

IL TEOREMA DI GÖDEL E LA FILOSOFIA DELLA MENTE

Aldo Antonelli
Università della California – Irvine

Siena, 24–25 maggio 2000

1 Il teorema di Gödel-Turing-Kleene

Kleene comincia la sezione §60 di *Introduction to metamathematics* considerando la questione se la matematica informale, e specialmente la teoria intuitiva dei numeri sia formalizzabile. Il classico teorema di Gödel dimostra che una particolare formalizzazione della teoria dei numeri è necessariamente incompleta, ma Kleene è interessato alla domanda più generale se vi siano formalizzazioni della teoria intuitiva che sono complete.

Per semplificare il problema si consideri la teoria di un singolo predicato $P(x)$ della teoria intuitiva. Affinché un dato sistema formale F sia una formalizzazione della teoria di P , è necessario che F contenga «oggetti formali» (cioè: formule) corrispondenti alle proposizioni $P(0), P(1), P(2), \dots$. Possiamo, senza perdita di generalità, designare tali oggetti formali come ϕ_0, ϕ_1, \dots .

È anche necessario che vi siano degli oggetti formali di natura finita, d_0, d_1, d_2, \dots , chiamati «dimostrazioni», ognuno dei quali è associato con una particolare formula. L'unica condizione è che la relazione $\mathcal{P}(n, d)$ (« d è una dimostrazione di ϕ_n in F ») deve essere *effettivamente decidibile*. Dato che sia agli enunciati sia alle dimostrazioni di F possono venire assegnati numeri di Gödel, il predicato \mathcal{P} può essere identificato, a tutti gli effetti, con un predicato decidibile R che prende come argomenti numeri naturali.

Veniamo ora all'argomentazione di Kleene. È opportuno notare che in quanto segue si presuppongono nozioni elementari di teoria della ricorsività classica, e in particolare: la definizione di macchina di Turing; l'assegnamento di numeri di Gödel alle macchine di Turing come loro indici; il teorema dei parametri e il teorema di ricorsione; il lemma diagonale di Gödel.¹ In particolare si sa che esiste un predicato ricorsivo $T(x, y, z)$ che dice che z è una computazione completa della macchina di Turing con indice x su argomento y . Assumiamo anche la cosiddetta «tesi di Church»: TESI DI CHURCH Ogni funzione effettivamente calcolabile (e in particolare ogni predicato effettivamente decidibile) è Turing-computabile.

LEMMA Dato un predicato ricorsivo totale $R(x, y)$, esiste un numero p tale che

$$\exists y R(p, y) \neq \forall y \sim T(p, p, y).$$

Quindi i due predicati (in x): $\exists y R(x, y)$ e $\forall y \sim T(x, x, y)$ non possono coincidere, e il predicato $\forall y \sim T(x, x, y)$ non può essere espresso nella forma duale $\exists y R(x, y)$. Poiché il caso in cui il quantificatore $\exists y$ è vacuo non viene escluso, si ha in particolare che il predicato $\forall y \sim T(x, x, y)$ non è ricorsivo (decidibile).

¹Nella forma: per ogni predicato $B(y)$ c'è un enunciato θ tale che $\theta \leftrightarrow B(\ulcorner \theta \urcorner)$ è dimostrabile.

Dimostrazione Si usa un procedimento diagonale simile a quello di Cantor. Sia $R(x, y)$ un predicato decidibile. Allora c'è una macchina di Turing M tale che

$$R(x, y) \equiv M(x, y) = 1.$$

Sia p l'indice di una macchina M_p che, dato un argomento x , simula in successione le computazioni $M(x, 0), M(x, 1), M(x, 2), \dots$, finché per qualche y si ha $M(x, y) = 1$; se per nessun y la condizione è vera, allora $M_p(x) \uparrow$. Chiaramente, per ogni x , si ha:

$$\exists y R(x, y) \equiv \exists y T(p, x, y)$$

(la condizione di destra dice che y è una computazione completa di $M_p(x)$). In particolare, per $x = p$:

$$\exists y R(p, y) \equiv \exists y T(p, p, y)$$

da cui, usando la logica classica ($A \not\equiv \sim A$ e le equivalenze dei quantificatori), si ha immediatamente l'enunciato del lemma.

TEOREMA DI GÖDEL GENERALIZZATO (KLEENE 1943). Non c'è alcun sistema formale corretto e completo per il predicato $\forall y \sim T(x, x, y)$. Altrimenti detto, se un sistema formale è corretto per $\forall y \sim T(x, x, y)$ allora non è completo.

Dimostrazione. Procedendo in modo più generale, supponiamo che F sia un sistema formale corretto e completo per un predicato $P(x)$ dei numeri naturali. Questo significa che il linguaggio di F contiene enunciati ϕ_n tali che per ogni n , $P(n)$ vale se e solo se $F \vdash \phi_n$, e inoltre il numero di Gödel di ϕ_n può essere determinato uniformemente (cioè, ricorsivamente) in n . Siccome il predicato

$$P(p, \phi) \equiv p \text{ è una prova di } \phi$$

è decidibile, lo è anche il predicato $Q(n, m)$: « n è il numero di Gödel di una dimostrazione in F di un enunciato ϕ_n avente numero di Gödel m ».

Usando la tesi di Church, Q è equiestensionale a un predicato ricorsivo (Turing-computabile) $R(x, y)$. Ne segue che

$$\exists y R(x, y) \equiv P(x).$$

Quanto detto vale per un qualsiasi predicato P , e in particolare per il predicato $\forall y \sim T(x, x, y)$, ma questo contraddice il precedente lemma.

COROLLARIO Se il sistema formale F contiene almeno l'aritmetica di Robinson, Q , ed è formalmente coerente, allora non è completo.

Basta osservare che se F contiene Q allora è coerente se e solo se è corretto per il predicato $\forall y \sim T(x, x, y)$. Infatti, se F non è corretto per tale predicato, allora esiste almeno un indice i tale che $M_i(i) \downarrow$, ma $F \vdash \forall y \sim T(i, i, y)$; d'altra parte, se $Q \subseteq F$, allora esiste un q tale che $F \vdash T(i, i, q)$ e quindi $F \vdash \sim \forall y \sim T(i, i, y)$, cosicché F è incoerente.

Conversamente, se F è incoerente, allora $F \vdash \phi$ per ogni ϕ ; se i è (ad es.) l'indice della funzione costante uguale a 0, allora $F \vdash \forall y \sim T(i, i, y)$ mentre $M_i(i) \downarrow$, e quindi F non è corretto.

È importante osservare che le dimostrazioni di questa sezione possono essere completamente formalizzate in un sistema debole, come ad esempio l'aritmetica di Peano (PA).

2 L'argomentazione di Lucas-Benacerraf

J.R. Lucas, un filosofo di Harvard noto per le sue posizioni controverse, nel 1961 pubblica un articolo intitolato *Minds, Machines, and Gödel*, in cui sviluppa un'argomentazione tesa a dimostrare come, alla luce del teorema di Gödel, non sia possibile sostenere che la mente umana sia, in alcun senso, identificabile con una macchina.

Siccome le «macchine cibernetiche» costituiscono realizzazioni concrete di sistemi formali, il teorema di Gödel implica che, data una macchina in grado di eseguire semplici operazioni aritmetiche senza mai produrre una contraddizione, è possibile esibire una formula che non potrà mai comparire fra quelle che la macchina produce.

Secondo Lucas, questo fatto *puramente matematico* ha conseguenze importanti per la filosofia della mente:

Qualsiasi modello meccanico della mente deve includere un meccanismo capace di produrre verità aritmetiche, poiché questo è un compito che la mente è in grado di eseguire: infatti è facile produrre modelli meccanici che per molti aspetti possono produrre verità aritmetiche molto meglio di quanto non possa fare la mente. Ma c'è un singolo aspetto per cui tali modelli non possono fare meglio: e cioè per ogni macchina c'è un enunciato vero che essa non può produrre, mentre una mente può [riconoscerne la verità]. Ne segue che una macchina non può essere un modello completo e adeguato della mente (Anderson [1964], p. 47).

La mente può riconoscere la verità di enunciati fuori della portata delle «macchine cibernetiche». Si noti però la possibile ambiguità fra *tutti e qualsiasi* (*every* e *any*, come già osservato da Russell).

