

# Gödel, Penrose, e i fondamenti dell'intelligenza artificiale

Gian Aldo Antonelli\*  
Dicembre 1996

Il dibattito sul ruolo e le implicazioni del teorema di Gödel per l'intelligenza artificiale ha recentemente ricevuto nuovo impeto grazie a due importanti volumi pubblicati da Roger Penrose, *The Emperor's New Mind* [1989] e *Shadows of the Mind* [1994]. Naturalmente, Penrose non è il primo né l'ultimo a usare il teorema di Gödel allo scopo di trarne conseguenze per i fondamenti dell'intelligenza artificiale. Tuttavia il recente dibattito suscitato dai due libri di Penrose è significativo sia per ampiezza sia per profondità. In queste pagine si vuole dare una rassegna di tale dibattito, cominciando dai suoi precursori negli anni '60 (fra cui Lucas, Putnam, e Chihara), per passare poi alle complesse argomentazioni proposte da Penrose e le reazioni di una serie di commentatori (ad esempio Dennett, Feferman, McDermott, Davis).

## 1. Preistoria

Se non il primo, certamente il più noto fra coloro che per primi hanno cercato di applicare il teorema di Gödel per trarne conseguenze in filosofia della mente, è J.R. Lucas. In un articolo originariamente pubblicato nel 1961 con il titolo "Minds, Machines, and Gödel", Lucas 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.

Lucas comincia con l'osservare che le cosiddette "macchine cibernetiche" costituiscono realizzazioni concrete di sistemi formali. Ne segue 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, questa constatazione 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).

Proprio come nel caso dei sistemi formali, è certamente possibile aggiungere l'enunciato indecidibile alla lista delle verità aritmetiche prodotte dalla macchina. Ma così facendo otterremmo un'altra macchina, distinta dalla prima, che a sua volta avrà un altro enunciato indecidibile, e così via. E anche se potessimo costruire una macchina in grado di esibire tutti gli enunciati indecidibili ottenuti in questo modo, tale macchina avrà pur sempre un suo proprio enunciato indecidibile.

La presa di posizione di Lucas suscitò naturalmente molte reazioni. Particolarmente rappresentativi sono però due articoli, uno di C. Chihara [1972] e uno di D. Dennett [1972], perché riassumono il dibattito originato dall'articolo di Lucas più di dieci anni prima. L'articolo di Chihara, "On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results" comincia col notare il ruolo cruciale che nell'argomentazione di Lucas è ricoperto dalla idea che una macchina, o un sistema formale, o una certa persona può "produrre-in-quanto-vero" (*produce as true*) un certo enunciato. Chiaramente questa nozione è del tutto cruciale, dato che l'intera

---

\* Ringrazio i Professori Diego Marconi e Gabriele Lolli per le loro utili osservazioni su una precedente versione di questo articolo.

argomentazione di Lucas è tesa a dimostrare che la mente umana non è identificabile con una macchina, poiché comunque si tenti di identificare la mente con una certa macchina  $M$ , è possibile esibire enunciati che  $M$  non può "produrre-in-quanto-veri", mentre la mente è perfettamente capace di riconoscerne la verità.

In secondo luogo, osserva Chihara, c'è un'obiezione originariamente mossa da Putnam [1961] contro Newman & Nagel [1958], ma che si applica anche all'argomentazione di Lucas. Essenzialmente, Putnam osserva, 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.*

Si può dimostrare che, se  $M$  è coerente, allora  $U$  non può essere né dimostrato né refutato da  $M$ . E poiché  $U$  afferma di non essere dimostrabile da  $M$ , ne segue in particolare che  $U$  è vero. Tuttavia è assolutamente possibile che  $M$  sia in grado di dimostrare l'enunciato (1). Per poter ridurre 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.

Naturalmente, Lucas è pronto ad argomentare 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,

Nonostante la possibilità che io sia essenzialmente incoerente non possa essere esclusa tramite argomentazioni formali e possa forse risultare inconcepibile, è pur sempre una posizione di completo nichilismo matematico, che, sebbene io non sia né un matematico né un teologo, sono comprensibilmente riluttante ad adottare (p. 158).

Naturalmente, Chihara osserva che, in mancanza di un resoconto più dettagliato di cosa significhi "produrre-come-vero", le affermazioni di Lucas rimangono del tutto inverificabili.

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]. Qui riportiamo l'argomentazione di Benacerraf nella forma semplificata datane da Chihara. Si consideri il seguente insieme di numeri naturali:

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

È allora possibile ridurre 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). Chihara conclude, con Benacerraf, che 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 richiama la nostra attenzione sul fatto che la dimostrabilità di un enunciato dipende dal bagaglio di conoscenze di chi deve effettuare la dimostrazione, e tali conoscenze possono essere determinate solo empiricamente. In altre parole, anche se per ogni particolare insieme di conoscenze o informazioni  $\alpha$ , l'insieme  $S$  risulta ricorsivamente enumerabile relativamente ad  $\alpha$ , non è detto che tale enumerazione sia uniforme in  $\alpha$  (tale cioè che esista un'unica procedura meccanica che, dato  $\alpha$ , 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.

Un punto importante, sui cui si dovrà tornare, è sollevato da Dennett [1972] recensendo Lucas [1970]. Dennett comincia con l'osservare che chiunque voglia "estrarre conseguenze anti-meccaniciste" dal teorema di Gödel, si trova a dover affrontare la seguente difficoltà: il teorema di Gödel non concerne, così come è formulato, né oggetti materiali né menti. Conseguentemente, Lucas e gli altri anti-meccanicisti si trovano a dover chiudere lo iato che separa le verità sui sistemi formali dalle verità sui movimenti degli oggetti fisici.

