Università
Ca'Foscari
Venezia

Corso di Laurea
Magistrale

in Scienze Filosofiche
ordinamento (ex D.M. 270/2004)

Tesi di Laurea

Oltre gli argomenti gödeliani

Elenchos e l'infinito della mente umana

Relatore
Ch. Prof. Paolo Pagani

Correlatore
Ch. Prof. Enrico Jabara

Laureando
Stefano Sandini
Matricola 859487

Anno Accademico
2020 / 2021

Indice analitico

Introduzione …….4

Parte I

IL PROBLEMA MENTE-CORPO E IL FENOMENO DELLA COSCIENZA

Capitolo 1

Il materialismo e il principio di chiusura causale del mondo fisico ……...7

Capitolo 2

L'"interazionismo" e le critiche di Popper al materialismo ……10

Capitolo 3

Il "naturalismo biologico" e la critica di Searle alla tesi dell'intelligenza artificiale ……...15

Parte II

I TEOREMI DI INCOMPLETEZZA DI GÖDEL: UNA SVOLTA NEL DIBATTITO INTORNO ALLA NATURA DELLA MENTE UMANA

Capitolo 1

La "crisi dei fondamenti" e il programma formalista di Hilbert ……..19

Capitolo 2

I teoremi di incompletezza di Gödel ……...31

Capitolo 3

Le conseguenze dei teoremi di incompletezza per il programma formalista ……...42

Capitolo 4

La macchina di Turing e la formalizzazione della nozione di algoritmo …….45

Capitolo 5

La *Gibbs Lecture*: Gödel a Providence ……...48

Capitolo 6
Gli argomenti gödeliani contro la tesi meccanicistica dell'identità tra mente umana e macchina di Turing ……53

6.1 L'argomento di Lucas ……53
6.2 Alcune obiezioni alla tesi di Lucas ……57
6.3 Il carattere dialettico dell'argomento di Lucas ……...63
6.4 L'argomento nuovo di Penrose ……..69
6.5 La critica di Shapiro e Lindström al nuovo argomento di Penrose ……..79

Parte III
PROSPETTIVE ULTERIORI

Capitolo 7
Dalla logica formale all'ontologia: logica della dimostrabilità e logica dell'evidenza ……89

Capitolo 8
Élenchos e l'infinito dell'anima ……..97

8.1 *Élenchos* e il realismo ontologico ……..97
8.2 L'impossibilità di una autofondazione della logica per via elenctica ……...101
8.3 L'infinito trascendentale e l'anima ……..106

Conclusioni ……..112

Bibliografia ……..115

Introduzione

Con il presente elaborato intendiamo riprendere un lungo ed acceso dibattito riguardante la natura della coscienza umana, che ebbe inizio intorno alla metà del XX secolo, e che rappresenta un'importante specificazione di quello assai più ampio ed annoso riguardante il problema del rapporto tra le mente ed il corpo. Uno dei punti fondamentali della questione che qui prenderemo in esame, è se la coscienza dell'uomo possa, nella sostanza, venire identificata o meno con quei processi chimici neuronali che, sulla base degli innumerevoli progressi e delle scoperte avvenute in campo neurofisiologico specialmente a partire dalla seconda metà del secolo scorso, sembrano stare in stretta correlazione con l'emergenza degli stati psichici interni, e che dunque paiono dover giocare quantomeno un ruolo chiave nel tentativo da parte della scienza e della filosofia di fornire una spiegazione al fenomeno della coscienza. In secondo luogo, grazie alla nascita delle teorie computazionali come conseguenza degli sviluppi del calcolo formale avvenuti nei primi decenni del Novecento, il problema mente-corpo vede allargarsi il proprio dominio fino ad includere la possibilità di teorizzare un rapporto di identità, o perlomeno di equiparabilità, tra mente umana ed intelligenza artificiale. Infatti, se da un lato la teoria fisicalista, o più in generale meccanicista, prevede che ogni evento o fenomeno della realtà sia riconducibile, o meglio, riducibile ad un'interazione di tipo causale tra gli elementi ultimi che costituiscono il mondo fisico, dall'altro le teorie della computazione prevedono la possibilità, in linea di principio, di costruire una simulazione, attraverso modelli logico-matematici, di ogni sistema fisico complesso. Va ricordato che da queste due posizioni prende le mosse, a partire dalla seconda metà del XX secolo circa, il progetto di ricerca avente di mira la realizzazione di forme di intelligenza artificiale, che qualcuno ha distinto in "Intelligenza Artificiale forte" (IA forte) e "Intelligenza Artificiale debole" (IA debole). Il primo, in linea con la visione meccanicistica, intende la coscienza come qualcosa di interamente riconducibile all'insieme dei processi neurochimici che hanno sede nel cervello, e intende tali processi come schemi (*pattern*) ricorrenti di informazione. Per questo motivo i teorici dell'IA forte ritengono idealmente possibile realizzare una replica esatta, o un modello, della struttura e del comportamento di quei processi nei termini di una funzione o algoritmo che traduca certi segnali in entrata (*input*) provenienti dal corpo (umano quanto artificiale) in determinati segnali in uscita (*output*) verso il corpo e quindi, in ultima istanza, verso l'ambiente con cui quello interagisce. Ad essere prospettata, qui, è l'idea che la mente dell'uomo, e nella sua essenza e nel suo modo di funzionare, sia interamente spiegabile in dimensione fisica e risulti perciò replicabile attraverso procedure algoritmiche implementabili in una macchina di Turing. Il secondo, invece, se da un lato concepisce il fenomeno della coscienza come simulabile, anche se soltanto dal punto di vista funzionale e comportamentale, dall'altro, al contrario del primo,

ne rifiuta la totale riducibilità a processo meccanico o elettrochimico. In altri termini, ciò che il programma dell'IA debole non accoglie è la proporzione propugnata dai difensori dell'IA forte, tale per cui come il corpo sta allo *hardware* di una macchina, la mente sta al suo *software*.

Nella Parte prima, a scopo introduttivo, esporrò i principali aspetti concernenti la tematica del problema mente-corpo, in cui prenderò in esame testi di Popper e Searle. Nella Parte seconda, di natura più tecnica, procederò a ricapitolare i punti salienti del dibattito sull'equiparabilità tra la mente ed una macchina di Turing. A questo proposito, ritengo che il punto d'avvio della discussione debba essere fatto risalire al 1931, ovvero all'anno di pubblicazione del testo *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I*, di Kurt Gödel. Questo testo, infatti, ha sancito il ridimensionamento del programma hilbertiano degli anni Venti, contribuendo notevolmente così alla presa di consapevolezza circa l'impossibilità per la matematica e la logica di autofondarsi come sapere coerente e completo. Successivamente, in occasione della celebre Gibbs Lecture del 1951, Gödel torna sugli esiti dei suoi teoremi di limitazione e sul fenomeno dell'incompletezza della matematica, e sviluppa alcune originali riflessioni riguardanti le capacità dimostrative della mente umana da una parte e quella di una macchina di Turing dall'altra. I risultati dei teoremi di incompletezza sono stati poi utilizzati a più riprese, da autori come il filosofo Lucas e il fisico Penrose, nel tentativo di dimostrare la profonda differenza sussistente tra il funzionamento della mente e quello della macchina di Turing. Tali argomentazioni (detti anche "argomenti gödeliani") sono state a loro volta oggetto di severe critiche da parte di molti pensatori appartenenti a diverse correnti di pensiero. Pensiamo, tra gli altri, a Putnam, Benacerraf, Lindström e Shapiro. Come cercheremo di mostrare, lo scontro tra la posizione meccanicistica da un lato e quella anti-meccanicistica dall'altro, termina in una sorta di stallo teorico, da cui la soluzione non può emergere se non in senso "negativo" dall'interno dei confini della logica, ovvero come rimando ad una dimensione ulteriore, quella dell'ontologia, che trascende e fonda la prima.

Allora, ciò che i teoremi di Gödel hanno da insegnarci, non in quanto "dimostrazioni", bensì da un punto di vista filosofico, è che i linguaggi formali si trovano costretti ad abbandonare, almeno in alcuni casi, quella pretesa di autofondazione sostenuta da una linea di pensiero il cui sviluppo, tra la seconda metà del XIX secolo e la prima metà del XX secolo, va dal logicismo di Frege e Russell, attraverso il programma formalista di Hilbert, fino al neopositivismo, rappresentato da pensatori come Carnap ed Otto Neurath.

Nella Parte terza, in particolare attraverso l'analisi in dettaglio della celebre figura argomentativa dell'*élenchos*, per la quale faremo affidamento ai testi di autori come Galvan e Pagani, argomenteremo a favore della necessità di postulare una dimensione – quella dell'"infinito trascendentale" – ulteriore rispetto a quella dell'"infinito potenziale" in cui è costretta

l'argomentazione logico-matematica, e che sottende l'intera struttura del pensiero; un orizzonte ultimo di senso che la nostra mente è in grado di scorgere, mai come qualcosa di determinato, ma sempre e soltanto come condizione trascendentale di ogni discorso. Da questo punto di vista, riteniamo che la ricerca, sia quella presente che quella futura, che intenda inserirsi nel dibattito intorno ai limiti della logica formale e al rapporto di identità-differenza tra mente umana ed intelligenza artificiale, possa trovare nel discorso ontologico un fertile terreno di crescita e di sviluppo.

