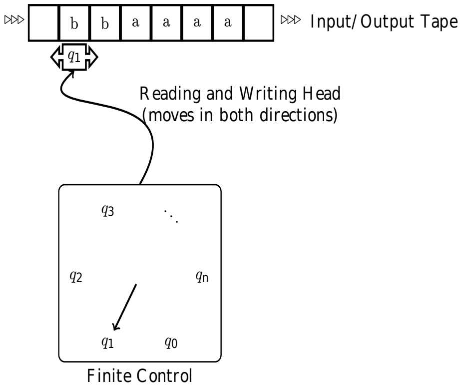


Figure: Una macchina di Turing. Qualsiasi sistema che condivide la capacità di calcolo della *macchina universale* è detto Turing completo. Un modo per dimostrare che un sistema è Turing completo è dunque dimostrare che può simulare una macchina Turing universale.

La mente è una macchina?

Menti, macchine e il teorema di Gödel.

- Una macchina può pensare? La mente stessa può essere concepita come una macchina? La teoria conosciuta come *computazionalismo*, sostiene che in effetti la mente è simile ad un computer:

  "L'idea alla base del modello computazionale della mente è che la mente sia il programma e il cervello l'hardware di un sistema computazionale. Uno slogan che si incontra spesso dice: La mente è per il cervello ciò che il programma è per l'hardware...Ciò che mandò i brividi su e giù per la spina dorsale di un'intera generazione di giovani ricercatori nel campo dell'intelligenza artificiale è il pensiero seguente: Immagina che il cervello sia una macchina di Turing universale. "(J. Searle)

La mente è una macchina?

Menti, macchine e il teorema di Gödel.

- I primi autori a formulare una teoria computazionale della mente furono McCulloch e Pitts nel 1943, che tuttavia utilizzarono il modello delle reti neurali. Questo è poi diventato il modello di riferimento per l'indirizzo connessionista nel campo delle scienze cognitive.
- Kleene (1956) dimostrò che il modello di McCulloch e Pitts è equivalente ad un *automa a stati finiti* (una macchina con una quantità fissa di memoria, mentre una macchina Turing ha una quantità illimitata di memoria).

La mente è una macchina?

Menti, macchine e il teorema di Gödel.

- Siegelmann and Sontag (1992) dimostrano che esiste una versione di questo modello (Recurrent Neural Network) che è Turing-completo e può quindi implementare qualsiasi algoritmo.
- Nel 1948 Turing aveva anch'egli indagato su un modello di rete neurale, denominato *B-type unorganized machine*, nell'articolo *Intelligent Machinery*.

La mente è una macchina?

- Tra coloro che avversano la concezione computazionalista, ve ne sono alcuni che talvolta definiscono “Church-Turing fallacy” l’ipotesi, a loro avviso errata, che i concetti tratti dalla teoria della moderna computabilità, dovuta in gran parte a studiosi come Alonzo Church, Emil Leon Post, Stephen Cole Kleene ed Alan Turing, supportino questa interpretazione.
- Si è tentato di utilizzare i risultati di incompletezza di Gödel allo scopo di confutare il computazionalismo. In questa discussione la nozione di *sistema formale* è spesso identificata con quella di *procedura meccanica per enumerare teoremi*: del resto ogni sistema formale assiomatico il cui insieme dei teoremi è computabilmente enumerabile è convertibile in una macchina e viceversa.

Autoreferenza e diagonalizzazione

La mente è un computer?

- Alcune versioni del teorema di incompletezza (Kleene, Penrose) che ricorrono nella discussione sulle sue implicazioni per questo problema, utilizzano il concetto di *computabilità*, piuttosto che quello di *dimostrabilità*.
- Sotto certe condizioni, in effetti la nozione di "sistema formale assiomatico" i cui teoremi siano enumerabili, è intertraducibile con quella di "macchina" o programma.
- In particolare vi sono legami fra il teorema di Gödel e quello di insolubilità dello *halting problem* di Turing (per un approfondimento, vedi S. Feferman, *Penrose's Gödelian argument*).