Un modo alquanto naturale per mettere in relazione oggetti fisici e sistemi formali (quali ad esempio le macchine di Turing) è quello per cui un certo oggetto viene *interpretato* come se "seguisse" le istruzioni che costituiscono il 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 stormir delle sue 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).

I matematici non sarebbero fondamentalmente diversi: ma allora ricade su Lucas l'onere di far vedere come la verità dell'enunciato di Gödel (relativo a questa interpretazione privilegiata) possa essere riconosciuta dalla mente umana.

## 2. L'argomento di Penrose

L'argomentazione principale proposta da Penrose [1989, 1994] è costituita da una generalizzazione del teorema di Gödel, nello stile del "teorema della fermata" di Turing, dovuta a Kleene. L'argomentazione in sé, in quanto dimostrazione matematica (corretta), è inattaccabile. Sono le conseguenze filosofiche che Penrose vuole trarne che hanno attratto l'attenzione dei critici. Vale dunque la pena di presentare l'argomentazione nella sua interezza, se non altro per cercare di individuare esattamente dove finisce la matematica e dove inizia la filosofia.

Seguiamo la versione data in *Shadows of the Mind*, ma le differenze fra tale versione e quella in *The Emperor's New Mind* sono minime. 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  $\Pi_1$

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 effettivamente  $C(n)$  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. In altre parole, l'arrestarsi della computazione  $A(q, n)$  costituisce una "dimostrazione" dell'enunciato " $C_q(n)$  non termina", e siccome  $A$  è 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  $C_1, C_2, C_3, \dots, C_n, \dots$  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

$G$  I matematici umani per determinare verità matematiche non usano una procedura che possa essere riconosciuta come corretta (*a knowably sound algorithm*).

Infatti, se i matematici usassero una tale procedura  $A$ , essi saprebbero anche 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.

Penrose passa a poi a prendere in considerazione una serie di possibili obiezioni che potrebbero essere (e in alcuni casi sono state) mosse alla conclusione  $G$ . Queste obiezioni sono numerate **Q1-Q20**, e qui ne presentiamo solo un paio delle più interessanti.

In **Q10-Q20** Penrose affronta una serie di obiezioni collegate fra loro, che hanno a che fare con la relatività della nozione di verità matematica. A questo riguardo Penrose sottolinea come qualsiasi matematico che accetti gli assiomi e le regole di inferenza di un dato sistema formale  $F$  accetta per ciò stesso anche l'enunciato di Gödel relativo a quel sistema (cioè l'enunciato  $G(F)$  che afferma di non essere dimostrabile in  $F$ ). Il passaggio dall'accettazione di  $F$  all'accettazione di  $G(F)$  non richiede alcun incremento dei principi e metodi di dimostrazione che si accettano, anche se a prima vista non ci se ne rende conto. Se si è convinti della correttezza di  $F$ , allora, in particolare, si è convinti della coerenza di  $F$ , e quindi dell'indimostrabilità di  $G(F)$ ; siccome  $G(F)$  asserisce precisamente di non essere dimostrabile,  $G(F)$  è vera. Quindi accettare  $F$  come un sistema coerente significa accettare la verità di  $G(F)$ . Ma siccome  $G(F)$  non è dimostrabile, ne segue che le deduzioni formali eseguite sulla base di  $F$  non possono esaurire la totalità delle credenze dei matematici.

In **Q19** Penrose prende in considerazione una procedura introdotta da Turing [1939] e in seguito studiata da Feferman [1962]. Si supponga di avere un certo sistema formale  $F$ , sufficientemente potente da avere un enunciato di Gödel indecidibile  $G(F)$ . Si può allora pensare di aggiungere  $G(F)$  come nuovo assioma ad  $F$ , ottenendo un nuovo sistema formale  $F_1$ , che avrà un ulteriore enunciato indecidibile  $G(F_1)$ , che, aggiunto agli altri assiomi, produce un ulteriore sistema formale  $F_2$ , e così via. Si ottiene in questo modo una progressione transfinita di teorie assiomatiche, la cui lunghezza può essere misurata da un ordinale ricorsivo (e conversamente, per ciascun ordinale ricorsivo si può definire una progressione di tale lunghezza). È possibile far vedere che ogni enunciato  $\Pi_1$  vero può essere dimostrato per mezzo di questo processo di gödelizzazione iterata. Tuttavia non vi è alcuna procedura che ci permetta di iterare la gödelizzazione per *tutti* gli ordinali ricorsivi simultaneamente (e quindi non vi sono progressioni la cui lunghezza è misurata dal minimo ordinale non ricorsivo). Ne segue che non è possibile in questo modo ottenere una procedura che ci permette di dimostrare tutti gli enunciati  $\Pi_1$  veri, e questa via per aggirare la tesi  $G$  rimane bloccata.