Da ultimo, riteniamo qui opportuno anticipare un punto importante su cui cercheremo di porre l'attenzione nella parte finale della trattazione. A nostro avviso, è fondamentale, oggigiorno, contribuire alla demistificazione di un particolare tipo di retorica catastrofista e fantascientifica relativa agli sviluppi nel campo dell'intelligenza artificiale, di cui sono imbevuti gran parte degli attuali canali pubblicistici non-accademici, e in particolare diverse opere di stampo divulgativo e commerciale. In altri termini, risulta cruciale ridimensionare la portata di cui si è caricata la promessa tecnologica, comunque epocale, dell'avvento di un'autentica forma di intelligenza artificiale capace di operare attivamente nella realtà e di collaborare alla pari con l'essere umano. Seguendo questa linea, con il presente lavoro intendiamo: in primo luogo, mettere in luce i limiti che caratterizzano la tesi meccanicistica, la quale ipotizza che la mente sia, in un certo senso, un algoritmo che opera secondo regole e schemi di tipo finitario; in secondo luogo, mostrare l'incompatibilità tra la logica della dimostrabilità (quella che determina i processi computazionali) e la logica dell'evidenza (che, invece, regola il procedere razionale di tipo intuitivo); infine, sottolineare l'importanza di recuperare alcuni elementi-cardine della riflessione filosofica intorno all'anima e alla sua natura infinita, di contro all'appiattimento categoriale promosso dal programma riduzionista sostenuto dalla dominante visione fisicalistica della realtà.

Parte I

IL PROBLEMA MENTE-CORPO E IL FENOMENO DELLA COSCIENZA

Capitolo 1
Il materialismo e il principio di chiusura causale del mondo fisico

Il materialismo classico, che risale remotamente all'atomismo di Leucippo e Democrito, è quella particolare prospettiva filosofica che pone a fondamento ultimo della realtà un qualche elemento di tipo materiale, a sua volta non ulteriormente scomponibile o riducibile ad altro da sé come sua causa. Con ciò il materialismo pone, implicitamente, anche l'identità essenziale tra uomo e macchina. Infatti, se, come i suoi sostenitori affermano, ogni fenomeno è spiegabile facendo unicamente appello ad una serie di interazioni fisiche di tipo causale tra elementi materiali, allora anche il fenomeno della coscienza, che tradizionalmente ha funto da discrimine tra l'essere umano e le altre specie viventi (nonché, e a maggior ragione, tra l'uomo e gli oggetti inanimati), non potrà che essere riconducibile ad una ragion d'essere appartenente per intero al dominio della scienza naturale. Ciononostante, come rileva Popper nel primo volume de *L'Io e il Suo Cervello*, nel corso dei secoli la fisica è giunta, in misura sempre maggiore e attraverso numerose e rivoluzionarie scoperte (come, ad esempio, quella dell'attrazione gravitazionale o delle particelle subatomiche), a spiegare la materia come una forma di energia e quindi non più nei termini di una sostanza od essenza ultima, bensì come realtà proteiforme e processuale. Attraverso il progresso scientifico, dunque, «[i]l materialismo ha in tal modo trasceso se stesso».[1] Questo, tuttavia, significa non che il materialismo sia stato superato, bensì soltanto che esso ha subito una modificazione parziale delle sue premesse (dalla materia come sostanza alla materia come energia in costante trasformazione), senza che con ciò ne sia stato rigettato il nucleo teorico centrale, ovvero la riducibilità di ogni fenomeno od evento ad una interazione di tipo fisico, e quindi l'identificabilità, in ultima analisi, tra uomo e macchina (questa volta intesa non in senso meramente meccanico, bensì elettrochimico). È in questo senso che, in particolar modo a partire

[1] Cfr. K. R. Popper – J. C. Eccles, *L'Io e il Suo Cervello*, a cura di B. Continenza, trad. it. di G. Mininni e B. Continenza, Armando Editore, Roma 1981, vol. I, p. 19; tit. orig. *The Self and Its Brain, An Argument for Interactionism*, Springer-Verlag, 1977.

dal Novecento in poi, si può parlare di una nuova connotazione del materialismo, ovvero del "fisicalismo" (termine che risale all'epoca del Circolo di Vienna e al neopositivismo, e che d'ora in avanti utilizzerò a indicare il materialismo in generale nella sua veste contemporanea). Va sottolineato, inoltre, che all'interno della dottrina materialista si sono venute differenziando quattro posizioni principali che prenderemo brevemente in considerazione seguendo la chiara esposizione che ne dà Popper: il fisicalismo radicale, il panpsichismo, l'epifenomenismo, la teoria dell'identità.

1. Il fisicalismo radicale, o funzionalismo radicale, propone una soluzione di tipo monista al problema mente-corpo negando l'esistenza degli stati psichici (o quantomeno ritenendo non rilevante interrogarsi su di essi) e restringendo il dominio ontologico al solo mondo fisico, la cui descrizione ci è fornita dalle scienze naturali.
2. Il panpsichismo, o parallelismo (da distinguersi dallo spiritualismo o dall'idealismo), ritiene che esistano due lati complementari e non interagenti della materia: l'uno esterno e l'altro interno. Quest'ultimo sarebbe in un certo qual modo simile allo stato cosciente di un'anima. In altri termini, ogni ente sarebbe dotato di anima, in grado minore o maggiore a seconda della complessità del suo corpo o del suo organismo. Analogamente al fisicalismo radicale, anche il panpsichismo, in un certo senso, adotta una concezione monista della realtà.
3. Per l'epifenomenismo, a differenza del panpsichismo, gli stati mentali non esistono se non come fenomeni che emergono dal piano delle interazioni causali, senza però farne parte, e soltanto in relazione ad organismi complessi quali l'uomo o gli animali superiori. In questo senso, gli stati di coscienza non interagiscono con gli eventi del mondo fisico.
4. La teoria dell'identità sostiene, da un lato e in maniera analoga all'epifenomenismo, l'emergere degli stati di coscienza in relazione esclusivamente ad organismi biologici complessi, e dall'altro, ma al contrario di quello, il loro interagire causalmente con i processi fisici. Ciononostante, tale interazione non avviene tra due dimensioni totalmente distinte, pena la ricaduta nell'aporeticità del dualismo cartesiano, bensì mediatamente, ovvero per tramite dei processi che occorrono nel cervello a cui gli stati coscienti corrispondono. Da questo punto di vista, perciò, l'identità tra l'insieme degli stati mentali e quello dei processi cerebrali va intesa soltanto in un senso lato, come una forma di parallelismo. Come fa notare Popper, intesa in questo senso, la teoria dell'identità va incontro alle medesime difficoltà a cui è soggetto il parallelismo.

Un punto fondamentale del programma fisicalista è quello per cui il mondo fisico deve essere inteso come causalmente chiuso, ovvero tale per cui o non esiste alcun piano ulteriore ad esso, o, anche se tale piano esistesse, quest'ultimo non potrebbe in alcun modo entrare in relazione con il primo. Tuttavia, l'indipendenza del mondo fisico, argomenta Popper, oltre ad essere incompatibile con la nostra visione *naïve* della realtà, secondo la quale vige un rapporto, sebbene problematico, tra la mente e il corpo, non sarebbe nemmeno sufficiente a rendere conto del perché si manifesti qualcosa di così difficilmente negabile o trascurabile come la coscienza.

Più in generale, un aspetto molto importante della visione materialista è il suo allinearsi con il programma riduzionista, per il quale «è necessario che tutte le possibilità che si sono realizzate nel corso del tempo e dell'evoluzione siano state potenzialmente preformate o pre-stabilite fin dall'inizio».[2] In altri termini, il riduzionismo elimina dal dominio della storia la possibilità che qualcosa di nuovo effettivamente accada. Da questo punto di vista è evidente che il concetto di evoluzione si riduce in tal modo al semplice svolgersi nel tempo di quelle potenzialità già insite da sempre nella materia da cui l'universo ha avuto inizio. È utile notare, anche se solo di passaggio, che una simile posizione, per cui tutti gli eventi della storia sono da sempre pre-determinati, risale già al pensiero greco antico.

In particolare, si pensi all'atomismo classico, secondo cui ogni mutamento, sia quantitativo che qualitativo, è dovuto ad un mero cambio disposizionale da parte di quelle particelle indivisibili che costituiscono il fondamento ultimo della realtà fisica. All'interno di questo quadro, naturalmente, trova posto la concezione laplaciana, per la quale un demone dalle capacità sovrannaturali che conoscesse le esatte leggi che regolano il divenire insieme all'intero stato di cose in un dato momento (ovvero, in termini moderni, la posizione e il momento di ogni singola particella), sarebbe in grado di conoscere, in una sola volta, tutto il passato, il presente e il futuro del cosmo.

Quanto ora detto sintetizza a grandi linee ciò in cui consiste la visione deterministica del mondo abbracciata dal materialismo classico e, in particolare, dal fisicalismo. Si tratta di una concezione assai diffusa e, per certi versi, molto intuitiva ed elegante nella sua semplicità. La perenne ricerca da parte della scienza di leggi e costanti universali, infatti, non fa che rispecchiare l'aspirazione del programma riduzionista di spiegare, attraverso cause elementari appartenenti ai livelli inferiori della realtà (come ad esempio quello fisico o chimico), eventi che si manifestano ai livelli superiori (ad esempio quello biologico o psicologico): eventi tra cui rientrano, naturalmente, gli stati psichici o mentali che costituiscono nel loro insieme il fenomeno della coscienza umana.

[2] Cfr. *ivi*, p. 27.