Autoreferenza e diagonalizzazione

Lo stesso Gödel del resto afferma che:

> *A formal system can simply be defined to be any mechanical procedure for producing formulas, called provable formulas. For any formal system in this sense, there exists one in the usual sense that has the same provable formulas (and likewise vice versa).*

Ogni sistema formale assiomatico il cui insieme dei teoremi è computabilmente enumerabile è convertibile in una macchina e viceversa.

Mente e macchine

- Già a partire dal 1961 J. R. Lucas aveva avanzato la tesi secondo cui dai risultati di Gödel circa 'incompletezza dell'aritmetica seguisse la confutazione della tesi secondo cui la mente umana può essere accuratamente modellata da un computer digitale, ovvero da una "macchina di Turing" ("impossibilità dell'intelligenza artificiale").
- la mente umana non è simulabile da una macchina –si diceva– poiché se fosse simulabile da una certa macchina A, per il primo teorema di Gödel è possibile esibire enunciati che questa A (intesa come concreta realizzazione di un sistema formale) non può dimostrare, mentre la mente è capace di riconoscerne la verità.

Mente e macchine

- Una obiezione, dovuta originariamente a Putnam, a questa versione naïve è che io so che una certa proposizione indimostrabile nel sistema formale è vera, *se conosco la consistenza (cioè la non-contraddorietà) del sistema formale*.
- Questo dice il primo teorema di Gödel: una proposizione condizionale. Questa premessa non può essere elusa: ovviamente un sistema inconsistente dimostra tutto, inclusa la proposizione di Gödel.
- Ma se dicendo che A rappresenta le capacità dimostrative umane intendo che dimostra le stesse proposizioni matematiche umanamente dimostrabili, allora se io dimostro la consistenza di A, esso stesso dovrà dimostrarla, contro il *secondo* teorema di Gödel: la consistenza di un sistema formale A non è dimostrabile nel sistema stesso.

La mente è una macchina?

- Gli argomenti prodotti da Roger Penrose sono più sofisticati, ma ugualmente problematici. La versione del teorema di Gödel che ricorre nel libro di R. Penrose, *Shadows of the Mind: A search for the missing science of con sciousness*, discende in realtà da un'altra versione dovuta a Kleene (1943, pubblicata nel 1952), che si basa sulla analogia fra sistemi formali assiomatici e programmi (macchine di Turing).
- Sottolineando che si può dare una enumerazione esaustiva dei programmi che calcolano funzioni *parziali* computabili (non solo di quelle *totali*!, ed è un'altra applicazione della diagonalizzazione), in particolare quelle con un solo argomento:

$$C_0(x), C_1(x), C_2(x)...$$

.

La mente è una macchina?

esso è equivalente a quest'altro teorema:
Teorema. Sia A una macchina di Turing che, quando si arresta su input (q, n), allora la "macchina" $C_q(n)$ non si arresta. Allora esiste un numero e tale che $C_e(e)$ non si arresta, ma al contempo A non si arresta su (e, e).

che questa forma ha una è dimostrazione simile a quella del teorema di Alan Turing (vedi S. Feferman, *Penrose's Gödelian argument*, e A. Antonelli, *Gödel, Penrose, e i fondamenti dell'intelligenza artificiale*, per un approfondimento del legame con la versione di Kleene del teorema di incompletezza).

- Intuitivamente, interpretiamo l'arrestarsi della computazione $A(q, n)$ come una *dimostrazione* dell'enunciato: "la computazione $C_q(n)$ non termina".

La mente è una macchina?

- Supponiamo adesso che A sia un programma o "macchina di Turing" che incorpora *tutti i metodi disponibili presso la comunità matematica*. Assumiamo inoltre che A sia corretta, cioè non dimostri cose false;
- in particolare tale programma saprà dimostrare teoremi della forma:

  "La computazione di $C_q(n)$ non termina."

- Dunque, quando A riceve in input (q, n), se dopo un po' si arresta (cioè la sua computazione termina, ovvero "dimostra" l'enunciato suddetto), allora, per la *correttezza*, realmente la computazione di $C_q(n)$ non termina.