È interessante notare come tale iterazione del procedimento di “gödelizzazione” fosse già stata considerata da Hofstadter [1979] a proposito dell'argomentazione di Lucas. Hofstadter offre un interessante e caratteristico rovesciamento di una delle obiezioni degli anti-meccanicisti. Questi ultimi, proprio a partire da Lucas, indicano il fenomeno per cui sembra che non vi sia limite alla capacità umana di “gödelizzare” sistemi formali

a riprova del fatto che i poteri inferenziali umani non possono essere rappresentati da alcuna macchina. Formalmente, questo viene espresso citando un teorema, dovuto a Alonzo Church e Stephen Kleene, secondo cui non vi è alcun sistema ricorsivo di notazioni in grado di assegnare un nome a tutti gli ordinali costruttivi. Il rovesciamento compiuto da Hofstadter consiste nel considerare ciò indicazione del fatto che “qualsiasi essere umano semplicemente raggiungerà a un certo punto i limiti della propria abilità di ‘gödelizzare’. Da tale punto in avanti, i sistemi formali aventi tale complessità, per quanto certamente incompleti per le ragioni indicate da Gödel, avranno altrettanto potere quanto il dato essere umano” (p. 476).

### 3. La logica dell'argomentazione gödeliana

L'argomento gödeliano sviluppato da Penrose sia in *The Emperor's New Mind* sia in *Shadows of the Mind* ha naturalmente suscitato molte reazioni. Alcune di queste si sono concentrate sugli aspetti computazionali dell'argomento gödeliano.

Nel 1990 Penrose pubblica su *Behavioral and Brain Sciences* una versione sommaria del suo libro [1989] (cfr. Penrose [1990]). Come è costume per tale rivista, l'articolo di Penrose fu accompagnato da una serie di risposte e recensioni da parte di eminenti scienziati e filosofi della mente e da una replica da parte di Penrose. Tale dibattito fu continuato nel 1993, quando *Behavioral and Brain Sciences* raccolse altre recensioni che pubblicò con un'ulteriore replica da parte di Penrose. Molti altri interventi sono stati pubblicati più recentemente sulla rivista elettronica *Psyche*, pubblicata dalla Monash University in Australia e distribuita in tutto il mondo attraverso Internet (una versione cartacea sarebbe in preparazione presso la MIT Press). Per completezza, altri interventi e recensioni, come ad esempio Tieszen [1996] e Faris [1996] sono riportati in bibliografia.

Nel cominciare una rassegna di questo dibattito, prendiamo in considerazione prima di tutto gli aspetti strettamente computazionali dell'argomentazione di Penrose. A questo riguardo sono particolarmente interessanti gli interventi di logici importanti come George Boolos [1990], Martin Davis [1990, 1993] Hilary Putnam [1995] e Solomon Feferman [1995].

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  $\Pi_1$ , e l'enunciato "Se  $F$  è corretto, allora  $G(F)$  è vero" non ha tale forma. Tuttavia 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. In altre parole, l'argomentazione è la seguente: dato un sistema formale  $F$  si consideri la collezione  $T$  degli enunciati  $\Pi_1$  (nel linguaggio di  $F$ ) che sono *veri*. Tale collezione non è ricorsivamente enumerabile, mentre la collezione degli enunciati  $\Pi_1$  che sono *dimostrabili* in  $T$  è ricorsivamente enumerabile. Ne segue che se  $F$  è corretto, allora c'è un enunciato  $\Pi_1$  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*, così come non si 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  $\Pi_1$  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. Si noti però che Gödel è ben cosciente che il suo platonismo non è in contraddizione con il meccanicismo in filosofia della mente.

Un'analisi puntuale (e per molti versi impietosa) dell'esposizione tecnica di Penrose si può trovare in Feferman [1995]. Feferman si dice convinto della "estrema implausibilità" del modello computazionale della mente, ma afferma anche che l'argomento gödeliano di Penrose non riesce a stabilire tale punto di vista conclusivamente. Ciò è dovuto in gran parte all'alto numero di imprecisioni che si possono riscontrare 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*  $\Pi_1$ . Al tempo stesso, a pp. 90-92, Penrose nota che se un sistema formale è corretto, allora è certamente  $\omega$ -coerente. Feferman osserva che, mentre la semplice coerenza è equivalente alla correttezza per enunciati  $\Pi_1$ , la nozione di  $\omega$ -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  $\omega$ -coerente mediante l'abbreviazione  $\Omega(F)$ , e procede poi a enunciare il secondo teorema di incompletezza di Gödel nella forma: se  $F$  è  $\omega$ -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 la coerenza di  $F$ , *a fortiori*  $\Omega(F)$  non può essere un teorema di  $F$ . L'ipotesi della  $\omega$ -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  $F$ .

Feferman nota altri punti in cui l'esposizione di Penrose lascia a desiderare. Ad esempio, c'è una certa confusione fra il secondo teorema di Gödel e il teorema di Rosser (§ 3.5); a p. 92 Penrose sembra pensare che per poter esprimere la terminazione di una computazione sia necessario usare l'operatore del minimo  $\mu$  (§ 3.6); a p. 96 Penrose dichiara che sia  $G(F)$  sia  $\Omega(F)$  sono enunciati  $\Pi_1$ , mentre solo il primo lo è (§ 3.7).

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 of the Mind*. 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  $\omega$ -coerenza di  $F$ , in mancanza della quale è possibile esibire un controesempio.

Altre inesattezze sono riscontrate da Feferman in §§ 3.9-11, una delle quali relativa al trattamento dato da Penrose alla teoria delle progressioni ricorsive di teorie assiomatiche sviluppata da Turing e Feferman. 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  $\omega$ -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  $\Pi_1$  (come osservò Hilbert), 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  $\text{non-}p$ ) 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, così come esso viene effettivamente prodotto, non è meccanico" (§ 4.2).

Come nota Feferman, 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  $\Pi_1$  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).