Capitolo 2

L'"interazionismo" e le critiche di Popper al materialismo

Nell'opera sopra citata Popper svolge un'estesa e puntuale critica alla dottrina materialista nelle diverse versioni sopra elencate, avanzando, di contro, alcuni argomenti a favore della teoria dell'"interazionismo". Come cercherò di mostrare soffermandomi brevemente soltanto su alcuni dei passaggi di maggior interesse, a mio avviso le sue critiche non sempre colgono nel segno, in quanto non tengono adeguatamente conto del fatto che la linea argomentativa del fisicalismo e del riduzionismo indica un programma di principio, e non sembra efficacemente confutabile riferendosi allo stato attuale delle nostre conoscenze. Come Popper stesso riconosce, infatti, ciò che viene messo in luce è che non è possibile fornire una vera e propria confutazione del materialismo, se non assumendo come presupposto una determinata prospettiva ontologica.

Innanzitutto, l'interazionismo, almeno relativamente al fenomeno della coscienza, va distinto dall'epifenomenismo, poiché, a differenza di quest'ultimo, esso assume come proprio punto centrale l'interazione costante e reciproca tra quello che Popper chiama il «Mondo 1» (il mondo fisico) e quello che egli chiama il «Mondo 2» (cioè il dominio della soggettività e del pensiero). Sintetizzando, potremmo dire che, mentre per il neodarwinismo la selezione naturale è esclusivamente frutto di mutazioni genetiche casuali interne all'organismo e dell'azione di forze ambientali esterne all'organismo, al contrario, per l'interazionismo anche determinate preferenze e comportamenti, che nel tempo si radichino sino a divenire abitudini e tradizioni, svolgono un ruolo decisivo nell'influenzare «indirettamente le pressioni selettive alle quali [l'organismo] è esposto e, con esse, il risultato della selezione naturale»,[3] fino a creare nuove nicchie ecologiche e forme di vita. Tale teoria, conclude Popper, «mostra quindi il valore selettivo di una certa libertà comportamentale innata», e «può farci capire meglio in che modo sia emersa la mente umana».[4]

Passiamo dunque alle critiche che Popper muove al materialismo nelle versioni che abbiamo elencato. In particolare, cominciamo da un esempio le cui implicazioni torneranno utili anche più avanti, quando discuteremo del realismo platonico in matematica e in logica in relazione ai teoremi gödeliani. Popper, giustamente, afferma che i sostenitori del materialismo non sono in grado di fornire una spiegazione soddisfacente di come certi prodotti della mente e gli oggetti astratti che fanno parte di quello che egli chiama «Mondo 3» (come, ad esempio, le teorie o le relazioni logiche che regolano la comunicazione e il pensiero) possano retroagire su di essa. L'esempio che egli propone è quello della "scoperta" da parte di Russell di una contraddizione implicita nell'opera di Frege *Grundgesetze*

[3] Cfr. *ivi*, p. 24.
[4] Cfr. *ivi*, p. 25.

der Arithmetik (*I principi dell'aritmetica*). «Qualora la sua teoria non fosse stata *oggettivamente* incoerente» – dice Popper – «Frege non vi avrebbe potuto applicare la dimostrazione russelliana dell'incoerenza, per cui non si sarebbe convinto della sua insostenibilità».[5] Dall'oggettività propria dei concetti logici, risulta evidente, secondo l'autore, una certa forma di autonomia del Mondo 3 rispetto al Mondo 2, la quale implica una dipendenza reciproca che vige sul piano causale tra i due mondi. Ora, va notato che tale posizione si basa su di uno specifico presupposto [ontologico] [, ossia, per usare le parole di Quine, su di un certo "impegno ontologico"], tale per cui viene ammessa l'esistenza di qualcosa come il Mondo 3. Di contro, un fisicalista od un comportamentista potrebbe riformulare l'esempio precedente come segue: gli stimoli sensoriali dovuti alla lettura dell'opera di Frege da parte di Russell vengono tradotti dal suo sistema nervoso periferico in segnali elettrochimici diretti in entrata verso il cervello (*input*). Questi contribuiscono a modificare e rimodellare la struttura delle connessioni neurali creando una precisa configurazione (*pattern*) di scariche a livello sinaptico. Al termine del processo, il cervello invia a sua volta dei segnali in uscita (*output*) verso le zone periferiche del sistema nervoso, che li traduce in nuovi stimoli sensoriali diretti agli organi destinatari, deputati alla risposta. In questo caso la risposta consisterà nello scrivere una lettera con cui Russell informa Frege della contraddizione contenuta nel suo lavoro. Il fisicalista potrebbe continuare affermando che ciò che consente a Russell di fare questo – la memoria, ovvero ciò che ci consente di attribuire a noi stessi gli eventi di cui facciamo esperienza, e quindi di formare il senso unitario del nostro io – è spiegabile come un'estensione nel tempo del caso limite di una singola interazione con l'ambiente da parte del cervello. In altri termini, in linea di principio, non può immediatamente escludersi l'ipotesi che la struttura sinaptica, in una maniera a noi ancora sconosciuta nel suo complesso, sia in grado di conservare traccia delle configurazioni di connessioni sinaptiche pregresse e, quando necessario, di riattivarle sotto determinate condizioni che innescano il processo. Si noti che il livello a cui è possibile spiegare ciascuna interazione nella riformulazione materialistica dell'esempio di Popper è esclusivamente quello chimico e fisico.

Un'ovvia obiezione a tale ricostruzione riguarda la libertà creativa del pensiero di riportare alla mente ricordi del passato o di progettare con l'immaginazione eventi nel futuro, indipendentemente da ciò che accade all'esterno. A questa il fisicalismo potrebbe rispondere però dicendo che ciò che noi crediamo di poter pensare liberamente e, per così dire senza limiti, all'interno della nostra mente, altro non è che il modificarsi di una determinata configurazione della struttura sinaptica in un dato momento a causa di un'interazione complessa che risulta dall'unione di un condizionamento esterno (una nuova interazione ambiente-organismo) e di un condizionamento interno (la precedente configurazione sinaptica). Per fare un esempio, assumiamo che io mi trovi in

[5] Cfr. *ivi*, p. 76. Il corsivo è nostro.

una stanza e che in un dato tempo *t* il mio pensiero abbia un contenuto determinato, diciamo la percezione *x* di trovarmi proprio in questa stanza, la quale corrisponde ad una specifica configurazione *y* delle connessioni neurali all'interno del mio cervello. Successivamente, al tempo *t*+1, il contenuto del mio pensiero trapassa nel ricordo *z* di quando mi trovavo in gita a Venezia a cinque anni. La precedente configurazione *y* si trasforma allora in *w*, anche se il contesto in cui mi trovo, all'apparenza, non è cambiato per nulla. Come si spiega allora il fatto che io abbia richiamato alla memoria un ricordo della mia infanzia? Di certo, sostiene il fisicalista, non attraverso un atto di libertà da parte della mia mente. Infatti, il passaggio da *x* a *z*, e quindi il passaggio dalla configurazione *y* a *w*, è completamente determinato da un'interazione complessa che è il risultato, da un lato, di un mutamento delle condizioni esterne talmente minimo (poniamo pure della durata di frazioni di secondo) da essere impercettibile al mio organismo; dall'altro, del condizionamento interno costituito dalla stessa configurazione *y*. In questo modo possiamo vedere che il materialismo espunge totalmente la libertà dal dominio dell'azione umana, trincerando quest'ultima all'interno dei confini del determinismo.

Per quanto macchinosa e poco convincente possa sembrare una simile posizione, essa mostra come l'argomentazione di Popper, nel modo in cui è formulata, non riesca davvero nell'intento di confutare la visione fisicalista della chiusura del Mondo 1.[6] Naturalmente, resta una difficoltà non irrilevante, su cui torneremo tra poco, legata al fatto che il materialismo, qualora si ponga come una tesi determinata, entra in un rapporto dialettico con ciò che gli si oppone: esso è costretto ad elaborare una particolare spiegazione dei fenomeni mentali proprio a causa del loro darsi come qualcosa di esistente. In altre parole, il materialista, nel ricostruire i meccanismi fisici che sarebbero alla base del pensiero, non può prescindere, cadendo così in una sorta di circolarità, dal considerare, e dall'utilizzare quegli stessi oggetti astratti (*definiendum*) che i processi cerebrali (*definiens*) dovrebbero spiegare; oggetti che possiedono, peraltro, uno statuto ontologico specifico, che eccede quello delimitato dalla teoria materialista.[7]

---

[6] Anche l'osservazione di Popper circa le «conseguenze disastrose» che deriverebbero dall'intendere, alla maniera del fisicalista o del comportamentista, il linguaggio esclusivamente nella sua funzione espressiva o comunicativa, ovvero prescindendo dalla sua funzione descrittiva e argomentativa, anche questa osservazione, dicevamo, non costituisce una vera e propria critica alla prospettiva fisicalista. Tali conseguenze a cui Popper si riferisce sono costituite nient'altro che dal tralasciare «proprio tutto ciò che è caratteristico del linguaggio umano e lo distingue da quello animale: la sua capacità di fare asserzioni vere e false e di produrre argomenti validi e invalidi» (Cfr. *ivi*, p. 79). È evidente che questo non possa fare problema al fisicalista, poiché anzi è proprio ciò che egli è già disposto ad ammettere, assumendo, come abbiamo già detto, un impegno ontologico differente da coloro che invece credono nell'esistenza di qualcosa come gli stati coscienti e i suoi prodotti.

[7] Come vedremo nel prossimo paragrafo, Searle avanza un'obiezione molto simile al fisicalismo, in particolare alla tesi dell'IA forte, che riguarda il fatto che i concetti e le relazioni sintattiche attraverso cui opera un calcolatore non apparterebbero alla fisicità di quest'ultimo ma, al contrario, sarebbero soltanto «relative all'osservatore».