Anche aggiungendo, per «forza bruta» l'enunciato indecidibile alla lista di verità aritmetiche prodotte dalla macchina, si ottiene un'altra macchina, che a sua volta avrà un'altro enunciato indecidibile, e così via.

Charles Chihara nel 1972 pubblica un articolo «On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results» in cui riprende un'argomentazione originariamente formulata da Putnam [1961] contro Newman & Nagel [1958], ma che si applica anche all'argomentazione di Lucas.

Putnam osserva che, data una macchina M , è possibile trovare un enunciato indecidibile U tale che si può dimostrare (ad esempio nell'aritmetica di Peano) che

- (1) Se M è coerente, allora U è vero ma non decidibile da M .

Per poter risurre all'assurdo l'ipotesi che la mente umana sia identificabile con M , è necessario poter dimostrare la coerenza di M , il che potrebbe non essere affatto banale.

Lucas ribatte che la coerenza di M può essere in effetti determinata empiricamente (cfr. Lucas [1968]): se la mente fosse in effetti un sistema formale contraddittorio, saremmo pronti ad asserire qualsiasi enunciato (p. 157) e nessun pensiero sarebbe possibile (p. 158). Inoltre, la posizione che la mente sia incoerente è pur sempre una posizione di «completo nichilismo matematico».

Un'argomentazione simile a quella di Lucas, ma con un notevole «incremento in rigore e chiarezza» (Chihara [1972], p. 511) viene proposta da Paul Benacerraf [1967]. Si consideri il seguente insieme di numeri naturali:

$$S = \{x \mid x \text{ è il numero di Gödel di un enunciato aritmetico che io posso dimostrare}\}.$$

Secondo Benacerraf, il teorema di Gödel riduce all'assurdo l'ipotesi che S sia ricorsivamente enumerabile e che io possa scoprire l'indice o il programma di una macchina di Turing in grado di enumerare S (insieme ad altre ipotesi ausiliarie). Quindi, o S non è ricorsivamente enumerabile, o è impossibile che io ne possa scoprire il programma.

In risposta a queste argomentazioni di Lucas e Benacerraf, Chihara indica come la dimostrabilità di un enunciato dipende non solo dalle «procedure inferenziali», ma anche dal bagaglio di conoscenze «di sfondo» di chi deve effettuare la dimostrazione, e tali conoscenze possono essere determinate solo empiricamente.

Cioè, anche se per ogni particolare insieme di conoscenze o informazioni α , l'insieme S risulta ricorsivamente enumerabile relativamente ad α , non è detto che tale enumerazione sia uniforme in α (tale cioè che esiste un'unica procedura meccanica che, dato α , genera il programma della macchina che enumera S). Se ciò non fosse il caso, la derivazione della contraddizione verrebbe bloccata, e con essa ogni pretesa refutazione del meccanicismo.

Più in particolare: assumendo che S sia ricorsivamente enumerabile, ne segue che c'è un sistema formale assiomatico F tale che $S = \{\phi : F \vdash \phi\}$. Considerando l'enunciato indecidibile $G(F)$ per F , si ha che $F \not\vdash G(F)$, ma anche che $G(F) \in S$ (poiché io capisco la dimostrazione del teorema di Gödel e so che per ogni F l'enunciato $G(F)$ è indecidibile in F e vero). Quindi, $S \neq \{\phi : F \vdash \phi\}$.

Tuttavia, anche relativizzando il sistema formale a un insieme di conoscenze di sfondo α , le cose non sembrano cambiare molto. Infatti, se α è finito (o addirittura ricorsivamente enumerabile), allora può essere incorporato nella struttura assiomatica di F , e ci si riduce al caso precedente. Se invece α non è nemmeno ricorsivamente enumerabile, allora resta da spiegare com'è che abbiamo accesso a un insieme altamente non computabile di informazioni, specialmente dato che la nostra esperienza (sia come individui che come specie) è comunque finita. Insomma, questa mossa di Chihara, di relativizzare F ad α non sembra condurre da nessuna parte.

Recensendo Lucas [1970], Dennett [1972] osserva che il teorema di Gödel non concerne, così come è formulato, né oggetti materiali né menti, e quindi chiunque ne voglia «estrarre conseguenze anti-meccaniciste» debba chiudere lo iato che separa le verità sui sistemi formali dalle verità sui movimenti degli oggetti fisici.

Ora, è possibile «interpretare» un oggetto *come se* «seguisse» le istruzioni del programma della macchina. Se tale interpretazione è possibile, allora l'oggetto fisico in questione costituisce una realizzazione di quella particolare macchina di Turing o sistema formale.

L'osservazione cruciale qui è che non solo ogni data macchina astratta o sistema formale può avere una moltitudine di realizzazioni, ma che «qualsiasi oggetto fisico può essere *simultaneamente* interpretato come una varietà di macchine di Turing diverse» (p. 528, enfasi aggiunta).

C'è naturalmente la possibilità che per ciascun oggetto fisico esista una descrizione privilegiata, «a grana massimamente fine». In tal caso, ciascun oggetto fisico dovrebbe essere interpretato come un'unica macchina di Turing, per la quale si darebbero precise limitazioni logiche dovute al corrispondente teorema di Gödel. Come osserva Dennett (p. 529), anche se vi fosse una tale interpretazione privilegiata (certamente un'assunzione pesante), essa non sarebbe ancora sufficiente a stabilire il punto di vista anti-meccanicista di Lucas:

Preso in tale senso, il teorema di Gödel ha implicazioni rispetto alle capacità dimostrative, diciamo, delle querce: nonostante ogni singola quercia con lo stormire delle fronde e il cadere delle ghiande possa dimostrare innumerevoli teoremi (!), ve ne è uno che essa non può dimostrare: il suo enunciato di Gödel (p. 530).

Potrebbe essere interessante confrontare questo passaggio di Dennett con il famoso brano delle *Ricerche filosofiche* di Wittgenstein, dove le urla e calci dei due giocatori sono interpretate come mosse del gioco degli scacchi.

3 L'argomentazione «induttiva» di Putnam

Putnam [1963] dà un'argomentazione diagonale relativo alla logica «induttiva» invece che su quella deduttiva, come nel teorema di Gödel. Centro dell'analisi è il concetto di *misura induttiva*, cioè la concezione *quantitativa* del «grado di conferma» di una teoria o ipotesi. L'articolo si sviluppa in tre fasi:

1. Formulazione delle *condizioni di adeguatezza* che un qualsiasi procedimento induttivo deve soddisfare.
2. Dimostrazione che nessun procedimento induttivo quantitativo (misura) può soddisfare queste condizioni.
3. Dimostrazione che vi sono procedimenti induttivi non quantitativi che possono invece soddisfare le condizioni di adeguatezza.

Solo i primi due punti sono pertinenti al presente discorso.

Prima di tutto bisogna osservare che l'obbiettivo polemico di Putnam è la nozione carnapiana di una funzione di conferma, o «*c-function*», cioè di una funzione che assegna probabilità $p \in [0, 1]$ a certe ipotesi generali dato un segmento iniziale di dati. In quanto tale, l'argomentazione di Putnam non viene subito applicata alla filosofia della mente, ma in almeno un caso (Putnam [1985]) Putnam stesso indica il possibile parallelo fra quest'argomentazione e le difficoltà gödeliane di una concezione meccanicista della mente.

Cominciamo con l'identificare un linguaggio \mathcal{L} che assumiamo *sufficientemente ricco* da rappresentare l'ordinamento spazio-temporale degli individui, oltre che l'aritmetica elementare al prim'ordine.

DEFINIZIONE Un'ipotesi h è *effettiva* se e solo se:

1. h è esprimibile in \mathcal{L} ;
2. ogniqualvolta h implica $M(c)$ (per M un predicato «osservabile» di \mathcal{L}), allora $h \rightarrow M(c)$ è *dimostrabile* in \mathcal{L} ;
3. h è equivalente a un insieme di enunciati atomici $M(c)$ o loro negazioni $\sim M(c)$.

PRIMA CONDIZIONE DI ADEGUATEZZA Se h è un'ipotesi che è effettiva e vera, allora la «conferma specifica» (*instance confirmation*) di h tende a 1 man mano che un numero sempre maggiore di individui viene esaminato. Questa condizione può essere indebolita un po' richiedendo che la conferma specifica rimanga $> \frac{9}{10}$ oppure $> \frac{1}{2}$.