Altrettanto impietosa dell'analisi di Feferman, anche se meno dettagliata, è la recensione di Putnam [1995], secondo cui l'opera di Penrose "rappresenta un triste episodio nella nostra presente vita intellettuale" (p. 370). Putnam prende le mosse dall'osservazione che Lucas e Penrose, per quanto simili nel loro utilizzo del teorema di Gödel, vogliono raggiungere conclusioni diverse. Infatti, mentre Lucas è interessato a dimostrare che "la mente umana è un'entità misteriosa che ha poco a che fare con la fisica o la chimica del cervello" (p. 370), Penrose ritiene invece che nel cervello si verifichino dei processi non computazionali che possono essere spiegati solo con una "nuova fisica".

Sia l'argomentazione di Lucas che quella di Penrose contengono degli errori cruciali, ma proprio come le conclusioni che le due argomentazioni si prefiggono di raggiungere sono diverse, così gli errori sono dissimili. Infatti, la conclusione di Lucas sembrerebbe dipendere, per Putnam, da una confusione fra due proposizioni diverse, a cui ci si può riferire come "l'enunciato che il sistema formale  $S$  è coerente": la prima è la proposizione, espressa in linguaggio ordinario, che i metodi impiegati dalla comunità matematica sono coerenti; la seconda è la proposizione che un dato sistema formale è coerente. Quest'ultima è una proposizione matematica assai complessa, ad esempio l'enunciato che una data funzione primitiva ricorsiva non assume mai valore zero.

L'errore di Penrose è meno tecnico. Supponiamo di accettare la conclusione di Penrose che "nessun programma la cui correttezza noi possiamo conoscere può simulare tutte le competenze matematiche umane" (p. 371). Ovviamente, ciò non stabilisce ancora che tali competenze non possano essere simulate adeguatamente: infatti esiste sempre la possibilità che tale programma non sia sufficientemente perspicuo e trasparente affinché la correttezza possa venire stabilita. Inoltre, tale stato di cose è compatibile con l'ipotesi che ciascuna delle regole a cui si affidano i matematici sia non solo corretta, ma anche conoscibilmente tale. Infatti, supponiamo che ci venga esibito un programma per calcolatore composto da centinaia di migliaia, se non milioni, di righe, ad esempio un programma che, stampato, occupa un libro delle dimensioni dell'elenco telefonico di New York. È assai improbabile che tale programma risulti sufficientemente perspicuo affinché ne si possa stabilire la correttezza, e tale possibilità è sufficiente a bloccare la conclusione di Penrose.

#### 4. La "nuova" argomentazione

Almeno due commentatori (Chalmers [1995] e McCullough [1995]) hanno creduto di riconoscere, contenuto 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 of the Mind*. Tale nuova argomentazione si presenta come un rafforzamento di quella originale, in quanto la premessa dell'argomentazione viene indebolita da "Io sono il sistema conoscibilmente corretto F" (l'ipotesi cioè che vi sia un algoritmo conoscibilmente corretto che incorpora le mie capacità inferenziali) a "Io sono F". Naturalmente, sta a Penrose ora far vedere che la prima premessa in è realtà già implicata dalla seconda. E infatti, secondo Penrose, sarebbe impossibile scoprire l'algoritmo sottostante ai nostri processi inferenziali, senza al tempo stesso riconoscere che esso incorpora procedure corrette di dimostrazione. È interessante allora ritornare ad alcune delle osservazioni sulla coscienza sviluppate da Penrose nella prima parte di *Shadows*; infatti, tali osservazioni diventano tanto più pertinenti adesso, nel momento in cui possono servire a gettare luce sulle implicazioni della "nuova argomentazione".

Tali osservazioni di Penrose sono sviluppate a cominciare dalla sezione 1.12 di *Shadows*, dove vengono distinte le varie nozioni di "consapevolezza" (*awareness*), "comprensione" (*understanding*), "coscienza" (*consciousness*) e "intelligenza" (*intelligence*). Pur senza fornire precise definizioni di tali nozioni, Penrose comincia con il notare che non vi può essere comprensione senza consapevolezza: ad esempio, non vi può essere comprensione di una deduzione matematica, senza che si abbia una qualche consapevolezza di ciò che la deduzione riguarda. Analogamente, non vi può essere vera intelligenza senza un qualche tipo di comprensione. La consapevolezza è un aspetto importante della coscienza (l'altro aspetto è il libero arbitrio, ma persino Penrose si ritrae di fronte a questa *vexata quaestio*). Come si vede, Penrose prepara la strada per il passo decisivo della sua "nuova argomentazione", quello cioè in cui fa vedere che non vi può essere vera comprensione senza consapevolezza. In particolare, io non potrei sapere di essere F senza essere consapevole delle particolari procedure che F incorpora, che verrebbero quindi a essere riconosciute come corrette. Infatti, io non posso non pensarmi come corretto, perché altrimenti non potrei giustificare le mie stesse conclusioni. Ne segue che qualsiasi sistema che incorpora le mie stesse procedure inferenziali deve a sua volta essere corretto.

Come si è detto, sono due i commentatori che hanno posto particolare enfasi sulla "nuova argomentazione" di Penrose, e cioè McCullough e Chalmers. McCullough comincia con l'argomentazione che ormai conosciamo bene. Si supponga che i ragionamenti di un matematico, ad esempio Penrose stesso, possano essere adeguatamente rappresentati in un dato sistema formale F. Si supponga ugualmente che Penrose sia cosciente del fatto che F adeguatamente cattura i propri poteri inferenziali. Ora, secondo Penrose, tale convinzione implica la convinzione che F è corretto. Poiché Penrose conosce il teorema di Gödel, può concludere che  $G(F)$  è vero ma non dimostrabile in F. Abbiamo dunque esibito un enunciato la cui verità può essere inferita da Penrose ma non dimostrata in F.