Da ultimo, vorrei concentrarmi sull'osservazione avanzata da Popper in base alla quale «la concezione epifenomenista» – ma lo stesso vale a maggior ragione anche per il fisicalismo radicale – «se applicata alle argomentazioni e alla nostra valutazione delle ragioni, è suicida».[8] Ciò è di grande importanza anche per il prosieguo del nostro lavoro dove, analizzando i risultati dei teoremi di incompletezza gödeliani, mostreremo alcune limitazioni intrinseche alla logica, e vedremo come la loro considerazione segni il ridimensionamento del progetto della sua autofondazione come sapere coerente e completo. La linea argomentativa seguita da Popper, il quale a sua volta si rifà ad un argomento del biologo J.B.S. Haldane, è sintetizzabile pressoché nel modo seguente: in quanto il fisicalismo, o l'epifenomenismo, intende porsi come teoria vera, o quantomeno fondata su criteri logici di validità, in tanto esso si contraddice *in actu exercito* ed è costretto ad affermare proprio ciò che si propone di negare, ovvero l'esistenza di un piano astratto, di un orizzonte concettuale, che corrisponde a quello che Popper chiama Mondo 2. L'argomento di Haldane, nella sua forma originaria, recita:

> se il materialismo fosse vero, mi sembra che, allora, noi non potremmo sapere che esso è vero. Se le mie opinioni sono il risultato dei processi chimici che avvengono nel mio cervello, sono determinate dalle leggi della chimica e non della logica.[9]

Tuttavia, un computer, sebbene funzioni in base alle sole leggi della fisica, ciononostante esegue operazioni di tipo logico. Ma ciò, fa notare Popper, dipende dal fatto che un computer viene progettato dalla mente umana, per cui la logicità del suo operare è soltanto "per noi", ovvero tale solamente se considerato dal nostro punto di vista. In particolare, vorrei soffermarmi su un passaggio del dialogo, che Popper immagina come riformulazione dell'argomento di Haldane, tra un interazionista e un fisicalista in merito alla logica. Il punto centrale della discussione è che, dal punto di vista del fisicalismo, nozioni come quella di verità o di validità di un'inferenza appartengono ad una dimensione astratta (il Mondo 3), che esiste soltanto come interpretazione di quello che in realtà è il risultato complesso di un lungo processo e di apprendimento adattivo e di selezione naturale. In altre parole, la razionalità e la logica altro non sono che il frutto dell'adattamento selettivo dei nostri cervelli all'ambiente circostante e del condizionamento sociale che veicola l'apprendimento e l'istruzione. Se assumiamo questo, allora l'errore logico è tale relativamente a criteri che sono validi esclusivamente sulla base della loro utilità selettiva nella lotta per la sopravvivenza, e di conseguenza

[8] Cfr. K. R. Popper – J. C. Eccles, *L'Io e il Suo Cervello*, cit., p. 96.
[9] Ho ripreso la citazione dal testo di Popper. Cfr. *ivi*, p. 97.

la verità di una teoria viene a coincidere con la sua utilità.[10] Ciò però comporta l'impossibilità di disporre di un criterio per stabilire se ogni inferenza logica sia valida, e quindi utile. Infatti, come rileva Popper, un'inferenza logica di tipo deduttivo è valida se il «contenuto informativo vero» viene trasmesso dalle premesse alla conclusione, e in modo tale per cui esso non può essere contenuto in misura maggiore nell'ultima rispetto alle prime.[11] In altri termini, un'inferenza deduttiva è tautologica, ovvero possiede un contenuto informativo vero ma inutile. Per il fisicalista, dunque, l'unico modo per rendere conto dell'utilità di ogni singola inferenza, e quindi anche delle inferenze deduttive, è quello di dare per presupposta l'utilità dell'insieme di tutte le inferenze possibili, ovvero del sistema della logica nella sua totalità. Tuttavia, un tale insieme costituisce proprio quel tipo di oggetto astratto di cui il fisicalista nega l'esistenza. Quest'ultimo, perciò, rinunciando all'assunzione di un criterio astratto che non sia arbitrario o convenzionale, ma che piuttosto si rifaccia ad un orizzonte fondativo, non è in grado di rendere ragione dell'universalità della propria posizione. Quello che intendiamo dire è che, in quanto la visione fisicalista avanza la pretesa di porsi come teoria "vera di contro" alle altre, in tanto è costretta a trascendere i suoi stessi presupposti, e quindi a rigettare l'utilità quale unico criterio di validità/verità.

[10] Anche Penrose, mediante un ragionamento analogo a quello di Popper, rifiuta la riduzione della verità, seppur in riferimento alle proposizioni matematiche, all'utilità nella lotta per la sopravvivenza. Si veda il paragrafo §6.4.

[11] In proposito, infatti, va ricordato che oltre a quelle di tipo deduttivo, esistono inferenze di tipo induttivo, dialettico e apagogico, che non si limitano ad esplicitare, analiticamente, il contenuto implicito nelle premesse. Inoltre, va notato che esistono verità analitiche che, sebbene tautologicamente conseguenti, non sono immediatamente evidenti a partire dai postulati iniziali.

Capitolo 3

Il "naturalismo biologico" e la critica di Searle alla tesi dell'intelligenza artificiale

Riteniamo opportuno, a questo punto, esaminare anche alcuni aspetti chiave della visione del filosofo americano J.R. Searle in merito al fenomeno della coscienza. Egli, infatti, è noto non solo per i suoi contributi in filosofia del linguaggio, ma anche per la sua posizione originale in merito al dibattito intorno al problema mente-corpo e al programma dell'intelligenza artificiale: i due temi principali che sono oggetto di interesse del presente lavoro.

Uno dei problemi filosofici contemporanei centrali, per Searle, è costituito dalla relazione, o interazione, tra quelle che egli definisce «realtà umana» e «realtà di base». La realtà umana è rappresentata dall'insieme di fenomeni mentali, linguistici, e culturali che intessono il nostro modo di vivere individuale e di comunità, e corrisponde, all'incirca, ai Mondi 2 e 3 del discorso popperiano. La realtà di base, invece, è rappresentata dal mondo fisico come le teorie fisiche di cui oggi disponiamo ce lo descrivono, ovvero come costituito da particelle subatomiche fondamentali che interagiscono tra loro secondo delle leggi dettate dalla specifica natura dei campi di forza in cui esse sono immerse. Searle sposa una concezione che chiama «naturalismo biologico», secondo la quale l'insieme dei fenomeni o degli eventi coscienti non costituisce un dominio della realtà indipendente e totalmente altro da quello naturale ma, al contrario, ne è parte integrante. Da questo punto di vista, la coscienza non è altro che un fenomeno biologico, risultante dallo sviluppo naturale del mondo fisico. Con ciò, i concetti di realtà e di natura, secondo Searle, vengono a coincidere.

Vi sono due presupposti alla base di questa visione: la teoria fisica delle particelle e quella biologica dell'evoluzionismo neodarwiniano. Come abbiamo visto, anche Popper basa la propria teoria dell'interazionismo emergente sui medesimi assunti. Tuttavia, mentre l'interazionismo si propone di risolvere il problema mente-corpo assumendo un'ontologia che non è né monista né dualista, bensì pluralista, per la quale esistono tre diversi domini della realtà, irriducibili l'uno all'altro ma pur sempre interagenti, il naturalismo biologico, a nostro avviso, si presenta come una versione della teoria dell'identità, e quindi, stando alla classificazione che ne dà Popper, come una forma di materialismo parallelista. Infatti, sebbene Searle riconosca l'influenza che l'emergere della coscienza esercita sui processi di selezione naturale, nonché la capacità degli stati mentali di produrre degli effetti sulla realtà fisica, ciononostante egli intende il fenomeno della coscienza come una proprietà fisica del mondo che è «*causata dai* processi neuronali che avvengono nel cervello, e che […] *esiste*

nel cervello».[12] Una determinata interazione causale tra stati mentali e ambiente costituisce un evento che «può essere descritto ad almeno due livelli: un livello cosciente […], e un livello neurobiologico. È assolutamente essenziale rendersi conto che qui non si tratta di due eventi distinti, ma piuttosto di un unico evento che ammette due livelli di descrizione».[13] Al pari dei teorici dell'identità dunque, anche Searle tenta di dissolvere, per così dire, il problema filosofico del rapporto tra mente e corpo ammettendo una relazione di identità, anche se non di tipo logico, tra gli stati coscienti e i processi neurochimici che sembrano starne alla base.[14] In questo modo, egli crede di evitare quello che ritiene l'errore commesso da due grandi tradizioni di pensiero – il dualismo cartesiano e il materialismo scientifico –, ovvero aver negato alla coscienza l'appartenenza al mondo naturale. Tuttavia, come fa notare Popper, nemmeno la teoria dell'identità, per così dire, sfugge alla critica da lui mossa al materialismo, poiché essa concepisce il rapporto tra coscienza e processi neuronali non come una vera e propria identità logica (il che, come abbiamo detto, sarebbe comunque insensato), ma come una forma di parallelismo di tipo spinoziano. Spinoza, infatti, parlava di due differenti *modi*, entrambi veri, di pensare l'unica Sostanza o Natura: l'uno dall'interno (come anima o mente), l'altro dall'esterno (come corpo fisico). Per i teorici dell'identità, dunque, non sussiste una vera e propria interazione se non tra il mondo fisico e i processi cerebrali paralleli e sottostanti ai corrispettivi stati di coscienza. Essi, perciò, non sono in grado, come vorrebbero, di rendere conto del ruolo fondamentale che la coscienza ha nell'influire il corso della selezione naturale nel processo evolutivo dell'uomo.