Quindi la condizione di adeguatezza, dice che se un'ipotesi generale è vera, prima o poi questo verrà scoperto nel sistema.

Un punto terminologico: siccome per Carnap il grado di conferma di un'ipotesi universale è sempre 0, ne *I fondamenti logici della probabilità* egli introduce la conferma specifica di un'ipotesi generale h , definita come la conferma dell'ipotesi particolare che l'individuo seguente che viene esaminato sia un'esemplificazione di h . Putnam critica questa nozione notando che la conferma specifica è compatibile con un numero qualsiasi di eccezioni e inoltre non cattura adeguatamente l'uso pre-teorico degli scienziati secondo è perfettamente sensato parlare della conferma o disconferma di ipotesi universali.

Torniamo ora all'argomentazione principale. Per semplificare assumiamo che esista solo una dimensione discreta e con un punto iniziale. Assumiamo cioè che le locazioni spazio-temporali abbiano

lo stesso tipo d'ordine dei numeri naturali. Gli individui corrispondono alle posizioni x_0, x_1, x_2, \dots . Sia M una proprietà osservabile degli individui, ad es., $M(x) = x$ è rosso. Supponiamo di avere una funzione (misura) c che dà il grado di conferma di un'ipotesi data una condizione, in analogia al concetto di probabilità condizionale. In altre parole, $c(h, e)$ è il grado di conferma dell'ipotesi h data un certo «corpo di evidenza» e . Inoltre, senza perdita di generalità si può assumere che $c(h, e)$ sia un valore reale nell'intervallo $[0, 1]$.

SECONDA CONDIZIONE DI ADEGUATEZZA Per ogni predicato M

$$\forall k \exists m \ c(M(x_{k+m+1}), \bigwedge_{i=1}^m M(x_{k+i})) > \frac{1}{2}.$$

Questa è una versione del requisito che il grado di conferma di $M(x)$ deve alla fine convergere al valore della frequenza relativa di M nel campione in questione. Più in particolare, la condizione dice che data l'ipotesi $\forall x M(x)$, non importa quante eccezioni possano esserci nei primi k casi, c'è sempre un numero m così grande che se tutti gli m individui x_{k+1}, \dots, x_{k+m} hanno M , allora la probabilità che anche x_{k+m+1} abbia M è maggiore di $\frac{1}{2}$. Qualsiasi numero iniziale di eccezioni all fine non conta per la conferma di un'ipotesi.

TEOREMA Non c'è nessuna funzione computabile c che soddisfa entrambe le condizioni di adeguatezza.

Dimostrazione. Si segue un *procedimento diagonale*. Si osserva prima di tutto che esiste un insieme infinito $C = \{n_1, n_2, n_3 \dots\}$ di numeri naturali con la seguente proprietà:

(*) Per ogni i : se per ogni j t.c. $n_{i-1} < j < n_i$ il grado di conferma di $M(x_j)$ è maggiore di $\frac{1}{2}$, allora anche il grado di conferma di $M(x_{n_i})$ è maggiore di $\frac{1}{2}$.

Per vedere che tale insieme C esiste definiamo n_i per induzione su i , usando la seconda condizione di adeguatezza. Per $i = 1$, ponendo $k = 0$ si ha che deve esserci un numero m che se i primi m individui hanno M allora la probabilità (conferma) di $M(x_m)$ è maggiore di $\frac{1}{2}$. Si definisce $n_1 = m$. Supponendo ora definito n_i , si usa ancora la seconda condizione di adeguatezza con $k = n_i$ ottenendo un m t.c. se tutti gli x_j strettamente compresi fra x_{n_i} e x_{n_i+m+1} hanno M con probabilità maggiore di $\frac{1}{2}$, allora anche x_{n_i+m+1} ha M con probabilità maggiore di $\frac{1}{2}$. Si pone allora $n_{i+1} = n_i + m$.

Si considera ora l'ipotesi h :

$$\forall n \ M(x_n) \leftrightarrow n \notin C.$$

Si osserva a questo punto (usando l'ipotesi che c è computabile) che C è *decidibile* (ricorsivo): infatti, C può essere enumerato in modo effettivo e in ordine ascendente: $n_1 < n_2 < n_3 < \dots$. Quindi, poiché \mathcal{L} contiene l'aritmetica, h è *effettiva* nel senso della definizione data sopra.

Ora supponendo che h sia di fatto *vera*, si ottiene ora una contraddizione con la prima condizione di adeguatezza. Infatti, se h è vera, allora per la prima condizione di adeguatezza

$$\lim_{n \rightarrow \infty} c(M(x_n) \leftrightarrow n \notin C) = 1$$

Facciamo le seguenti trasformazioni: riscriviamo il bicondizionale come una congiunzione di implicazioni; usiamo il fatto che la probabilità di una congiunzione ha un limite se e solo se le probabilità dei congiunti hanno lo stesso limite; e infine riscriviamo la probabilità dell'implicazione come probabilità condizionale, otteniamo:

$$\lim_{n \rightarrow \infty} c(\sim M(x_n), n \in C) = 1.$$

Otteniamo la contraddizione desiderata facendo vedere che ci sono un numero infinito di n tali che $c(\sim M(x_n), n \in C) < \frac{1}{2}$. Infatti, se h è vera allora per ogni j strettamente fra n_i e n_{i+1} si ha

$$c(M(x_j)) > \frac{1}{2},$$

e quindi per la seconda condizione di adeguatezza

$$c(M(x_{n_{i+1}})) > \frac{1}{2},$$

cioè $c(\sim M(x_{n_{i+1}})) < \frac{1}{2}$. La conclusione desiderata segue immediatamente.

Un altro modo di considerare l'argomentazione appena data è che se c è effettivamente computabile, allora c'è un'ipotesi effettiva h che, se vera, non è possibile imparare. La prima condizione dice infatti che c eventualmente ci porta a imparare ogni ipotesi vera. Tuttavia noi possiamo arrivare a scoprire che h è vera, ad esempio considerando la definizione di C , quindi noi non usiamo c . Data la generalità di c sembra possibile concludere che noi non usiamo alcuna misura ricorsiva.

In [1985], Putnam trae le conclusioni concernenti la filosofia della mente in modo simile:

In un articolo pubblicato già nel 1963, ho dimostrato che se la relazione di conferma di una logica induttiva \mathbf{I} è ricorsiva, allora data la descrizione computazionale di tale relazione, è possibile produrre in modo effettivo un'altra logica induttiva \mathbf{I}' che è più adeguata di \mathbf{I} , nel senso che usando \mathbf{I}' è possibile scoprire regolarità che \mathbf{I} non sarebbe in grado di scoprire, anche se i dati procedessero all'infinito ... La morale è che se \mathbf{I} fosse tale che si può scoprirne la correttezza per mezzo di un'argomentazione intuitivamente valida (dimostrativa o non-dimostrativa), allora si sarebbe giustificati in usare \mathbf{I}' e non solo \mathbf{I} e quindi a considerare una certa regolarità come confermata ... ma solo confermata nel senso di \mathbf{I}' . Poiché la conclusione che siamo giustificati nello scommettere su tale regolarità è intuitivamente corretta, e \mathbf{I} non registra tale conclusione, allora questo già mostra come \mathbf{I} non possa essere una formalizzazione *completa* della nostra competenza induttiva prescrittiva. (pp. 145–46).

4 L'argomentazione di Penrose

L'argomentazione principale proposta da Penrose [1989, 1994] è del tutto parallela al teorema di Gödel-Turing-Kleene. Ovviamente l'argomentazione in sé, in quanto dimostrazione matematica (corretta), è inattaccabile; è tutto un altro discorso quando se ne vogliono trarre conseguenze filosofiche.

Penrose comincia con il supporre (p. 73) che esista una procedura effettiva A che incorpora tutte le procedure a disposizione della comunità matematica allo scopo di dimostrare teoremi. In particolare, A permette di dimostrare teoremi della forma «la computazione $C(n)$ non termina». Gli enunciati di questa forma sono detti Π_1 poiché sono equivalenti a enunciati della forma «per ogni m : m non è il codice di una computazione completa del programma C con argomento n ».