La mente è una macchina?

- In particolare, se avviene che A ricevendo come input (n, n) si arresta, cioè dimostra l'enunciato "la computazione di $C_n(n)$ non termina", allora $C_n(n)$ realmente non si arresta (poiché A per ipotesi non dimostra cose false).
- Ma in questo caso, siccome A dipende dall'unico parametro n, potrà essere espressa come $C_e(n)$, per qualche e, cioè farà parte della enumerazione dei programmi per funzioni unarie.

Ma che succede allora se diamo in input ad A lo stesso indice e? Ecco la situazione:

1. Se A su input (e, e) si arresta, allora $C_e(e)$ si arresta,
2. Se A su input (e, e) si arresta, allora $C_e(e)$ *non* si arresta

Se A su input (e, e) si arresta si ha dunque una contraddizione; pertanto A su input (e, e) non si arresta.

La mente è una macchina?

Dunque A è incapace di dimostrare che C_e, su input e, non si arresta, mentre noi lo abbiamo or ora dimostrato! Ergo la sua pretesa di racchiudere le procedure dimostrative "umane" è infondata.

Riepilogando:

- siamo partiti dall'ipotesi che A racchiudesse tutte le procedure dimostrative "umane" e dall'ipotesi della conoscenza da parte di "noi umani" della correttezza di A,
- ma da ciò abbiamo costruito una computazione che noi comprendiamo non arrestarsi e tuttavia A non è in grado di dimostrarlo.
- Dunque i matematici umani non usano una procedura che possa essere riconosciuta come corretta e una tale A non può incorporare tutte le procedure dimostrative di cui dispone la comunità matematica.

La mente è una macchina?

Di questo argomento vi sono varie versioni, e tutte hanno ricevuto fondate obiezioni; ma cosà pensava lo stesso Gödel al riguardo? Egli prende posizione in particolare nelle *Gibbs lectures* del 1951, dove trae in particolare dal suo secondo teorema queste conclusioni, come "a mathematically established fact", in forma di una disgiunzione:

> *O la mente umana supera il potere di ogni macchina finita, oppure esistono problemi indecidibili con ogni tipo di procedura che la mente possa concepire.*

Benché per sua convinzione filosofica egli sia incline a rifiutare l'idea che vi siano problemi insolubili con ogni mezzo razionale (e dunque ad accettare l'idea che la mente non sia una macchina), Gödel non afferma che ciò segua dai suoi teoremi.

La mente è una macchina?

- Per *macchina finita* Gödel intende qui una macchina di Turing, e con *procedura finita* intende una procedura che può essere eseguita da tale macchina. Inoltre, in accordo con la tesi di Church-Turing che identifica la computabilità effettiva con la computabilità su una macchina di Turing.
- Gödel ammette dunque che esistano procedure mentali finite, in qualche senso "costruttive", ma non "meccaniche", cioè non algoritmiche, che consentono alla mente di superare i poteri di ogni macchina.

La mente è una macchina?

Meritevole di approfondimento è il confronto sul tema del rapporto mente-macchina fra la posizione di Turing e quella, diversa, di Gödel:

- Gödel in una breve nota del 1972 intitolata *A Philosophical Error in Turing's Work* sintetizzava la posizione di Turing nello slogan:

  "mental procedure cannot go beyond mechanical procedures"

  e ravvisava tra i punti cardine dell'analisi di Turing i seguenti:

  (A) Non c'è una mente separata dalla materia; e
  (B) il cervello funziona essenzialmente come un computer digitale.

La mente è una macchina?

- Muovendo da una posizione antidualista, secondo cui non c'è una "mente" separata dal cervello (gli stati mentali sono sostanzialmente stati fisici), Turing giungeva alla conclusione che il numero di stati mentali di un essere umano deve essere finito, con questo curioso argomento:

  if we admitted an infinity of states of mind, some of them will be 'arbitrarily close' and will be confused.

  Turing riteneva che dunque anche le macchine potessero sviluppare una forma di pensiero, e che un giorno si sarebbe potuta creare una macchina dotata di una mente propria.