McCullough quindi osserva che vi è una inerente ambiguità 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, così che questa conoscenza empirica potrebbe non essere adeguatamente riflessa in F. Penrose aggira tale problema considerando (nella sezione 3.16 di *Shadows of the Mind*) 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. Questa è, essenzialmente, la "nuova argomentazione centrale" di *Shadows*.

Ne segue che il teorema di Gödel non dimostra che il ragionamento umano non è computabile. Dimostra solo che se il ragionamento umano è computabile allora o non è corretto, oppure è impossibile per noi scoprire quali siano i nostri poteri inferenziali e giungere alla conclusione che essi sono corretti. Per Penrose è impossibile scoprire quali siano i nostri poteri inferenziali senza anche giungere alla conclusione che sono corretti (cfr. *Shadows of the Mind*, sezione 3.2).

Per McCullough questa posizione di Penrose è più un'affermazione di psicologia che di matematica. Anche se Penrose considera alcune delle sue convinzioni (in matematica) come "incontrovertibilmente vere" (*unassailably true*), non ne segue che i ragionamenti di Penrose sono incomputabili, ma solo che egli non potrà mai convincersi della loro computabilità.

Prima di tornare alla "nuova argomentazione", è opportuno presentare alcuni punti interessanti sviluppati da McCullough. Prima di tutto, McCullough osserva che, a certe condizioni, il teorema di Gödel si applica anche a teorie non computabili. Tali condizioni sono (a) che sia possibile codificare le formule della teoria T come termini e rappresentare operazioni sintattiche come la sostituzione; (b) che il predicato "x è un teorema di T" sia definibile in T. Nessuna teoria T che soddisfa sia (a) che (b) può essere simultaneamente corretta e completa, come è facile dimostrare. Ottenuto un enunciato G dimostrabilmente equivalente in T a "G non è un teorema di T", si ragiona come al solito: se G fosse un teorema di T, allora T sarebbe incorretta. Ne segue che se T è corretta, allora G non può essere un teorema di T. Siccome G è, in particolare, vera, ne segue che se T è corretta allora è incompleta. (L'ipotesi di correttezza può essere sostituita - come dimostrato da Löb - dall'ipotesi di semplice coerenza, che è definibile direttamente in T). Come esempio di una teoria non computabile che soddisfa (a) McCullough cita l'esempio della teoria ottenuta dalla teoria degli insiemi di Zermelo-Fraenkel aggiungendo tutti gli enunciati aritmetici veri. Ciò è dovuto al fatto che per ogni enunciato A, A è vero se e solo se vero a qualche livello della gerarchia cumulativa degli insiemi, e tale nozione può essere rappresentata nella teoria degli insiemi.

La seconda 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 questa 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. Conversamente, 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 in verità non lo sarebbe. (Una versione matematicamente più rigorosa della stessa argomentazione si può trovare in McCullough.)

McCullough conclude che 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.

La "nuova argomentazione" di Penrose è ripresa e sviluppata da Chalmers [1995]. Come già McCullough, Chalmers comincia con una valutazione critica della prima argomentazione gödeliana. Supponendo che i nostri poteri inferenziali siano adeguatamente catturati da un sistema formale corretto F, ne segue, per il teorema di Gödel, che F non può dimostrare il proprio enunciato di Gödel G(F); tuttavia, F può dimostrare l'enunciato condizionale "se F è coerente allora G(F)", e quindi, in particolare, F non dimostra la propria coerenza. Fin qui l'argomentazione standard, in essenza non dissimile da quella proposta ad esempio da Lucas. Chalmers osserva che il sostenitore dell'intelligenza artificiale ha una replica immediata a questo modo di argomentare: non c'è ragione di credere che noi siamo in grado di riconoscere la verità dell'enunciato G(F).

Tuttavia, secondo Chalmers, Penrose è molto più cauto di Lucas a questo riguardo. Penrose sostiene, correttamente, che i nostri poteri inferenziali non possono essere catturati da un sistema che è "conoscibilmente corretto". Questo naturalmente lo obbliga a sviluppare una argomentazione ausiliaria tesa a dimostrare che se noi fossimo un sistema formale corretto F, la correttezza di F sarebbe conoscibile da parte nostra. La ragione è che, dal punto di vista di F, gli assiomi di F sono certamente veri, e che è immediato verificare che le regole preservano la verità. A questo riguardo Chalmers osserva che non è affatto detto che il

sistema formale  $F$  si presenti nella forma di assiomi-e-regole: in generale, ovviamente, ogni macchina di Turing è equivalente a un sistema formale, ma non è detto che il sistema formale sia immediatamente discernibile dato il programma della macchina di Turing, in modo tale per di più che la verità degli assiomi e la validità delle regole risulti immediatamente evidente (alla macchina di Turing stessa).