Un'importante assunzione relativa ad A è che A deve essere *corretta*, nel senso che A non dà mai risposte sbagliate. Se A permette di stabilire che la computazione $C(n)$ non termina, allora $C(n)$ non termina. Inoltre, le diverse procedure computazionali a un argomento, in quanto rappresentabili come parole su di un alfabeto finito (ad esempio, programmi per calcolatore), possono essere enumerate esaustivamente in una lista:

$$C_1, C_2, C_3, \dots, C_n, \dots$$

Si può assumere che A abbia la seguente caratteristica: se A termina con argomenti q e n , allora $C_q(n)$ non termina. Ciò è possibile perché se A dovesse fermarsi con *output* «la computazione termina», si può sempre modificare A in modo da metterla in un *loop* infinito.

Con questa ipotesi, l'arrestarsi della computazione $A(q, n)$ costituisce una «dimostrazione» dell'enunciato « $C_q(n)$ non termina», e siccome A è per ipotesi corretta, abbiamo che

(1) Se $A(q, n)$ termina, allora $C_q(n)$ non termina.

In particolare, per $q = n$, si ha che

(2) Se $A(n, n)$ termina, allora $C_n(n)$ non termina.

A questo punto si osserva che la procedura $A(n, n)$ dipende solo più dall'unico parametro n , e quindi deve a sua volta essere una delle procedure $C_1, C_2, C_3 \dots$, ad esempio C_k . Ne segue che

(3) $C_k(k)$ termina se e solo se $A(k, k)$ termina.

Da (2), con $n = k$, si ha

(4) Se $A(k, k)$ termina, allora $C_k(k)$ non termina.

e da (3) e (4),

(5) Se $C_k(k)$ termina, allora $C_k(k)$ non termina.

Questo implica che la computazione $C_k(k)$ non può terminare. Ma allora, dato (3), neanche la computazione $A(k, k)$ può terminare. Ne segue che la procedura A è incapace di determinare la mancata terminazione di $C_k(k)$.

D'altro canto, noi abbiamo appena dimostrato matematicamente che la computazione $C_k(k)$ non può terminare, e quindi A non può contenere o rappresentare la totalità delle procedure dimostrative a disposizione della comunità matematica. In particolare, Penrose ritiene che sia stata dimostrata la seguente proposizione

\mathcal{G} I matematici umani per determinare verità matematiche non usano una procedura riconosciuta come corretta (*a knowably sound algorithm*).

Infatti, se i matematici usassero una tale procedura A , essi saprebbero in particolare che A è corretta, e quindi che esiste una procedura computazionale di cui non può predire la non-terminazione. Ma essendo i matematici giunti alla conclusione che la computazione non termina, devono poter trascendere il potere predittivo e dimostrativo di A , che quindi non può incorporare tutti i metodi dimostrativi dei matematici.

Osserviamo che l'argomentazione dipende in modo cruciale dal fatto che i matematici non solo (i) usano una procedura corretta, ma anche (ii) che *sanno* di usare una procedura corretta. Entrambe le parti sono necessarie per l'argomentazione. Infatti, se (ii) venisse meno, avremmo che il teorema di Gödel ancora implica l'esistenza di un enunciato vero indimostrabile in F , ma noi non potremmo applicare il teorema a F concludendo che l'enunciato è indimostrabile e *vero*. L'unica base che abbiamo per riconoscere la verità dell'enunciato indimostrabile è il teorema di Gödel; venendo a mancare questo, la conclusione che noi conosciamo un enunciato inaccessibile a F viene a cadere.

In modo simile, se la condizione (i) venisse a mancare, cioè se l'ipotesi fosse solo che i matematici *credono* di usare una procedura corretta, allora il teorema di Gödel non si applica a F (in quanto F potrebbe essere incorretto), e di nuovo non c'è alcuna garanzia di scoprire la verità di un enunciato inaccessibile a F . Vedremo dopo che Penrose pensa di poter indebolire (ii) eliminando la qualifica «corretta».

5 I commentatori

L'argomento gödeliano di Penrose ha naturalmente suscitato molte reazioni. Alcune di queste si sono concentrate sugli aspetti computazionali dell'argomento gödeliano. Queste repliche si possono trovare principalmente in due raccolte particolari. Nel 1990 Penrose pubblica su *Behavioral and Brain Sciences* un riassunto dell'*Emperor's New Mind* [1989] (cfr. Penrose [1990]), riassunto accompagnato da una serie di risposte e recensioni di eminenti scienziati e filosofi della mente e da una replica da parte di Penrose.

Il dibattito fu continuato nel 1993, quando *Behavioral and Brain Sciences* raccoglie altre recensioni pubblicate con un'ulteriore replica di Penrose. Altri interventi sono stati apparsi nel 1995 sulla rivista elettronica *Psyche*, pubblicata dalla Monash University in Australia² Qui mi limito a segnalare i contributi principali.

5.1 Martin Davis

Nei suoi due *commentaries* su Penrose, Davis mostra come la posizione anti-meccanicista in filosofia della mente sostenuta sia da Gödel sia da Penrose non sia una conseguenza del teorema di Gödel (e infatti Gödel, a differenza di Penrose, è sempre stato molto attento a non sostenere che lo fosse). L'argomento gödeliano mostra che, per ogni sistema formale F , l'intuizione in base alla quale si riconosce come vero l'enunciato «Se F è corretto, allora $G(F)$ è vero» non può essere incorporata nel sistema F stesso.

Ora, secondo Davis, è fondamentalmente scorretto da parte di Penrose chiamare tale enunciato una «intuizione» (*insight*), dato che si tratta di un *teorema*, e nemmeno dei più difficili. La ragione per cui tale teorema non può essere incorporato è che per ipotesi si è ristretto F a produrre enunciati Π_1 , e l'enunciato «Se F è corretto, allora $G(F)$ è vero» non ha tale forma. Tuttavia, come si è visto, l'enunciato è facilmente dimostrabile in ogni sistema formale sufficientemente potente da rappresentare l'aritmetica elementare, come ad esempio l'aritmetica di Peano. Ad esempio, si può dimostrare nell'aritmetica di Peano che se la teoria degli insiemi di Zermelo-Fraenkel è coerente allora l'enunciato di Gödel per tale teoria non è dimostrabile nella teoria stessa.

Tale affermazione implica che se la teoria di Zermelo-Fraenkel è corretta allora l'enunciato di Gödel è vero. più in particolare, l'argomentazione è la seguente: dato un sistema formale F si consideri la collezione T degli enunciati Π_1 (nel linguaggio di F) che sono veri. Tale collezione non è ricorsivamente enumerabile, mentre la collezione degli enunciati Π_1 che sono dimostrabili in F è ricorsivamente enumerabile. Ne segue che se F è corretto, allora c'è un enunciato Π_1 vero che non è dimostrabile. Non solo: la collezione T è un insieme produttivo, cioè tale che esiste una procedura meccanica che, dato un qualsiasi sottoinsieme ricorsivamente enumerabile di T , produce un enunciato in T che non è nel sottoinsieme dato (tale enunciato è un «testimone» del fatto che T non è ricorsivamente enumerabile; se un insieme è produttivo, tali testimoni possono essere generati in modo meccanico).

Un altro contributo interessante di Martin Davis [1993] aiuta a capire quale fosse la posizione di Gödel stesso nei confronti della meccanizzabilità della mente. Nella famosa Gibbs lecture del 1951, Gödel fece la seguente famosa affermazione:

D'altro canto, sulla base di ciò che è stato stabilito finora, è possibile che esista (e che possa persino essere empiricamente scoperta) un macchina per la dimostrazione di teoremi che è di fatto equivalente all'intuizione matematica, ma che non lo è dimostrabilmente, cosiccome non si

²Si veda <http://psyche.cs.monash.edu.au/psyche>.

può dimostrare che essa produce solo teoremi corretti della teoria finitaria dei numeri. (Citato in Wang [1974], p.324)

Gödel certamente non condivideva la concezione meccanicista della mente, ma era anche cosciente del fatto che tale concezione è perfettamente compatibile con il platonismo matematico. La tesi che Gödel presenta nella Gibbs lecture ha la forma di una disgiunzione: o la mente umana è equivalente a una macchina di Turing, e quindi esistono enunciati Π_1 assolutamente indecidibili, oppure la mente ha un aspetto non meccanizzabile, e quindi dobbiamo accettare una posizione di tipo vitalistico. Entrambe le conclusioni sono compatibili con il platonismo matematico: infatti, l'argomento di Gödel è precisamente che entrambe le alternative comportano l'accettazione di una forma di platonismo matematico. Si noti però che Gödel è ben cosciente che il suo platonismo non è in contraddizione con il meccanicismo in filosofia della mente.