Un altro importante aspetto da rilevare è che la posizione di Searle ammette la possibilità di ricreare artificialmente la coscienza in una macchina (sia di tipo organico che non-organico) attraverso la ricostruzione dettagliata dei meccanismi che regolano il funzionamento del cervello. Nonostante egli riconosca che, di fatto ci sia ancora molta strada da fare, tuttavia, egli crede non soltanto che la direzione intrapresa dalla ricerca nel campo della neurobiologia sia quella corretta, ma che i suoi risultati debbano essere tenuti in forte considerazione da parte della riflessione filosofica, la quale deve servirsi di essi nell'indagine intorno al fenomeno della coscienza da un punto di vista logico e metafisico. È però doverosa una precisazione in merito a questo punto: sebbene la ritenga attuabile in linea di principio, Searle nega che la creazione di un'autentica forma di intelligenza

[12] Cfr. J.R. Searle, *Il Mistero della Realtà*, trad. it. di P. Di Lucia e L. P. Glazel, Raffaello Cortina Editore, Milano 2019, p. 42; tit. orig. *The Basic Reality and the Human Reality*, 2019. Il corsivo è dell'autore.

[13] Cfr. *ivi*, p. 43.

[14] «La relazione tra questi due eventi [lo stato cosciente e il corrispondente processo neurobiologico] è una relazione di identità, e questo è uno dei fatti più importanti relativi a come funziona il mondo» (Cfr. J. R. Searle, *Il Mistero della Realtà*, cit., p. 44). Nell'introduzione al testo nella traduzione italiana leggiamo: «Compito della filosofia è […], secondo Searle, quello di indagare, avvalendosi anche delle scoperte della neurobiologia, la *forma logica* dei processi e delle esperienze della coscienza […]» (Cfr. *ivi*, p. XXII). Il corsivo è dell'autore.

artificiale (che sia perciò cosciente) possa essere conseguita soltanto sulla base di un modello computazionale. Egli, infatti, sostiene che un programma (*software*), implementato in una macchina o in un computer, che esegue delle computazioni regolate da procedure algoritmiche, sia limitato essenzialmente alla manipolazione sintattica di segni e simboli primitivi del cosiddetto "linguaggio macchina"; esso sarebbe, in altre parole, privo della semantica e dell'intenzionalità che rappresentano invece caratteristiche fondamentale del pensiero e del linguaggio umani. Inoltre, egli giunge perfino a dire che la macchina, in sé stessa, non avrebbe nemmeno una sintassi, né eseguirebbe veri e propri calcoli computazionali: tutte queste cose, in realtà, sarebbero interpretazioni di un soggetto osservatore. Quest'ultimo sarebbe il solo ad avere una sintassi e una semantica, e ad essere perciò capace, per così dire, di "imprimerle" nei circuiti di un calcolatore. Per questo motivo, sulla base di quanto appena detto, Searle rigetta la versione cosiddetta "forte" del programma dell'intelligenza artificiale. Quest'ultimo consiste in un progetto di ricerca diretto alla realizzazione di programmi implementabili in una macchina (o un computer), che siano in grado di eseguire compiti (in particolare modo di tipo cognitivo) con lo stesso grado (se non addirittura maggiore) di accuratezza ed efficienza con il quale vengono eseguiti dagli esseri umani. La versione forte di tale progetto, dunque, ritiene che la mente umana sia essenzialmente "identica" ad un programma, o ad un insieme di istruzioni algoritmiche, talmente complesso da non essere comprensibile per la mente stessa. La critica di Searle, invece, non è rivolta alla versione debole del programma dell'IA, la quale sostiene soltanto la possibilità, in linea di principio e non sulla base degli attuali risultati della ricerca, di realizzare una "simulazione" computazionale dei fenomeni mentali. A questo proposito, è perciò fondamentale distinguere tra la nozione di duplicazione e quella di simulazione. La duplicazione della coscienza, al contrario della sua simulazione, sarebbe possibile, secondo Searle, solo a patto di ricostruire esattamente l'insieme di meccanismi che determinano i processi cerebrali e da cui emergono gli stati mentali; per questo motivo, un semplice modello (o simulazione) computazionale della mente non costituirebbe una condizione né sufficiente né necessaria a tale scopo.

Da ultimo una precisazione. Nel corso di questa prima parte del lavoro, ci siamo permessi di introdurre soltanto provvisoriamente una serie di termini e nozioni quali "algoritmo", "programma dell'intelligenza artificiale", "computazione", al fine di poterli riprendere e di fornirne un'analisi più dettagliata nella seconda parte, dove svilupperemo il tema centrale che funge da *fil rouge* per il presente lavoro: il dibattito contemporaneo intorno a quelli che vengono chiamati "argomenti gödeliani", ovvero quegli argomenti che, prendendo le mosse dai risultati dei celebri teoremi di incompletezza del logico austriaco Kurt F. Gödel, affermano l'impossibilità logica di realizzare una simulazione computazionale esatta e completa della mente umana. Le argomentazioni del filosofo oxoniense John. R. Lucas e del fisico britannico Roger Penrose sono in proposito le più note, nonché

quelle che prenderemo in esame. Oltre a queste, analizzeremo anche alcune delle più importanti critiche che sono state mosse loro da diversi autori, come, ad esempio, Hilary Putnam, Paul Benacerraf e Stewart Shapiro. Per essere in grado di seguire le diverse fasi di questo dibattito è necessaria, tuttavia, una preliminare comprensione non solo del teorema di Gödel e del suo corollario (ormai noti, rispettivamente, come il primo e il secondo teorema di incompletezza), ma anche del retroterra culturale in cui il lavoro di Gödel si inserisce, almeno per quanto riguarda i loro aspetti fondamentali. Il prossimo capitolo, dunque, sarà dedicato alla loro ricostruzione.

Parte II

I TEOREMI DI INCOMPLETEZZA DI GÖDEL: UNA SVOLTA NEL DIBATTITO INTORNO ALLA NATURA DELLA MENTE UMANA

Capitolo 1
La "crisi dei fondamenti" e il programma formalista di Hilbert

Prima di introdurre i teoremi di incompletezza ed analizzare le loro profonde implicazioni sul piano logico e filosofico, è necessario illustrare in che cosa consiste quello che viene definito il "programma della coerenza", o anche "programma formalista", introdotto dal matematico tedesco David Hilbert, per acquisire una migliore comprensione del terreno di ricerca su cui il successivo lavoro di Gödel si è delineato.

A partire dai primi anni del Novecento, Hilbert incominciò ad interessarsi del problema inerente alla fondazione del sapere matematico e logico; questione che era divenuta notevolmente rilevante già nel corso dell'Ottocento, quando era stato messo in discussione il metodo assiomatico euclideo.[15] Quest'ultimo si fonda sulla «convinzione tradizionale che gli assiomi della geometria (o […] di qualunque sistema) possano essere provati dalla loro apparente autoevidenza»;[16] in altri termini, che la verità degli assiomi, e quindi la loro legittimità come fondamento ultimo della geometria, derivi dalla loro stessa evidenza. Va qui sottolineato che la nozione di evidenza, legata a quella di verità, rimanda ad una forma di "intuizione intellettuale" attraverso la quale si suppone la capacità di attestare la corrispondenza tra determinati assiomi e teoremi di una teoria, da un lato, e la realtà dall'altro.

Nagel e Newman evidenziano che la dimostrazione (da parte di Gauss, Bolyai e Lobacevskij) dell'«*impossibilità* di dedurre l'assioma delle parallele dagli altri assiomi»[17] della geometria euclidea, conduceva alla problematizzazione dell'assiomatica classica e alla formulazione delle geometrie

[15] Per un approfondimento della questione riguardante la crisi dei fondamenti in matematica e logica e dei contributi apportati da Hilbert, raccomandiamo la lettura dei seguenti testi: E. Nagel, J.R. Newman, *La prova di Gödel*, trad. it. di L. Bianchi, Bollati Boringhieri, Torino 2013; tit. orig. *Gödel Proof*, 1958; E. Agazzi, *The Rise of the Foundational Research in Mathematics*, in «Synthese», XXVII, 1974, no. 1/2, pp. 7-26; P. Mancosu, *Il programma di Hilbert e i teoremi di incompletezza di Gödel*, in «Rivista di Filosofia Neo-Scolastica», XCVIII, 2006, no. 3, pp. 489-531.

[16] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 21.