Più complessa, invece, la "seconda argomentazione", che si trova sparsa nel capitolo 3 di *Shadows*, ma riassunta e condensata nel dialogo immaginario (§3.23) fra AI (*Albert Imperator*) e il robot MJC (*Mathematically Justified Cybersystem*). L'argomentazione è tesa a dimostrare che miei poteri inferenziali non possono essere rappresentati da alcun sistema formale  $F$ . Si supponga dunque, per una *reductio ad absurdum*, che  $F$  sia invece un tale sistema ("io sono  $F$ "). Si può allora identificare un ulteriore sistema formale  $F'$ , ottenuto da  $F$  aggiungendo l'affermazione "io sono  $F$ ". Siccome l'aggiunta di un enunciato vero a un sistema corretto dà un sistema corretto, anche  $F'$  è corretto. Riproducendo l'argomentazione gödeliana per  $F'$ , io posso giungere a riconoscere la verità dell'enunciato  $G(F')$ , che è tuttavia indimostrabile in  $F'$ . Tuttavia, se io sono  $F$ , nel momento in cui divento cosciente di tale fatto, divento equivalente a  $F'$ . Come nella prima argomentazione, ciò è impossibile, dunque l'ipotesi che io (so che io) sono  $F$  deve essere falsa. Secondo Chalmers,

il potere di questa argomentazione è dovuto al fatto che essa non dipende dalla nostra capacità di riconoscere che un sistema  $F$  è corretto, o di scoprire che noi siamo  $F$ . Piuttosto, essa usa l'*assunzione* che noi siamo  $F$  per raggiungere le necessarie conclusioni. A prima vista sembrerebbe che tale assunzione può solo mostrare che il sistema più esteso  $F'$  può dimostrare  $G(F)$ , ma l'intuizione alla base dell'argomentazione è che si può creare una situazione in cui  $F'$  vede la verità del proprio enunciato di Gödel" (§3.4).

Secondo Chalmers, l'assunzione critica dell'argomentazione di Penrose non è che noi possiamo sapere di essere  $F$ , ma che possiamo sapere che siamo coerenti. Infatti, Chalmers presenta una rielaborazione dell'argomentazione, sviluppata insieme a McCullough, che segue le linee del teorema di Löb. In tale versione, una contraddizione è ottenuta solo sulla base dell'ipotesi che noi sappiamo di essere coerenti (indipendentemente dal fatto che le nostre conclusioni siano ottenute algebricamente o meno). Lasciando da parte i dettagli della costruzione (peraltro simile a quella del teorema di Löb), consideriamo un enunciato  $G$  che esprime il fatto che  $G$  non può essere creduto. Al solito, io posso raggiungere la conclusione che se io sono coerente, allora non posso credere  $G$ , e quindi che se io sono coerente,  $G$  è vero. Ora se io sapessi anche di essere coerente, ne seguirebbe che io potrei sapere che  $G$  è vero, e quindi saprei di credere  $G$ , e quindi che  $G$  è falso. Saprei quindi di credere una contraddizione, contro l'ipotesi.

Questa versione ricostruita da Chalmers (e McCullough) mostra come Penrose abbia indicato l'ipotesi sbagliata come responsabile della contraddizione. La contraddizione non deriva dal fatto che noi sappiamo di essere un sistema formale corretto, ma dal fatto che noi sappiamo di essere un sistema (algoritmico o meno) coerente.

McDermott [1995] è forse il commentatore di Penrose meno simpatetico. McDermott comincia con il notare che "Penrose fa dipendere tutto dalla sua analisi del teorema di Gödel, ma tale analisi è tutta sbagliata" (§2.1). Inoltre, la *pars construens* del libro, quella cioè in cui Penrose suggerisce una connessione fra il fenomeno della coscienza e certe interazioni quantistiche a livello neuronale, "deriva la propria forza dalla patetica debolezza dell'alternativa computazionale, così come è descritta nella prima parte del libro. Il minimo difetto nella prima parte svuoterebbe del tutto la seconda" (§ 2.2).

Più specificamente, McDermott distingue, nell'argomento di Penrose, un "teorema 1" e un "non-teorema 1". Il primo è il tradizionale argomento gödeliano, e mentre ovviamente corretto, è privo di conseguenze devastanti per le prospettive dell'intelligenza artificiale. Il secondo è essenzialmente la "nuova argomentazione" messa in luce da McCullough e Chalmers, ed è essenzialmente scorretto. Il teorema 1 non si applica all'intelligenza artificiale perché la costruzione di un sistema formale corretto non è mai stata fra gli scopi di quest'ultima. Piuttosto, l'intelligenza artificiale ha concentrato le sue risorse sulla costruzione di sistemi che, magari non necessariamente corretti, riproducono più o meno integralmente i processi cognitivi umani, fra cui quelli che portano alla dimostrazione di teoremi matematici.

Più interessante è la critica di McDermott al cosiddetto "non-teorema 1". Proprio per venire incontro all'obiezione sopracitata al teorema 1, Penrose vuole trovare il modo di isolare una classe ristretta di enunciati matematici la cui correttezza è garantita. Per far ciò, come abbiamo visto, Penrose introduce il concetto di "incontrovertibilità" (*unassailability*). Ora naturalmente Penrose non può identificare il concetto di incontrovertibilità con la dimostrabilità in un qualche sistema formale. Tale concetto deve essere, come nota McDermott, informale ma con una garanzia di accuratezza, e ciò non è ovviamente facile da ottenere.

Proprio per venire incontro a tali difficoltà Penrose introduce la sua seconda argomentazione, quella del non-teorema 1. Tale argomentazione è distinta da quella classica, in quanto ha diverse premesse e diversa conclusione. Piuttosto che ridurre all'assurdo l'ipotesi che noi siamo un sistema formale conoscibilmente corretto (cioè un sistema formale corretto e tale che noi siamo in grado di riconoscerne la correttezza), Penrose, come abbiamo visto, vuole ridurre all'assurdo l'ipotesi che noi possiamo identificarci con un qualsiasi sistema formale, corretto o meno. Un passo cruciale in tale argomentazione è che se io sapessi di "essere" un certo sistema formale F, allora non potrei non credere che F sia corretto. Naturalmente, è proprio questo punto che, secondo McDermott, viene meno: "credere che un certo sistema formale sottostia ai miei processi inferenziali *non* implica credere che tale sistema formale sia corretto" (§ 4.10).