5.2 Feferman

Un'analisi puntuale (e per molti versi impietosa) dell'esposizione tecnica di Penrose si può trovare in Feferman [1995]. La posizione di Feferman è interessante perchè non viene certo da un sostenitore della teoria computazionale della mente, tantomeno da un esponente dell'intelligenza artificiale forte. Feferman si dice convinto della «estrema implausibilità» del modello computazionale, ma afferma anche che l'argomento gödeliano di Penrose non riesce a stabilire in modo conclusivo tale punto di vista. Ciò è dovuto in gran parte all'alto numero di imprecisioni nell'esposizione matematica di Penrose.

Feferman nota subito (§3.3) che ci sono due diverse nozioni di correttezza (per un sistema formale) che Penrose utilizza in *Shadows of the Mind*. A pp. 74-75 Penrose spiega che se un sistema formale è corretto e dimostra che una certa computazione non termina, allora in effetti tale computazione non termina. Il contesto rende chiaro che qui Penrose ha in mente una nozione di correttezza per enunciati Π_1 , quella esposta nella prima sezione e equivalente alla semplice coerenza. Al tempo stesso, a pp. 90-92, Penrose nota che se un sistema formale è corretto, allora è certamente ω -coerente. Feferman osserva che, mentre come si è visto la semplice coerenza è equivalente alla correttezza per enunciati Π_1 , la nozione di ω -coerenza è più forte della semplice coerenza, e quindi le due nozioni di correttezza non coincidono.

Inoltre, Penrose si riferisce all'affermazione che un dato sistema formale F è ω -coerente mediante l'abbreviazione $\Omega(F)$, e procede poi a enunciare il secondo teorema di incompletezza di Gödel nella forma: se F è ω -coerente allora $\Omega(F)$ non è un teorema di F . In realtà, il secondo teorema di Gödel dice che se F è semplicemente coerente, allora « F è coerente» non è un teorema di F . Siccome $\Omega(F)$ implica (dimostrabilmente in F) la coerenza di F , a fortiori $\Omega(F)$ non può essere un teorema di F . L'ipotesi della ω -coerenza di F è necessaria solo se si vuole dimostrare che « F non è coerente» (o non- $G(F)$, equivalentemente) non è un teorema di F .

Altre inesattezze sono più sostanziali, e forniscono ulteriori esempi della *slapdash scholarship* di Penrose già lamentata da Davis [1993]. Ad esempio, in §3.8 Feferman affronta un punto relativo a p. 108 di *Shadows*. Lì Penrose considera i due sistemi formali ottenuti aggiungendo ad un sistema F rispettivamente l'enunciato $G(F)$ e la sua negazione. Penrose ritiene che se F è coerente allora entrambi i sistemi sono coerenti. In realtà, osserva Feferman, questo è vero solo per il secondo di tali sistemi; la coerenza del primo richiede la ω -coerenza di F , in mancanza della quale è possibile esibire un controesempio.

La conclusione (in §3.12) è che queste inesattezze suggeriscono che Penrose stia pericolosamente esplorando aree lontane dalle proprie competenze. Tuttavia Feferman non ritiene che tali errori e imprecisioni, di per se stessi, implicino che l'argomento principale di Penrose sia scorretto.

È quindi importante che Feferman si preoccupi di ricostruire l'argomento gödeliano in modo da evitare scorrettezze e imprecisioni. Prima di tutto, occorre riconoscere che l'argomento gödeliano richiede soltanto la prima metà del teorema di Gödel, cioè che se F è coerente, allora $G(F)$ non è dimostrabile in F . La seconda metà, che se F è ω -coerente allora la negazione di $G(F)$ non è dimostrabile, è del tutto estranea agli scopi di Penrose (nelle parole di Feferman, a *red herring*). Siccome F è coerente se e solo se F è corretto per enunciati Π_1 , questo è l'unico senso di «correttezza» di cui Penrose ha bisogno.

Feferman continua osservando che è il modello della ricerca esaustiva delle dimostrazioni che è sbagliato come modello della pratica matematica. Certamente i matematici non partono da una proposizione p , generando dimostrazioni finché una prova di p (o di non- p) viene scoperta. Le dimostrazioni sono ottenute attraverso una «meravigliosa combinazione di ragionamento euristico, intuizione e ispirazione (sulla base naturalmente delle conoscenze e esperienze precedenti)» (§4.2). Ne segue che «il pensiero matematico, cosiccome esso viene effettivamente prodotto, non è meccanico» (§4.2).

Come Feferman nota, da ciò non segue che il pensiero matematico non possa essere «ri-rappresentato», post-factum, in termini di sistemi formali. Negare ciò, come fa Penrose, è adottare un punto di vista in ultima analisi improduttivo. Ad esempio, come si è visto, la correttezza di F per enunciati Π_1 è tutto ciò di cui Penrose ha bisogno. Tuttavia egli continua a insistere sulla nozione più generale di correttezza per enunciati qualsiasi, e a collegare tale nozione con le proprie simpatie platonistiche in filosofia della matematica. Ma ciò non è affatto necessario. Ad esempio, potrebbero esserci altri metodi, oltre a una nozione globale di verità per gli enunciati di F , con cui si può riconoscere la verità dell'enunciato indecidibile $G(F)$. Feferman qui ha in mente metodi essenzialmente di teoria delle dimostrazioni, da lui stesso studiati e sviluppati. Con tali metodi, è a volte possibile ridurre la coerenza di un sistema formale adeguato per la formalizzazione di gran parte della matematica classica alla coerenza di sistemi formali costruttivi (come ad esempio l'aritmetica di Peano).

5.3 Dennett

Un'altra reazione a *Shadows* che vogliamo considerare è invece quella di Dennett [1995]. In *Darwin's Dangerous Idea*, e più precisamente in un capitolo intitolato «The Emperor's New Mind, and Other Fables», Dennett procede a smontare, per l'ennesima volta, l'argomentazione gödeliana. Dennett stesso confessa di essere sorpreso del fatto che alla base dell'argomentazione gödeliana, così come usata da Penrose prima e Lucas poi, vi sia un errore tutto sommato semplice. Certamente non una svista che ci si sarebbe aspettati dal *Rouse Ball Professor of Mathematics* all'Università di Oxford. Il punto è che il teorema di Gödel ci dice qualcosa solo su una sottoclasse estremamente ristretta della totalità degli algoritmi, e cioè su quegli algoritmi espressamente progettati (modulo un certo schema interpretativo) per produrre ed esibire dimostrazioni di teoremi matematici in un dato sistema formale corretto. Oltre a questi vi sono, in principio, innumerevoli altri algoritmi, alcuni dei quali, benché non dimostrabilmente corretti, possono tuttavia produrre teoremi in modo del tutto soddisfacente, pari ad esempio a quello di un vero matematico. Sono questi, e solo questi, gli algoritmi che l'intelligenza artificiale si è posta lo scopo di scoprire.

L'argomentazione gödeliana per Dennett è parallela alla seguente (p. 440): x ha una straordinaria abilità a giocare a scacchi e dare scacco matto; non esiste (per ragioni di complessità computazionale) alcun algoritmo praticabile per garantire la vittoria agli scacchi; quindi la capacità di x non può essere spiegata dal fatto che x segue un algoritmo. Si tratta, come è evidente, di un *non sequitur*. Esistono programmi per il gioco degli scacchi che sono di livello pari ai migliori

giocatori umani, e tali programmi non usano alcun algoritmo che garantisca la vittoria agli scacchi (nonostante tale algoritmo, in principio, esista, dato che gli scacchi sono comunque un gioco finito). In modo simile, è possibile riconoscere come Penrose non riesca a chiudere le «scappatoie» che egli stesso prospetta per l'argomentazione gödeliana.

La prima di queste è che l'algoritmo di fatto usato dai matematici sia inconoscibile a essi (ad esempio perché orrendamente complesso): Dennett sottolinea come questa sia una possibilità familiare ai ricercatori in intelligenza artificiale; un algoritmo così complesso simulerebbe la competenza di un umano ma sarebbe «invisibile» al suo beneficiario (quando diciamo che abbiamo risolto un problema usando l'intuizione, vogliamo in realtà dire che non sappiamo come l'abbiamo risolto).