La mente è una macchina?

È oramai popolarmente noto il test di Turing basato sul "gioco dell'imitazione". Turing sosteneva che le macchine che superano il test possono dirsi dotate di pensiero. Gödel era invece assertore di una forma marcata di dualismo, riteneva che il punto (A), ossia che non si può dare mente senza materia, fosse solo "a prejudice of our time":

> *What Turing disregards completely is the fact that mind, in its use, is not static, but constantly developing...Therefore, although at each stage of the mind's development the number of its possible states is finite, there is no reason why this number should not converge to infinity in the course of its development*

La mente è una macchina?

- Ma la critica centrata su una presunta scarsa attenzione di Turing agli aspetti dinamici della conoscenza è ingenerosa e non si basa su una corretta interpretazione del suo pensiero, men che mai del periodo post-bellico.
- Nei lavori tra il 1940 e il 1950 (con cui gettò le basi della odierna intelligenza artificiale), Turing descriveva le macchine che esibiscono un comportamento intelligente come dotate di una struttura più complessa, rispetto a quella di una singola macchina di Turing.
- In particolare, in questi scritti viene posta in evidenza la capacità di interagire con l'ambiente e viene formulata l'ipotesi che la mente sia rappresentabile come una serie di macchine distinte in stadi diversi, o come una macchina capace di apprendere dall'esperienza, modificando le proprie istruzioni.

Macchine e organismi viventi

Il meccanicismo classico. Narra un celebre aneddoto che quando Cartesio espose a Cristina di Svezia l'ipotesi secondo cui gli animali costituiscono una forma di automa meccanico, ella, additando un orologio, esclamò: "Vediamo se produce un figlio". La visione degli animali come macchine (la bête machine), in altre parole, si imbatteva in serie difficoltà, di fronte all'obiezione che gli organismi viventi, in generale, a differenza delle macchine, si diceva (e come si è ritenuto a lungo), hanno la capacità di riprodurre sé stessi:

> *Mettez une machine de chien et une machine de chienne l'une auprès de l'autre, et il en pourra résulter une troisième petite machine, au lieu que deux montres seront auprès l'une de l'autre, toute leur vie, sans jamais faire une troisième montre. (Bernard Le Bouyer de Fontenelle, 1742)*

Autoriproducibilità delle macchine

Gli organismi, a differenza delle macchine, sono sistemi auto-organizzanti che si autoriproducono. Dal tempo di Cartesio e La Mettrie molte cose (a partire dallo stesso concetto di 'macchina') sono cambiate, ma la contrapposizione fra concezioni vitalistiche, organicistiche e meccanicistiche della vita è ancora forte.

- Certe tendenze meccanicistiche concepiscono gli organismi viventi come macchine molto complesse programmate da un software genetico.
- Le concezioni anti-meccanicistiche sottolinano invece la irriducibile diversità fra organismi e macchine.

Autoriproducibilità delle macchine

Oggi, quella intorno ai sistemi autoreplicanti e sulle condizioni generali che devono soddisfare, è diventata comunque una importante area di ricerca nel campo della vita artificiale. Alla fine del 1940, John von Neumann cominciò ad occuparsi della autoriproducibilità dei sistemi artificiali, considerando poi, nei primi anni '50, in particolare il modello matematico costituito dai cosiddetti automi cellulari, uno strumento assai potente, capace di simulare le macchine di Turing. In particolare, si trattava di verificare la possibilità, logica e pratica, di estendere i fenomeni osservati nei sistemi biologici, alla teoria degli algoritmi e degli automi.

Autoriproducibilità delle macchine

Un sistema fisico non biologico autoreplicante fu presentato per la prima volta da John von Neumann nel 1948. Ma vista la difficoltà di darne un'analisi matematica soddisfacente, si concentrò successivamente su un modello astratto degli automi cellulari bidimensionali. Von Neumann tracciò comunque uno schema generale del suo automa auto-riproducente che anticipava alcuni concetti della attuale biologia cellulare, come quelli di traduzione e trascrizione (i due stadi fondamentali del processo di sintesi delle proteine).
Di fatto, il modello di Von Neumann anticipò le scoperte di Watson e Crick (1953) relative al funzionamento del DNA. In questa ricerca si avvalse di alcuni concetti introdotti da Turing, ad esempio nell'idea di un "costruttore universale", analogo alla "macchina di Turing universale".