[17] Cfr. *ibidem*. Il corsivo è degli autori.

cosiddette “non-euclidee”, oltre che alla revisione dei fondamenti di altre discipline. Da ciò nacque una concezione nuova della matematica come disciplina “pura” e “formale”, centrata sulla sola “coerenza” logica con cui le deduzioni devono essere sviluppate a partire dagli assiomi, la cui validità, quindi, non deve dipendere dal «particolare significato che può essere associato ai termini o alle espressioni contenute nei postulati».[18] La coerenza (o consistenza) di una teoria, come ad esempio una teoria geometrica non-euclidea, dipende, o meglio, è dimostrabile a partire da una certa “interpretazione”, o anche “modello”, che soddisfi i suoi postulati, ovvero tale per cui questi ultimi risultino veri. Inoltre, per una teoria formale, solitamente, esiste un modello *standard*, o “intuitivo”, come, ad esempio, l’insieme $\mathbb{N}$ dei numeri naturali rappresenta l’interpretazione intuitiva dell’aritmetica formalizzata (l’aritmetica ricorsiva primitiva). Il problema, perciò, divenne quello di stabilire la coerenza di un determinato sistema di sapere, ovvero quello di stabilire se sia possibile o meno dedurre al suo interno teoremi contraddittori (problema che, ovviamente, non si pone nel caso del metodo assiomatico classico). Come spiega Evandro Agazzi, l’abbandono dell’intuizione, quale criterio necessario per la verità della matematica, «condusse ad una reale crisi, quando venne provato che la geometria euclidea e [quella] non-euclidea avevano lo stesso grado di consistenza». «Questa crisi» – continua l’autore – «ebbe inizio nel 1868, quando Eugenio Beltrami sviluppò il primo modello euclideo di una geometria non-euclidea».[19] Nel caso della geometria non-euclidea riemanniana, ad esempio, è possibile mostrare che la sua coerenza dipende da uno specifico modello che è traducibile nei termini della geometria classica. Analogamente, anche l’interpretazione aritmetica della geometria (la geometria analitica introdotta già da Cartesio), la quale fu il punto di partenza da cui Hilbert tentò di risolvere la questione, non fa che riproporre il medesimo problema, e sposta la questione dalla prova della coerenza della geometria a quella dell’aritmetica (la quale, come spiega Mancosu, è equivalente, in un senso generale, alla teoria dei numeri reali).

In particolare, fu in seguito alla deriva del “programma logicista” portato avanti dal logico Gottlob Frege, a cavallo tra Ottocento e Novecento, che Hilbert decise di dirigere i propri sforzi nel tentativo di elaborare una soluzione alla cosiddetta «crisi dei fondamenti» sviluppatasi in quegli anni. L’obiettivo del logicismo era quello di fondare l’aritmetica sulla teoria degli insiemi. A quell’epoca, infatti, la nozione di insieme sembrava essere ancora più intuitiva e primitiva di quella di numero naturale. Tuttavia, come divenne evidente in seguito a un celebre scambio tra il matematico Bertrand Russell e Frege del 1902, da tale nozione, proprio a causa della vaghezza con cui era allora formulata, sorgeva una serie di paradossi e contraddizioni logiche imputabili, in ultima istanza, al carattere

[18] Cfr. *ibidem*.
[19] Cfr. E. Agazzi, *The Rise of the Foundational Research in Mathematics*, cit., p. 12.

"autoreferenziale" di alcuni enunciati.[20] Per risolvere un tale problema legato alla fondazione del sapere matematico e della logica, Hilbert si rese conto che, per ottenere una dimostrazione diretta della coerenza dell'aritmetica, era necessario abbandonare ogni tipo di appello all'evidenza e a nozioni apparentemente intuitive. Come riporta Mancosu, in un corso di lezioni del 1905, Hilbert fornisce una definizione generale della nozione di struttura che sta alla base della nuova assiomatica:

> I numeri diventano per noi solo una struttura di concetti alla quale, certamente, si arriva solo per mezzo dell'intuizione; tuttavia possiamo operare con questa struttura senza fare ricorso all'intuizione. Ma per garantire l'applicabilità agli oggetti che ci circondano, questo sistema concettuale è costruito in modo che esso sia ovunque in completa analogia con le più semplici intuizioni ed attraverso esse con i fatti d'esperienza.[21]

Inoltre, il metodo hilbertiano si pone in netto contrasto con quello seguito da altri matematici dell'epoca, in quanto prescinde dalla costruzione di un modello interpretativo che soddisfi l'insieme assiomatico della teoria. Come evidenzia Mancosu, «Hilbert non solo richiede una dimostrazione di consistenza per il sistema dei numeri reali ma vuol spingere il suo programma fino a render conto dei numeri naturali».[22] Ed è evidente che un simile intento derivi dal bisogno di non commettere il medesimo errore compiuto dai logicisti rispetto alla nozione intuitiva di insieme o classe.[23]

Ma è soltanto a partire dagli anni Venti che la riflessione di Hilbert, grazie anche alla collaborazione con il suo assistente Paul Bernays, acquisisce maggiore spessore e rigore. Nel saggio del 1922, intitolato *Nuovi fondamenti della matematica*, Hilbert definisce un elemento cardine del programma formalista: il "segno", inteso come entità primitiva pre-logica:

> Come prerequisito per l'uso delle inferenze logiche e per il funzionamento delle operazioni logiche, ci deve essere dato già qualcosa nella rappresentazione […]: certi oggetti discreti extralogici, che esistono intuitivamente come esperienze immediate prima di ogni pensiero. Se il ragionamento logico deve essere sicuro, questi oggetti devono essere completamente dominabili in tutte le loro parti e, insieme con gli oggetti, la loro esibizione, la loro distinzione, il loro susseguirsi ci sono dati in modo immediatamente intuitivo come qualcosa che non è riducibile ancora a qualcos'altro. Assumendo questa posizione, per me […] gli oggetti

[20] Cfr. F. Berto, *Tutti pazzi per Gödel!*, Laterza, Bari 2019, p. 36.

[21] La citazione si riferisce alla traduzione di Mancosu del passo contenuto nel dattiloscritto originale (non pubblicato) di Hilbert. Cfr. P. Mancosu, *Il programma di Hilbert e i teoremi di incompletezza di Gödel*, cit., p. 493.

[22] Cfr. *ivi*, p. 494.

[23] Mancosu spiega che questo punto era già stato messo in evidenza da H. Behmann, studente di dottorato di Hilbert, nella sua tesi di ricerca intorno ai fondamenti dell'aritmetica. Scrive Mancosu: «il progetto fondazionale difeso da Behmann consiste nel ridurre l'appello alle entità astratte a una costruzione simbolica senza alcuna portata ontologica» (Cfr. *ivi*, p. 503).

della teoria dei numeri sono proprio i segni, la cui forma può essere da noi riconosciuta e sicuramente, indipendentemente dallo spazio e dal tempo, e dalle condizioni particolari della costituzione del segno, così come da insignificanti differenze nell'esecuzione. […] *In principio c'è il segno*, questo è il motto.[24]

Sebbene lo statuto ontologico dei segni primitivi non sia stato mai completamente chiarito da Hilbert, tuttavia, dal passo sopra riportato, risulta evidente che non può trattarsi né di oggetti fisici (è possibile, infatti, estendere all'infinito l'insieme dei numerali, cioè i nomi con cui si denotano i numeri naturali) né di costrutti mentali come concetti o insiemi (dal momento che le loro proprietà sono oggettive). Dunque, gli assiomi di una teoria non devono più assumersi come espressivi di una verità intuitiva o immediata che rimanda ad una realtà esterna, come ad esempio nel caso della geometria euclidea; in questo senso, l'intuizione a cui si riferisce Hilbert non poggia su un dato d'evidenza intellettuale, bensì su un dato sensibile immediato, privo di referente oggettuale.

Hilbert credeva, inoltre, che il sapere matematico sia formalizzabile in un'unica teoria coerente e completa, tale per cui non sia derivabile alcuna contraddizione al suo interno e sia possibile risolvere "ogni" problema matematico (data una qualsiasi formula, cioè, decidere se tale formula sia un teorema oppure no). Ma, ciò che è più importante, egli riteneva necessario dimostrare la coerenza della matematica sulla base di sole procedure finitarie, che prescindono cioè, al contrario di quelle che Hilbert chiama "transfinite", dall'appello a nozioni intuitive, o di carattere semantico, che presuppongono un dominio oggettuale d'interpretazione infinito. Infatti, come Nagel e Newman mettono bene in luce, il problema di fondo riguardante la questione fondazionale «consiste nel fatto che gli assiomi vengono interpretati mediante modelli composti da un numero infinito di elementi». «Ciò rende impossibile» – continuano i due autori – «abbracciare i modelli con un numero finito di osservazioni, per cui la verità degli assiomi stessi è messa in dubbio».[25] In altri termini, per dimostrare in modo diretto la coerenza di una teoria, senza cioè che quest'ultima presupponga come già acquisita la coerenza di un'altra, sarebbe necessario disporre di un modello costituito da un insieme finito di elementi, tale per cui siano pienamente controllabili tutte le relazioni vigenti tra di essi.

A questo punto, riteniamo opportuno rifarci alla lezione di Sergio Galvan, oltre che alla terminologia da lui utilizzata, per meglio comprendere alcuni aspetti-chiave del programma formalista, e per spiegare i concetti di formalizzazione e di assiomatizzazione di una teoria, nonché esplicitare alcune caratteristiche dell'aritmetica primitiva ricorsiva e della finitarietà delle sue procedure dimostrative.[26]

---

[24] La citazione si riferisce alla traduzione di Mancosu del passo contenuto nello scritto originale hilbertiano. Cfr. *ivi*, pp. 508-509.

[25] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 31.

[26] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit.

Una teoria T si dice formalizzata quando è composta da:

a) un insieme di assiomi *A*(T);
b) un linguaggio artificiale formalizzato L su cui la teoria è impiantata, il quale è formato da un alfabeto A[27] e da regole di formazione F relativamente ai termini e alle formule di L;
c) un insieme di regole d'inferenza o deduttive D;
d) un calcolo logico *k* attraverso il quale è possibile dimostrare i teoremi della teoria, il quale è, a sua volta, determinato da L e D.