Questo ci porta molto vicini al cuore della questione, e a un punto che è stato più o meno trattato dagli altri commentatori, ma non così esplicitamente come da McDermott. 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*".

Prima di passare alle repliche di Penrose, vale la pena sottolineare l'ampiezza del dibattito svoltosi su *Psyche* che ha visto anche interventi non esplicitamente entrati in questa rassegna, come ad esempio Moravec [1995] e Maudlin [1995]. Moravec propone una continuazione del dialogo fra AI e MJC (*Shadows*, pp. 179-90) secondo linee non esattamente anticipate da Penrose, mentre Maudlin sviluppa un'argomentazione tesa a dimostrare come il teorema di Gödel, essendo un teorema concernente entità matematiche astratte quali i sistemi formali non può avere conseguenze fisiche, e in particolare non può avere conseguenze riguardo alla possibilità fisica di un sistema cibernetico intelligente. Poiché tali considerazioni, per quanto interessanti, o riprendono punti già trattati o cadono al di fuori dei limiti della presente rassegna, rimandiamo il lettore interessato direttamente al testo della recensione.

L'ultima reazione a *Shadows of the Mind* 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 Lucas prima e Penrose 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 garantisce 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 non essere 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.

## 5. Alcune repliche di Penrose

In [1996], Penrose replica ad alcune delle critiche che gli sono state mosse dai commentatori della rivista *Psyche*. In tale replica, Penrose essenzialmente reitera l'argomentazione gödeliana classica, sostenendo che essa è comunque cogente, ma lamentando al tempo stesso che pochi commentatori (tranne forse McCullough e Chalmers) hanno riconosciuto il ruolo e l'importanza della "nuova argomentazione" del terzo capitolo.

Penrose inizia la sua replica prendendo in considerazione gli appunti tecnici mossigli da Feferman. Ammettendo alcune "inaccuratezze", Penrose rettifica specialmente il ruolo attribuito in *Shadows* alla nozione di  $\omega$ -coerenza. Contrariamente a quanto affermato in *Shadows* (almeno nelle prime due ristampe), l'asserzione che un dato sistema formale  $F$  è  $\omega$ -coerente non ha la forma di un enunciato  $\Pi_1$ , e in particolare non è equivalente all'enunciato indecidibile di Gödel (come osserva Feferman, l'enunciato indecidibile di Gödel è equivalente alla semplice coerenza di  $F$ ). Penrose osserva anche, tuttavia, che tale errore non ha alcuna conseguenza per l'argomento sviluppato in *Shadows*, purché si faccia attenzione a sostituire l'enunciato  $G(F)$  ogniqualvolta la  $\omega$ -coerenza di  $F$  viene menzionata.

La seconda "inaccuratezza" riconosciuta da Penrose è l'ambiguità fra le due nozioni di correttezza identificate da Feferman: una è la nozione generale di correttezza per un dato sistema formale, l'altra la nozione di correttezza per i soli enunciati  $\Pi_1$  (quest'ultima di nuovo equivalente alla semplice coerenza di  $F$ ). In effetti, come vedremo in seguito in connessione con le critiche di Chalmers, Penrose ritiene che la sua posizione possa essere addirittura rafforzata passando alla nozione di correttezza più "locale". Resta comunque il punto, sottolineato da Penrose, che Feferman non ha preso in considerazione alcuna delle argomentazioni del terzo capitolo, e in particolare non ha espresso alcun parere sulla portata e l'efficacia della "nuova argomentazione".

L'intera terza sezione della replica è dedicata da Penrose alla difesa di questa "nuova argomentazione". Come si ricorderà, Chalmers ricostruisce tale argomentazione identificandone le premesse e la conclusione. La premessa che si vuole ridurre all'assurdo è che la mente umana possa essere conoscibilmente identificata con un sistema formale  $F$ , indipendentemente dalla sua correttezza; non è necessario infatti che tale correttezza compaia esplicitamente tra le ipotesi dell'argomentazione poiché essa può essere inferita. Infatti, nel momento in cui io considero l'ipotesi che i miei poteri inferenziali possano essere rappresentati da  $F$ , devo anche ammettere che  $F$  è corretto. In questo modo, secondo l'argomentazione ricostruita da Chalmers sarebbe

possibile aggirare le obiezioni che negano che F possa essere "conoscibilmente corretto".

Penrose osserva prima di tutto come l'ipotesi della nuova argomentazione possa essere indebolita da "io so che sono F" a "io sono F", dando in questo modo un'argomentazione corrispondentemente più forte. Penrose quindi prende in considerazione l'obiezione sviluppata da Chalmers (e McCullough), secondo cui sarebbe contraddittorio supporre di essere corretti. Come si ricorderà, Chalmers fornisce i dettagli di una dimostrazione (avente come modello il teorema di Löb), secondo cui qualsiasi sistema di credenze avente certi proprietà elementari non può dimostrare la propria coerenza, se coerente. (Vi è qui, in Penrose, un'oscillazione fra la nozione di correttezza e quella di coerenza, ma tale oscillazione è ereditata dall'esposizione di Chalmers.) Se ciò fosse vero, bloccherebbe il passaggio cruciale della nuova argomentazione di Penrose, quello in cui la correttezza (o coerenza) di F viene inferita dal semplice fatto che "io sono F". Tale passaggio è giustificato da Penrose facendo appello al fatto che F "incorpora procedure valide di dimostrazione matematica", il cui compito è precisamente quello di "instillare credenza" (§ 3.6).