La seconda scappatoia è che l'algoritmo potrebbe essere non corretto: per Penrose ciò è impossibile dati gli standard di rigore della comunità matematica. Dennett sottolinea come «le istituzioni sociali in cui i matematici presentano le proprie dimostrazioni, si controllano a vicenda, fanno errori in pubblico, contando poi sul pubblico per correggere quegli errori» (p. 443) conferiscano alla comunità matematica poteri inferenziali superiori a quelli di qualsiasi individuo. Ma ciò ovviamente non significa che non vi siano algoritmi all'opera in tale processo; al contrario, tale processo mostra le capacità virtualmente illimitate del procedere algoritmico.

6 La «nuova» argomentazione

Come notato da alcuni commentatori (vedi Chalmers [1995] e McCullough [1995]) è possibile riconoscere, contenuta nel terzo capitolo di *Shadows of the Mind*, un'argomentazione «nuova» rispetto a quella sviluppata in *The Emperor's New Mind* e nel secondo capitolo di *Shadows*. Questa nuova argomentazione di Penrose appare come rafforzamento di quella originale: nonostante la conclusione rimanga la stessa, la premessa viene significativamente indebolita.

Ricordiamo come l'argomentazione diciamo «classica» di Penrose è una riduzione all'assurdo della tesi «io sono il sistema conoscibilmente corretto F ». Questa tesi si compone di due parti: (i) io sono il sistema F ; (ii) io so che il sistema F è corretto (almeno per enunciati Π_1). Infatti, il teorema di Gödel dice che se F è coerente allora c'è un enunciato vero non dimostrabile; d'altra parte, se io so che F è corretto allora posso concludere per *modus ponens* che c'è un enunciato vero indimostrabile in F (questo ovviamente richiede non solo che il teorema di Gödel sia vero, ma che io *sappia* che è vero, ad esempio perché ne so riprodurre la dimostrazione). Siccome io so che l'enunciato indimostrabile è vero, ma F non può «saperlo», io non sono F . In particolare, osserviamo come il teorema di Gödel faccia una parte tutto sommato secondaria in questa argomentazione, e che vi è almeno un'altra premessa aggiuntiva che ha un ruolo essenziale.

Ora, nella cosiddetta nuova argomentazione di Penrose la seconda premessa della riduzione, «io so che F è corretto», viene a sostituita da «io so di essere F ». Si può discutere ovviamente se effettivamente si tratti di un indebolimento della premessa (e quindi di un rafforzamento dell'argomentazione). Ma una volta indebolita la premessa, sta a Penrose far vedere come la premessa originaria sia in realtà già implicata da quella apparentemente più debole. A questo riguardo Penrose fa tutta una serie di considerazioni volte a far vedere come non possa esserci vera comprensione senza consapevolezza, e quindi se io sapessi di essere F non potrei non riconoscere le procedure dimostrative incorporate da F come essenzialmente corrette.

In vero, è possibile persino spingersi su questa strada fino ad abolire completamente la seconda premessa della riduzione, e, basandosi sul fatto che comprensione implica consapevolezza, far vedere come, se io di fatto fossi F , non potrei — *cartesianamente* — non sapere di essere F e quindi che le procedure di F sono essenzialmente corrette.

Si tratta, come si vede, di punti alquanto delicati, in cui la dimostrazione puramente matematica del teorema di Gödel è lasciata sullo sfondo, e tutta l'azione si concentra su premesse e argomentazioni filosoficamente pregnanti, e quindi assai più soggette a essere messe in discussione del fatto matematicamente indiscutibile datoci da Gödel.

McCullough ha fatto notare che vi è una inerente ambiguità nella nuova argomentazione di Penrose rispetto al sistema formale F . Ad esempio, non è chiaro se F rappresenta i poteri inferenziali di un dato matematico o anche le sue conoscenze empiricamente acquisite. La differenza fra queste alternative diventa importante quando consideriamo la convinzione del matematico che F sia in grado di arrivare alle sue stesse conclusioni. Infatti, tale convinzione potrebbe essere stata acquisita empiricamente, cosicché questa conoscenza empirica potrebbe non essere adeguatamente riflessa in F .

Penrose aggira tale problema considerando (nella sezione 3.16 di *Shadows*) un sistema formale F' , ottenuto aggiungendo a F il fatto che F rappresenta i poteri inferenziali del matematico (prima che tale conoscenza sia acquisita). L'argomentazione gödeliana può essere ripetuta per F' senza ulteriori modifiche. Questo è, essenzialmente, la «nuova argomentazione centrale» di *Shadows*.

Ne segue che il teorema di Gödel non implica affatto che il ragionamento umano è non computabile. Piuttosto, l'unica conclusione che può essere raggiunta è che o il ragionamento umano non è computabile, oppure non è corretto, oppure è impossibile per noi scoprire quali siano i nostri poteri inferenziali e giungere alla conclusione che essi sono corretti. Naturalmente, questo obbliga Penrose a eliminare gli ultimi due disgiunti usando argomentazioni filosofiche sostanziali per stabilire che i nostri poteri inferenziali sono corretti e che è impossibile scoprire quali siano i nostri poteri inferenziali senza anche giungere alla conclusione che sono corretti (cfr. *Shadows*, sezione 3.2).

Come già notato da McCullough, nonostante a Penrose piaccia pensare che il teorema di Gödel è sufficiente, da solo, a smantellare la teoria computazionale della mente, questa sua posizione è in realtà più un'affermazione di psicologia che di matematica.

7 Ancora sulla diagonalizzazione

Presentiamo in questa sezione due argomentazioni «diagonali», in parte dovute a McCullough e Chalmers. In entrambi i casi si tratta di far vedere che i problemi inerenti alla diagonalizzazione non sono specifici dei sistemi computabili, e che a certe condizioni possono applicarsi anche a altri sistemi. Specialmente la seconda, ha una diretta pertinenza per la tesi di Penrose che se io fossi un sistema formale F , allora saprei di essere corretto. Come si è visto, tale tesi sembra basarsi sull'idea che io sono certo delle mie credenze e procedure inferenziali, e quindi *a fortiori* sono convinto che «se io sono *incontrovertibilmente certo* di ϕ , allora ϕ » (qui «incontrovertibilmente certo che ϕ » significa che ho soddisfatto le più stringenti condizioni (anche ideali) che mirano a assicurare la verità di ϕ).

La prima osservazione di McCullough che merita di essere riportata è mirata a stabilire che il teorema di Gödel sembra applicarsi anche agli esseri umani, indipendentemente dal fatto che essi siano «computabili» o meno. Penrose sottolinea come i matematici giungano a conclusioni che ritengono «incontrovertibilmente vere». Ora, idealmente, tutti gli enunciati veri dovrebbero essere riconosciuti come «incontrovertibili»; mentre questo è certamente una condizione assai forte, un requisito più debole, e quindi più facile a soddisfarsi è che nessun enunciato falso venga riconosciuto come incontrovertibile. Questo fatto, però, non può a sua volta essere una verità incontrovertibile. Si consideri infatti il seguente enunciato G :

Questo enunciato non può essere una convinzione incontrovertibile di Roger Penrose.

Se tale enunciato fosse una delle convinzioni incontrovertibili di Roger Penrose, allora sarebbe falso. Ne seguirebbe che fra gli enunciati incontrovertibili di Roger Penrose ve ne è almeno uno falso. Contrapponendo, se gli enunciati incontrovertibili di Roger Penrose sono corretti, G è vero. Siccome tale conclusione può essere riconosciuta da Roger Penrose, ne segue che se Roger Penrose pensasse di essere corretto, dovrebbe accettare G come incontrovertibile, e quindi G sarebbe falso. La conclusione è che se Roger Penrose pensasse di essere corretto, egli sarebbe in verità incorretto.

È persino possibile, a certe condizioni, rafforzare la conclusione «se Roger Penrose pensasse di essere corretto, egli sarebbe in verità incorretto», sostituendo la semplice *coerenza* per la correttezza. Dato G , possiamo concludere che se Roger Penrose fosse coerente allora G non può essere creduto, e quindi che G è, in particolare, vero. Se Roger Penrose sapesse di essere coerente, potrebbe usare il *modus ponens* per concludere G , e quindi che G non può essere incontrovertibile.

D'altra parte, una condizione plausibile sulla credenza è il «principio di introspezione», che se io credo ϕ , allora credo di credere ϕ . In questo caso il principio ci dà che se Roger Penrose crede G (come stabilito alla fine del paragrafo precedente) allora sa di credere G . Quindi Roger Penrose sa di credere G e di non credere G , e quindi Roger Penrose è incoerente.