In particolare, il linguaggio formale di una teoria formalizzata consta di un insieme finito di segni, o simboli, primitivi, la cui combinazione e manipolazione in base alle regole di formazione determinano delle stringhe, o sequenze, finite che si possono specificare in:

a) formule;
b) assiomi;
c) regole d'inferenza;
d) dimostrazioni formali (sequenze finite di formule);
e) teoremi (formule dimostrabili all'interno del sistema).

Come rileva Galvan, soltanto una teoria assiomatizzabile è puramente formale, e «una teoria è detta assiomatizzata se e solo se l'insieme degli assiomi *A*(T) è *decidibile* (ossia se esiste un procedimento meccanico capace di decidere, data una qualunque [formula] $\alpha \in L(T)$, se α è un assioma o no)».[28] All'interno di una teoria formale assiomatizzata, le relazioni tra le sequenze di

[27] L'alfabeto di un linguaggio formalizzato contiene:
variabili individuali (x, y, z, …);
costanti funzionali (o segni funzionali). f_k^n, ad esempio indica la k-esima costante funzionale ad n posti;
costanti predicative (o segni per proprietà o relazioni). P_k^n, ad esempio, indica la k-esima relazione tra n individui;
segni logici:
identità (=)
connettivi proposizionali ($\neg$, $\wedge$, $\vee$, $\rightarrow$)
quantificatori universale ($\forall$) ed esistenziale ($\exists$);
segni ausiliari (parentesi).

[28] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 56. Il corsivo è dell'autore. Quella di decidibilità è una nozione intuitiva, e indica una procedura meccanica che, rispettando le regole della grammatica del linguaggio di una teoria formale, consente di stabilire se una certa formula appartiene all'impianto assiomatico della teoria o deriva da essa sulla base di qualche regola d'inferenza. Mostreremo in seguito che, nelle teorie assiomatizzate, la proprietà di decidibilità di una formula corrisponde alla proprietà di ricorsività primitiva.

simboli sono di natura puramente sintattica, e non anche semantica, e sono quindi pienamente "visibili". Perciò, la procedura di assiomatizzazione di una teoria ne stabilisce l'indipendenza da dati dell'intuizione o d'evidenza, e semplifica l'analisi delle relazioni logiche vigenti tra gli elementi del sistema formale. Possiamo quindi introdurre le seguenti definizioni relativamente ad una teoria formale T:

- Coerenza (o consistenza) sintattica: equivale alla non derivabilità di due formule contraddittorie (α e $\neg\alpha$) a partire dagli assiomi di T. Va notato che l'incoerenza di una teoria implica la derivabilità di qualunque formula a partire dai suoi assiomi.
- Completezza sintattica: equivale alla decidibilità di ogni formula α espressa nel linguaggio di T, ovvero alla derivabilità o di α o di $\neg\alpha$.
- Coerenza semantica (o correttezza): equivale alla derivabilità in T delle sole proposizioni vere rispetto alla sua interpretazione standard. È opportuno sottolineare che la coerenza semantica è detta, in gergo, "più forte" di quella sintattica, dal momento che la derivabilità di una contraddizione implica necessariamente la falsità di almeno una delle due formule contraddittorie.
- Completezza semantica: equivale alla derivabilità in T di tutte e sole le proposizioni vere rispetto alla sua interpretazione standard.

Una teoria si dice numerica, poi, quando contiene la costante individuale «0» e il simbolo funzionale di successione «'». Mediante tali simboli, infatti, è possibile introdurre per induzione l'insieme dei termini numerali,[29] ovvero le rappresentazioni formali dei numeri naturali dell'insieme $\mathbb{N}$. Infine, una teoria numerica può dirsi "ω-consistente" ed "ω-completa".[30] Stando alle definizioni di Galvan:[31]

T è una teoria numerica ω-consistente $=_{def}$ (i) T è una teoria numerica e (ii) se $A(T) \vdash \alpha(0)$, $A(T) \vdash \alpha(\overline{1})$, … allora $A(T) \nvdash \exists\neg\alpha(x)$.

[29] I numerali si ottengono applicando un numero finito di volte (a partire da 0 volte) la funzione di successione alla costante zero. Solitamente con $\overline{k}$ si indica il k-esimo numerale, ossia il termine «0'… …'», ove la funzione di successione è applicata k volte (Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 59).

[30] Berto spiega che l'ω-coerenza di una teoria numerica si dimostra anche confutando la sua ω-incoerenza. In particolare, T si dice ω-incoerente se, e solo se, esiste una variabile x che soddisfa una formula α, ma per ogni numero naturale *n* vale: $\vdash_T \neg\alpha(x/\overline{n})$; ovvero $\vdash_T \neg\alpha(x/0)$, $\vdash_T \neg\alpha(x/\overline{1})$, $\vdash_T \neg\alpha(x/\overline{2})$, …, ecc. (Cfr. F. Berto, *Tutti pazzi per Gödel!*, p. 117).

[31] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 59. Si noti che la nozione di ω-consistenza è più forte di quella di consistenza. In altri termini, l'ω-consistenza è condizione sufficiente, ma non anche necessaria, per la consistenza.

T è una teoria numerica ω-completa $=_{def}$ (i) T è numerica e (ii) se $A(T) \vdash \alpha(0)$, $A(T) \vdash \alpha(\bar{1})$, … allora $A(T) \vdash \forall x\alpha(x)$.

Su questo è opportuna una breve digressione circa la nozione di decidibilità riferita a un insieme di elementi che soddisfano una certa relazione, proprietà, o funzione di una teoria formale. La tesi dovuta al matematico Alonzo Church congettura che le funzioni effettivamente calcolabili, ovvero decidibili, siano equivalenti alle funzioni ricorsive primitive.[32] Questa ipotesi, della cui plausibilità parleremo nel capitolo quarto, consente di estendere il predicato di ricorsività anche ai predicati e alle relazioni numeriche rappresentabili nel linguaggio di una teoria numerica, in quanto quest'ultime sono esprimibili mediante quella che viene definita la loro "funzione caratteristica", la quale è una funzione ricorsiva primitiva. Prendiamo ora in considerazione l'aritmetica ricorsiva primitiva, ovvero quella «parte dell'aritmetica che costituisce la base della teoria della computabilità effettiva e che trova la formalizzazione più soddisfacente nel sistema formale PRA [l'aritmetica ricorsiva primitiva di Skolem]».[33]. LPR è un linguaggio del primo ordine[34] con identità senza quantificatori, contenente tutte e sole le funzioni e le relazioni ricorsive primitive, ed è soddisfatto

[32] Per quanto riguarda la definizione di funzione ricorsiva primitiva, Galvan scrive: «le funzioni ricorsive primitive sono quelle che si possono introdurre induttivamente, a partire dalla serie dei numeri naturali (0, 1, 2, …) e dalla funzione di successione ('), tramite opportune equazioni definitorie» (Cfr. *ivi*, p. 63). E ancora «Oltre alla funzione di successione ', che è la funzione ricorsiva primitiva iniziale, si dicono *funzioni ricorsive primitive* tutte e sole le funzioni numeriche che si possono introdurre per mezzo degli schemi di sostituzione e di recursione» (Cfr. *ivi*, p. 65). Inoltre, spiega Galvan, «si dicono *termini ricorsivi primitivi* tutti e soli i termini appartenenti ad LPR. […] Naturalmente la nozione di termine ricorsivo primitivo è molto vicina a quella di funzione ricorsiva primitiva: ogni segno funzionale, seguito da una opportuna stringa di variabili, è un termine e ad ogni termine si può far corrispondere un segno funzionale» (Cfr. *ivi*, pp. 68-69). Per quanto riguarda le relazioni ricorsive primitive, Galvan spiega che «una relazione è ricorsiva primitiva se le sue condizioni di verità nel modello standard coincideranno con i valori di una determinata funzione ricorsiva primitiva» (Cfr. *ivi*, p. 77). Consideriamo, dunque, la definizione di "funzione caratteristica" di una relazione numerica generica: «sia data una relazione numerica $R(x_1 \ldots x_n)$; allora si dice funzione caratteristica di $R(x_1 \ldots x_n)$ una funzione, che si indica con $C_R(x_1 \ldots x_n)$, tale che:

$C_R(x_1 \ldots x_n) = 0$ se $R(k_1 \ldots k_n)$ è vero
$C_R(x_1 \ldots x_n) = 1$ se $R(k_1 \ldots k_n)$ è falso.

[…] Una relazione numerica è ricorsiva primitiva se e solo se la sua funzione caratteristica è ricorsiva primitiva, il che implica che la relazione sia anche decidibile. […] La funzione caratteristica fornisce, in altri termini, una procedura di decisione. È questo il motivo per il quale la ricorsività primitiva di un predicato [o relazione] implica la sua decidibilità» (Cfr. *ivi*, p. 78).

[33] Cfr. *ivi*, p. 63.

[34] Un linguaggio formale si dice del primo ordine quando i quantificatori universale ed esistenziale sono applicabili soltanto alle variabili, mentre si dice del secondo ordine quando i quantificatori sono applicabili anche ad insiemi e sottoinsiemi di variabili. In particolare, LPR contiene, oltre ai termini numerali, anche le costanti individuali numeriche 0,1,2, …. Naturalmente, rispetto al modello standard di LPR, i numerali designano gli stessi oggetti designati dalle costanti individuali numeriche o dai simboli metateorici corrispondenti (k, m, n, …). Perciò, k e $\bar{k}$ designano lo stesso numero naturale.

dall'interpretazione intuitiva sulla struttura (o modello) dei numeri naturali. L'insieme delle formule chiuse di LPR che sono vere rispetto a tale struttura costituisce quello che Galvan definisce il «sistema semantico ARP» (non assiomatizzato) dell'aritmetica ricorsiva primitiva. È evidente, dunque, che l'appartenenza di una proposizione al sistema ARP non possa essere determinata soltanto sulla base di procedure finitarie, bensì alluda ad un criterio semantico di verità. Come nota Galvan, ARP costituisce, perciò, una teoria né assiomatizzata né (sulla base dei teoremi di Gödel che analizzeremo in seguito) assiomatizzabile.[35] Ciononostante, uno degli obiettivi del progetto formalista era proprio quello di fornire un'assiomatizzazione dell'aritmetica ricorsiva primitiva al fine di dimostrarne la consistenza attraverso metodi finitari.