Ciò nonostante, Chalmers e McCullough vogliono mostrare l'incoerenza della nozione stessa che un sistema di credenza possa "credere in se stesso". Tale incoerenza 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  $\Pi_1$ , 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 "incontrovertibili", e a tale stadio solo enunciati  $\Pi_1$  sono ammessi. (Qui si vede anche l'importanza di limitarsi alla nozione di correttezza per i soli enunciati  $\Pi_1$  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  $\mathcal{G}$ , secondo cui "i matematici umani non usano un algoritmo conoscibilmente corretto allo scopo di scoprire verità matematiche".

Alcuni punti minori rimangono da prendere in considerazione. Nella sesta sezione, Penrose affronta l'obiezione secondo cui il fatto che i matematici umani commettono errori permette loro di sfuggire alla diagonalizzazione gödeliana. Penrose naturalmente non nega che i matematici facciano errori, ma osserva che la sua argomentazione concerneva la classe di enunciati che i matematici possono affermare come veri *in linea principio*, e che era questa "nozione ideale di comprensione matematica che giace al di là della computabilità" (§6.2). Tale nozione ideale fornisce uno standard che i matematici riconoscono, anche se non

riescono a raggiungerlo, e il fatto di riconoscerlo indica che in qualche modo essi "hanno accesso a concetti ideali non computazionali" (§6.4).

La settima sezione invece affronta l'altra scappatoia a disposizione del computazionalismo, e cioè l'inconoscibilità dell'algoritmo. Ma se l'algoritmo è inconoscibile deve essere stato il prodotto di un qualche processo cieco quale la selezione naturale (dal momento che non può certamente essere stato progettato), ma ciò è impossibile per Penrose a meno che tale meccanismo non abbia a propria volta accesso a un elemento non computazionale (come è noto, Penrose sostiene che tale elemento verrà fornito da una nuova teoria fisica che unifichi con successo relatività e meccanica quantistica, ma tale aspetto del programma di Penrose è oltre i limiti di questa rassegna).

## 6. Conclusione

Molti aspetti del dibattito suscitato sia da *The Emperor's New Mind* che da *Shadows of the Mind* sono stati omessi. Soprattutto, non abbiamo fatto menzione della già citata discussione sulla "nuova fisica" di Penrose (che invero occupa ad esempio tutta la seconda parte di *Shadows*). Si tratta oltre che di un argomento interessante, anche di una pregevole introduzione a concetti difficili come quelli della meccanica quantistica, ma che esula dai limiti di una rassegna sulla applicabilità del teorema di Gödel ai fondamenti dell'intelligenza artificiale.

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, che molti commentatori hanno in qualche modo isolato, ma che è stata riconosciuta nella sua forma più esplicita da McDermott. Infatti, ogniqualvolta i vari commentatori hanno indicato la possibilità che il sistema formale  $F$  potesse non essere corretto o riconoscibile come tale, essi hanno 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 di Chalmers, 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 esibita da Chalmers. Se le "conoscenze interne" del sistema danno adito a incoerenze, tali incoerenze si manifesteranno esternamente, ad esempio nella produzione, come "incontrovertibili" di coppie 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 Cliffs, NJ, 1964.

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.
- CHALMERS, D.J.  
 [1995] "Minds, Machines, and Mathematics", *Psyche*, vol. 2-1, 23 giugno 1995, <http://psyche.cs.monash.au/psyche>.
- CHIHARA, C.S.  
 [1972] "On Alleged Refutations of Mechanism Using Gödel's Incompleteness Theorems", *Journal of Philosophy*, LXIX, n. 17 (1972), pp. 507-26.
- 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.  
 [1972] 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.
- FARIS, W.  
 [1996] "Shadows of the Mind: A Search for the Missing Science of Consciousness", *Notices of the American Mathematical Society*, vol. 43, n. 2 (1996), pp. 203-208.
- 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>.
- HOFSTADTER, D.  
 [1979] *Gödel, Escher, Bach: an Eternal Golden Braid*, Basic Books, New York, 1979.
- LUCAS, J.R.  
 [1961] "Minds, Machines, and Gödel", *Philosophy*, vol. XXXVI (1961); ristampato in Alan R. Anderson [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.
- 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>.
- MAUDLIN, T.  
 [1995] "Between the Motion and the Act ...", *Psyche*, vol. 2-1, 2 maggio 1995, <http://psyche.cs.monash.au/psyche>.

MORAVEC, H.

- [1995] "Roger Penrose's Gravitonic Brains", *Psyche*, vol. 2-1, 12 maggio 1995,  
<http://psyche.cs.monash.au/psyche>.

NEWMAN, J.R. e NAGEL, E.

- [1958] *Gödel's Proof*, New York University Press, 1958.

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.

- [1961] "Minds and Machines", in S. Hook (a cura di), *Dimensions of Mind*, Collier, New York, 1961.  
[1995] Recensione di R. Penrose, *Shadows of the Mind*., Oxford University Press, 1994, *Bulletin of the American Mathematical Society*, nuova serie, vol. 32, n. 3 (1995), pp. 370-373.

TIESZEN, R.

- [1996] Recensione di R. Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press, 1994, *Philosophia Mathematica*, terza serie, vol. 4 (1996), pp. 281-90.

TURING, A.

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

WANG, H.

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