La conclusione, ottenuta anche da McCullough, è che è l'impossibilità di formalizzare i nostri ragionamenti in modo tale da essere certi che la teoria così ottenuta sia corretta non indica una limitazione inerente alle macchine (ma non agli umani). Al contrario, si tratta di una limitazione intrinseca delle nostre capacità di ragionare intorno ai nostri stessi processi inferenziali.

Il punto precedente può essere formulato in modo ancora più incisivo usando una versione di un'argomentazione originariamente dovuta a Löb, e variamente ripresa da alcuni commentatori di Penrose. La versione che ne diamo qui si concentra sulla mossa principale della nuova argomentazione e cioè che noi sappiamo di essere corretti, e quindi che se noi credessimo, anche solo ipoteticamente, di essere F , dovremmo concludere che F è corretto.

Si consideri l'insieme di credenze incontrovertibili di un dato soggetto, ad es., Roger Penrose. Dato un enunciato ϕ , abbreviamo con $\vdash \phi$ l'enunciato « ϕ è una credenza incontrovertibile di Roger Penrose». Naturalmente, Roger Penrose deve poter avere credenze sulle proprie credenze incontrovertibili, e quindi possiamo assumere che Roger Penrose abbia, fra i propri predicati, anche un predicato $\text{cr}(x)$ che esprime il fatto che x è una credenza incontrovertibile di Roger Penrose.

Assumiamo anche che il predicato cr soddisfi i seguenti principi per enunciati ϕ, ψ qualsiasi:

- (1) Se $\vdash \phi$ allora $\vdash \text{cr}(\ulcorner \phi \urcorner)$;
- (2) $\vdash \text{cr}(\ulcorner \psi \rightarrow \psi \urcorner) \rightarrow (\text{cr}(\ulcorner \phi \urcorner) \rightarrow \text{cr}(\ulcorner \psi \urcorner))$;
- (3) $\vdash \text{cr}(\ulcorner \phi \urcorner) \rightarrow \text{cr}(\ulcorner \text{cr}(\ulcorner \phi \urcorner) \urcorner)$.

Supponiamo ora che Roger Penrose creda di essere corretto, e quindi creda tutti gli enunciati della forma $\text{cr}(\ulcorner \phi \urcorner) \rightarrow \phi$, cioè:

$$\vdash \text{cr}(\ulcorner \phi \urcorner) \rightarrow \phi.$$

Fissiamo dunque un enunciato ϕ a piacere. Siccome Roger Penrose certamente sa abbastanza matematica da dimostrare il lemma di diagonalizzazione, c'è un enunciato θ che dice di se stesso che, se creduto incontrovertibilmente, implica ϕ :

$$\vdash \theta \leftrightarrow (\text{cr}(\ulcorner \theta \urcorner) \rightarrow \phi),$$

e quindi in particolare

$$\vdash \theta \rightarrow (\text{cr}(\ulcorner \theta \urcorner) \rightarrow \phi).$$

Applicando a quest'ultimo le condizioni (1) e (2) e *modus ponens*, si ottiene:

$$\vdash \text{cr}(\ulcorner \theta \urcorner) \rightarrow \text{cr}(\ulcorner \text{cr}(\ulcorner \theta \urcorner) \rightarrow \phi \urcorner);$$

usando la condizione (2) applicata al conseguente:

$$\vdash \text{cr}(\ulcorner \theta \urcorner) \rightarrow (\text{cr}(\ulcorner \text{cr}(\ulcorner \theta \urcorner) \urcorner) \rightarrow \text{cr}(\ulcorner \phi \urcorner)).$$

Ma la condizione (3) ci dice che

$$\vdash \text{cr}(\ulcorner \theta \urcorner) \rightarrow \text{cr}(\ulcorner \text{cr}(\ulcorner \theta \urcorner) \urcorner),$$

e quindi usando logica proposizionale

$$\vdash \text{cr}(\ulcorner \theta \urcorner) \rightarrow \text{cr}(\ulcorner \phi \urcorner).$$

Ma per ipotesi Roger Penrose pensa di essere corretto ($\vdash \text{cr}(\ulcorner \phi \urcorner) \rightarrow \phi$), e quindi per transitività

$$\vdash \text{cr}(\ulcorner \theta \urcorner) \rightarrow \phi.$$

Quest'ultimo è equivalente a θ , dato che θ è un punto fisso, e quindi

$$\vdash \theta,$$

da cui la condizione (1) dà

$$\vdash \text{cr}(\ulcorner \theta \urcorner),$$

e infine usando *modus ponens*

$$\vdash \phi.$$

Siccome ϕ era scelto a piacere, otteniamo che se Roger Penrose crede di essere corretto (e sa abbastanza matematica), allora crede qualsiasi enunciato ϕ , una conclusione non certo benvenuta per la posizione di Penrose.³ Si osservi tra l'altro, che Roger Penrose, se coerente, non sa di essere coerente, perché se sapesse, ad es., di non credere che $0 = 1$, cioè $\vdash \sim \text{cr}(\ulcorner 0 = 1 \urcorner)$, allora usando la logica proposizionale

$$\vdash \text{cr}(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$$

e per l'argomentazione appena data, $\vdash 0 = 1$.

Vale la pena sottolineare qui una replica fatta da Penrose a obiezioni di questo tipo, come ad esempio quelle di Chalmers e McCullough. Penrose sostiene che l'incoerenza esibita dalla dimostrazione può essere facilmente evitata, secondo Penrose, semplicemente restringendo il tipo di asserzioni che tali sistemi di credenze sono abilitati a produrre. In altre parole, Penrose propone (§3.8) di restringere tali sistemi di credenze alla produzione di enunciati Π_1 , osservando poi come l'enunciato incriminato («Questo enunciato non può essere creduto») non abbia tale forma. Naturalmente, se il sistema di credenze deve comunque incorporare le procedure valide di dimostrazione matematica, non vi è restrizione sui mezzi che esso può impiegare per giungere alle sue conclusioni. Ad esempio, tale sistema può utilizzare considerazioni relative a enunciati aritmetici arbitrariamente complessi, o a grandi cardinali in teoria degli insiemi, o infine alla propria coerenza o correttezza. La restrizione ha luogo solo nel momento in cui certi enunciati sono identificati come

³A questo punto il lettore è invitato a considerare la dimostrazione del teorema di Löb, ad es. in Boolos e Jeffrey [1989], pp. 187-89.

«incontrovertibili», e a tale stadio solo enunciati Π_1 sono ammessi. (Qui si vede anche l'importanza di limitarsi alla nozione di correttezza per i soli enunciati Π_1 oppure, equivalentemente, alla semplice coerenza.)

È anche importante osservare che Penrose riconosce, e anzi rivendica di avere osservato già in §7.9 di *Shadows*, come la diagonalizzazione gödeliana non si applichi solo a sistemi formali «computazionali» (cioè, sembra di capire, sistemi il cui insieme di conseguenze è ricorsivamente enumerabile), ma a teorie di arbitraria complessità, aritmetica, analitica, o del secondo ordine. Questo di per sé solleva interessanti questioni, ma soprattutto testimonia, per Penrose, della necessità di limitare i possibili *outputs* di ogni dato sistema di credenze.

Ritenendo comunque di aver «salvato» la seconda argomentazione, Penrose è tuttavia pronto a ribadire la cogenza dell'argomentazione gödeliana classica. Penrose elenca le manovre evasive messe in opera dal computazionalismo per sfuggire all'argomentazione classica:

forse agiamo e percepiamo secondo un algoritmo inconoscibile; forse la nostra comprensione matematica è intrinsecamente incorretta; forse potremmo venire a conoscere gli algoritmi con i quali comprendiamo la matematica, ma siamo incapaci di scoprire la loro funzione attuale. Va bene, tutte queste sono possibilità logiche, ma sono veramente spiegazioni plausibili? (§4.5)

La risposta naturalmente è che queste spiegazioni sono plausibili solo dal punto di vista del computazionalismo (§4.6). Comunque per Penrose, almeno un certo progresso è stato ottenuto, poiché nessuno dei suoi commentatori ha disputato la conclusione G, secondo cui «i matematici umani non usano un algoritmo conoscibilmente corretto allo scopo di scoprire verità matematiche».