Autoriproducibilità delle macchine

In campo informatico un interessante esempio di codici che si autoriproducono è dato per esempio dai virus informatici, definiti in effetti così da Cohen e Adleman:

> *A virus is a program that is able to infect other programs by modifying them to include a possibly evolved copy to itself.*

La forma più semplice di programma che si autoriproduce è però il *quine* (chiaramente dal nome del filosofo americano che studiò l'autoriferimento indiretto), ossia un programma che genera il proprio codice sorgente: quando il programma viene eseguito, stampa esattamente le istruzioni che il programmatore ha scritto.

Autoriproducibilità delle macchine

Nel libro di Hofstadter, creare un quine autoreplicante significa scrivere una frase la prima volta e poi scriverla una seconda volta, ma con le virgolette attorno. Ad esempio, considera la frase:

(A) stampa due copie della frase seguente di cui la seconda fra virgolette:

(B) “stampa due copie della frase seguente di cui la seconda fra virgolette:”

dove (A) è il codice e (B) i dati: applicando (A) a (B) si riproduce la frase stessa. In biologia cellulare, afferma Douglas Hofstadter, quello che abbiamo chiamato il ”codice” è la cellula, e i ”dati” sono il DNA della cellula: la cellula è in grado di creare una nuova cellula usando il DNA, e questo implica, tra altre cose, replicare il DNA stesso.

Autoriproducibilità delle macchine

Un quine si costituisce di due parti, una che chiamiamo il codice e una che chiamiamo i dati. Il codice utilizza i dati per stampare il codice, quindi utilizza i dati per stampare i dati.
Hofstadter ha sottolineato questa analogia con la biologia cellulare, istituendo un parallelismo fra cellule che si autoreplicano e programmi che si autoriproducono:

> *"... il codice di un quine è come una cellula, e i dati come fossero il DNA delle cellule: il DNA contiene tutte le informazioni necessarie per la replicazione delle cellule, d'altronde, quando una cellula usa il DNA per creare una nuova cellula, essa replica allo stesso tempo il DNA".*

In generale, in informatica, un quine è un algoritmo che riproduce il suo stesso codice sorgente (vedi alla voce "Quine(informatica)" di Wikipedia per esempi di "quines" in vari linguaggi).

Autoriproducibilità delle macchine

Cosa giustifica sul piano teorico i quines? Un'altra notevole applicazione della procedura formale di diagonalizzazione è proprio quella che consentì di confutare, da un punto di vista logico, l'obiezione circa la degenerazione nella costruzione delle macchine, attraverso i seguenti risultati dovuti a Stephen Cole Kleene (1938):
Teorema del punto fisso. Per ogni funzione totale computabile f esiste un numero e tale che:

$$C_e(x) = C_{f(e)}(x)$$

Teorema dei parametri. Consideriamo una funzione con due argomenti $A(x, y)$; allora esiste un algoritmo f che mi dà la possibilità di spostare il primo nel "programma", cioè:

$$A(x, y) = C_{f(x)}(y)$$

Autoriproducibilità delle macchine

Ecco che allora posso considerare la funzione $A(x, y) = x$, cioè che di due argomenti restituisce in output il primo; per il teorema dei parametri esisterà una funzione (totale computabile) f per cui:

$$A(x, y) = C_{f(x)}(y)$$

e per il teorema del punto fisso esisterà un numero e tale che:

$$A(e, y) = C_{f(e)}(y) = C_e(y) = e$$

Cioè a dire, questo programma stamperà sempre la propria descrizione, a prescindere dall'input.

Autoriproducibilità delle macchine

Autoriproducibilità delle macchine

Autoriproducibilità delle macchine

Autoriproducibilità delle macchine

Autoriproducibilità delle macchine

Autoriproducibilità delle macchine