8 Sistemi formali e calcolatori

Questo ci porta molto vicini al cuore della questione, e a un punto che è stato più o meno trattato dagli altri commentatori, e specialmente da D. McDermott [1995]. Si tratta del rapporto fra calcolatori e sistemi formali. Come dice McDermott, «i calcolatori digitali sono sistemi formali, ma i sistemi formali che essi sono sono quasi sempre distinti dai sistemi formali (o informali) con cui le loro computazioni sono connesse» (§4.1). In altre parole, data una macchina di Turing M , vi sono in principio non uno ma due sistemi formali associati con M : il sistema formale che rappresenta la macchina e il sistema formale rappresentato dalla macchina, e naturalmente in generale i due sistemi saranno non solo distinti ma anche molto diversi. Ad esempio, mentre il primo sistema ha per oggetto gli stati interni della macchina e le relative transizioni, il secondo può avere per oggetto i numeri naturali, o gli insiemi, o gli orari delle linee aeree. Solo il primo sistema deve essere coerente e corretto (ad esempio, non deve assegnare due diversi stati alla macchina allo stesso istante), ma il secondo può essere un sistema formale qualsiasi, anche un sistema il cui unico «teorema» è una contraddizione « p e non- p ».

Dato un sistema formale F , c'è sempre una macchina di Turing M_F che enumera i teoremi di F , e inversamente, data una macchina di Turing M , c'è sempre un sistema formale F_M che la descrive (e che permette di inferire, ad es., che M produce valore p su argomento q). Ma in generale *non* si ha

$$F = F_{M_F}.$$

Cioè, dato un sistema formale F , si ottiene una macchina che ne enumera i teoremi, che a sua volta può essere descritta da un sistema formale. Ma quest'ultimo non è in generale equivalente al sistema formale originario. Mentre F può essere un sistema che ci permette di inferire, ad es.,

teoremi geometrici, F_{M_F} è un sistema che dà teoremi concernenti gli stati interni di una macchina di Turing.⁴

9 Conclusioni

Come conclusione, ci limitiamo alle seguenti due osservazioni, aventi a che fare con la prima e la seconda argomentazione gödeliana, rispettivamente. Innanzi tutto, è opportuno notare come la prima argomentazione gödeliana possa essere interpretata così come propone Penrose solo a prezzo di effettuare una identificazione illegittima, quella fra F e F_{M_F} . Infatti, ogniqualvolta si è indicata la possibilità che il sistema formale F potesse non essere corretto o riconoscibile come tale, si è in realtà puntato il dito sul fatto che una macchina di Turing può avere associati non uno ma due sistemi formali: uno è il sistema formale che rappresenta la macchina, e il secondo è il sistema formale che la macchina deve rappresentare. Mentre il primo è necessariamente corretto, in quanto attribuisce alla macchina uno e un solo stato ad ogni istante, il secondo può utilizzare ogni sorta di euristiche, e non è necessariamente coerente.

La seconda osservazione ha a che fare con la nuova argomentazione gödeliana. Come abbiamo visto, Penrose propone una scappatoia alla dimostrazione di incoerenza sviluppata secondo le linee del teorema di Löb, secondo cui il «sistema di credenze» deve essere limitato nel suo *output* a enunciati di una particolare forma logica, mentre tale limitazione non è presente per quanto riguarda il funzionamento «interno» del sistema di credenze. Orbene, lasciando ad altro momento il compito di giudicare nel merito questa e altre proposte di Penrose, ci limitiamo qui a osservare che tale distinzione fra il «funzionamento interno» e il «comportamento esterno» di un sistema di credenze è perlomeno alquanto sospetta. Inoltre, non è affatto chiaro che tale drastica limitazione possa servire a evitare la contraddizione. Se le «conoscenze interne» del sistema danno adito a incoerenze, tali incoerenze si manifesteranno esternamente, ad esempio nella produzione, come «incontrovertibili» di enunciati contraddittori, per quanto entrambi della stessa ristretta forma logica.

Riferimenti Bibliografici

ANDERSON, A.R.

[1961] (a cura di), *Minds and Machines*, Prentice-Hall, Englewood, NJ 1961.

BENACERRAF, P.

[1967] «God, the Devil, and Gödel», *Monist*, vol. LI n. 1 (1967), pp. 9-32.

BOOLOS, G.

[1990] «On seeing the truth of the Gödel sentence», *Behavioral and Brain Sciences*, vol. 13 (1990), pp. 655-56.

BOOLOS, G. & JEFFREY, R.

[1989] *Computability and Logic*, terza edizione, Cambridge University Press.

CHALMERS, D.J.,

[1995] «Minds, Machines, and Mathematics», *Psyche*, vol. 2-1, 23 giugno 1995, <http://psyche.cs.monash.au/psyche>.

[1972] «On Alleged Refutations of Mechanism Using Gödel's Incompleteness Theorems», *Journal of Philosophy*, LXIX, n. 17 (1972), pp. 507-26.

⁴Sembra invece possibile argomentare che $M = M_{F_M}$, almeno in prima approssimazione.

DAVIS, M.

- [1990] «Is mathematical insight algorithmic?», *Behavioral and Brain Sciences*, vol. 13 (1990), pp. 659-60.
- [1993] «How subtle is Gödel's theorem? More on Roger Penrose», *Behavioral and Brain Sciences*, vol. 16 (1993), pp.611-12.

DENNETT, D.C.

- [1970] Recensione di J.R. Lucas, *The Freedom of the Will* (Oxford University Press, Oxford 1970), *Journal of Philosophy*, vol. LXIX, n. 17, pp. 527-31.
- [1978] *Brainstorms*, MIT Press - Bradford Books, Cambridge, Mass., 1978.
- [1995] *Darwin's Dangerous Idea*, Doubleday, New York 1995.

FEFERMAN, S.

- [1962] «Transfinite recursive progressions of axiomatic theories», *Journal of Symbolic Logic*, 27 (1962), pp. 364-84.
- [1995] «Penrose's Gödelian Argument», *Psyche*, vol. 2-1, 25 maggio 1995, <http://psyche.cs.monash.au/psyche>.

KLEENE, S.C.,

- [1952] *Introduction to Metamathematics*, Elsevier North Holland, Amsterdam and New York, 1952.

LUCAS, J.R.

- [1961] «Minds, Machines, and Gödel», *Philosophy*, vol. XXXVI (1961); ristampato in Alan R. Anderson (a cura di), *Minds and Machines*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [1968] «Satan Stultified: A Rejoinder to Paul Benacerraf», *Monist*, LII (1968).
- [1970] *The Freedom of the Will*, Oxford University Press, New York e Oxford, 1970.

MAUDLIN, T.

- [1995] «Between the Motion and the Act ...», *Psyche*, vol. 2-1, 2 maggio 1995, <http://psyche.cs.monash.au/psyche>.

MCCULLOUGH, D.

- [1995] «Can Humans Escape Gödel?», *Psyche*, vol. 2-1, 11 maggio 1995, <http://psyche.cs.monash.au/psyche>.

MCDERMOTT, D.

- [1995] «★ Penrose is Wrong», *Psyche*, vol. 2-1, 22 settembre 1995, <http://psyche.cs.monash.au/psyche>.

PENROSE, R.

- [1989] *The Emperor's New Mind*, Oxford University Press, Oxford 1989.
- [1990] «Précis of *The Emperor's New Mind*: Concerning computers, minds, and the laws of physics», *Behavioral and Brain Sciences*, vol. 13 (1990), pp. 643-55.
- [1994] *Shadows of the Mind*, Oxford University Press, Oxford 1994.
- [1996] «Beyond the Doubting of a Shadow», *Psyche*, vol. 2-1, 16 gennaio 1996, <http://psyche.cs.monash.au/psyche>.

PUTNAM, H.

- [1963] «Degree of Confirmation and Inductive Logic», in P.A. Schilpp (a cura di), *The Philosophy of Rudolf Carnap*, Open Court, 1963, pp. 761-783 (ristampato in H. Putnam, *Philosophical Papers*, vol. 1).

[1985] «Reflexive Reflections,» *Erkenntnis* 22 (1985), pp. 143–153.

TURING, A.

[1939] «Systems of logic based on ordinals», *Proc. London Math. Society*, 45 (1939), pp. 161-228.

WANG, H.

[1974] *From Mathematics to Philosophy*, Humanities Press, New York, 1974.