A proposito della finitarietà della prova di consistenza per l'aritmetica, bisogna esaminare un aspetto fondamentale del programma hilbertiano: la distinzione tra la matematica "reale" e quella "ideale". La prima si riferisce ad enti matematici finiti e concreti, come i numeri naturali, e le dimostrazioni formali concernenti tali oggetti. La seconda, invece, si riferisce ad oggetti matematici astratti non-esistenti, come ad esempio i concetti di insiemi infiniti, e le dimostrazioni che vertono intorno ad essi. Hilbert e Bernays non furono mai del tutto chiari intorno ai requisiti che avrebbero dovuto caratterizzare il concetto di "finitarietà" nell'ambito della matematica. Tuttavia, gli interpreti sembrano concordare sul fatto che, nell'ottica formalista, le uniche proposizioni ammissibili nella dimostrazione di consistenza per la matematica, cioè le proposizioni reali, non devono contenere elementi linguistici che presuppongano l'esistenza in atto di un insieme infinito di elementi. Le proposizioni che formano la matematica ideale, invece, ovvero quelle che, al contrario, possiedono un contenuto infinitario, devono poter essere ridotte, o giustificate, nei termini della matematica reale. Più precisamente, la parte reale della matematica, ossia quella che Galvan definisce "aritmetica reale" (AR), è costituita «dall'insieme delle Π_1-formule chiuse vere rispetto al modello standard di LPR»,[36] dove una Π_1-formula si ottiene mediante la generalizzazione universale (ovvero mediante la regola di introduzione del quantificatore universale) di tutte o solo alcune delle variabili individuali che occorrono in una formula ricorsiva primitiva (PR-formula) di LPR. Da ciò, risulta che AR coincide con ARP. Inoltre, le PR-formule chiuse sono «finitariamente significanti» (cioè hanno un «significato reale») – sostiene Galvan – poiché è possibile determinare il loro valore di verità mediante un computo effettivo riferito ai numeri naturali (che sono «oggetti concreti e finiti»), che «non coinvolge perciò il riferimento ad entità astratte o a totalità infinite».[37] Tuttavia, Galvan evidenzia che la

[35] «Tanto è vero che la teoria proposta a suo tempo da Skolem quale sistema formale entro il quale formalizzare l'aritmetica ricorsiva primitiva […] si rivelerà, insieme con ogni sua estensione assiomatizzabile, insufficiente a derivare la totalità delle proposizioni vere di ARP» (Cfr. *ivi*, p. 108).

[36] Cfr. *ivi*, p. 110.

[37] Cfr. *ivi*, p. 111.

definizione di proposizione reale non esaurisce quella di formula finitaria, perché il concetto di "finitariamente significante" possiede un'estensione più ampia di quello di "finitariamente derivabile". Infatti, il primo si riferisce al contenuto finitario di una formula, mentre il secondo si riferisce alla derivabilità di una formula mediante procedure finitarie.[38] Dunque, data l'equivalenza acquisita tra AR e ARP, e dato che, come spiega Galvan, «il metodo di individuazione delle formule appartenenti ad AR attraverso la richiesta della loro verità è un metodo altamente infinitario, di cui non si può far uso per generarne effettivamente l'insieme»,[39] risulta che l'aritmetica reale non è adatta alla definizione della matematica finitaria perseguita dal programma formalista. A questo scopo si rende necessaria l'assiomatizzazione di ARP e, come abbiamo già anticipato, il tentativo meglio riuscito (ma non soddisfacente) in questo senso consiste proprio nel sistema assiomatizzato PRA. Tale sistema, spiega Galvan, «intende catturare l'insieme delle proposizioni finitarie», ovvero le proposizioni finitariamente derivabili, «e si propone, al contempo di formalizzare il concetto di procedura finitaria di derivazione, nel senso che le procedure finitarie sono tutte e solo quelle formalizzabili entro tale sistema o suoi equivalenti».[40] PRA, inoltre, è particolarmente importante perché è dimostrabile che ogni teoria formale che ne costituisce un'estensione conservativa[41] può essere espressa mediante il linguaggio di PRA.

Occorre considerare a questo punto un ultimo aspetto di grande rilievo introdotto da Hilbert: la distinzione tra matematica formalizzata e metamatematica. Come rileva Mancosu, la formalizzazione della matematica costituisce soltanto la prima parte del programma, e consiste in quella che Hilbert definisce come "teoria della dimostrazione". La metamatematica consiste invece nell'insieme di «ragionamenti contenutistici […] o finitisti»,[42] ovvero dall'insieme di proposizioni che descrivono le relazioni tra i simboli che formano gli enunciati della matematica, e mediante cui è possibile, pertanto, produrre asserzioni dotate di significato intorno alla sintassi della matematica. La natura "contenutistica" o "finitista" dell'analisi metamatematica è da collegare: in primo luogo, a

[38] Cfr. *ivi*, p. 114.

[39] Cfr. *ivi*, p. 115.

[40] Cfr. *ivi*, p. 117. «PRA […] contiene solo quelle verità (verità queste espresse negli assiomi) che si possono direttamente appurare a proposito di oggetti concreti costruibili in modo finito, come sono i numeri naturali […], oppure che sono derivabili a partire dalle prime attraverso specifici passaggi formali totalmente esplicitabili e non problematici. Tali passaggi, sia quelli di natura puramente logica sia quelli di natura aritmetica, come le derivazioni per induzione, sono strutturati, cioè, in modo tale da riuscire a *mantenere* la verità *reale* delle premesse (non si esce dall'ambito delle Π_1-formule vere). Ciò si ottiene, da una parte, attraverso la riduzione (almeno in linea di principio) della componente logica alla sola logica dell'identità *senza quantificatori* e, dall'altra, attraverso il solo uso della regola d'*induzione ristretta* a proprietà ricorsive primitive» (Cfr. *ivi*, p. 115).

[41] Una teoria T' costituisce un'estensione conservativa di una teoria T quando il linguaggio L(T') comprende almeno l'alfabeto e le regole di formazione di L(T) e quando, data una formula α espressa in L(T), se α è derivabile dagli assiomi *A*(T), allora è derivabile anche dagli assiomi *A*(T').

[42] Cfr. P. Mancosu, *Il programma di Hilbert e i teoremi di incompletezza di Gödel*, cit., p. 513.

quella che Bernays definisce «rappresentazione intuitivo-primitiva che sta alla base dell'uso del segno; in secondo luogo, alla finitezza del dominio intorno a cui le proposizioni metamatematiche vertono. La distinzione tra teoria della dimostrazione e metamatematica, nell'ottica formalista, è ciò che dovrebbe consentire di stabilire una prova diretta della consistenza per la matematica formalizzata. Infatti, dal momento che le combinazioni simboliche interne al linguaggio formale matematico sono finite, risulta perciò possibile un'analisi completa delle relazioni sintattiche tra le formule di quest'ultimo, tale da poter stabilire in modo certo, data una qualsiasi tra quelle, se essa costituisca o meno un teorema della teoria. In altre parole, «un requisito essenziale del programma [hilbertiano]» – scrivono Nagel e Newman – «è che le dimostrazioni di coerenza si [basino] soltanto su quei procedimenti che non fanno appello a un numero infinito di proprietà strutturali delle formule»;[43] tali procedimenti, che formano l'analisi metamatematica, vengono perciò detti finitari. La necessità di fornire una fondazione formalista della matematica utilizzando un approccio finitario, secondo Mancosu, si deve al tentativo da parte di Hilbert di «mostrare che, aggiungendo ai formalismi [cioè alle teorie formalizzate] gli *assiomi transfiniti*, l'applicazione generalizzata delle leggi logiche non porterà mai a una contraddizione».[44] La dimostrazione di consistenza della matematica, perciò, consiste nel mostrare che tutte le proposizioni ideali della matematica sono riducibili a proposizioni reali. Hilbert, in un passo del saggio del 1923, *Ricerche sui fondamenti della matematica*, scrive in proposito:

> Si vede dunque che, rispetto al fine di una fondazione rigorosa della matematica, gli usuali procedimenti inferenziali dell'analisi non possono venir accettati come logicamente pacifici. Il nostro compito è piuttosto quello di conoscere perché e fino a che punto l'uso dei procedimenti inferenziali transfiniti […] porta pure a risultati corretti. Muovendosi sul terreno del finitario, si deve dunque garantire il libero uso e il pieno dominio del transfinito![45]

[43] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 41.

[44] Cfr. P. Mancosu, *Il programma di Hilbert e i teoremi di incompletezza di Gödel*, cit., p. 510. Il corsivo è dell'autore. In generale, gli assiomi transfiniti contengono elementi che si riferiscono ad un dominio oggettuale di interpretazione infinito, e «incorporano le leggi classiche per i quantificatori universale ed esistenziale» (Cfr. *Ivi*, p. 512). È importante notare, inoltre, che Hilbert usa il termine "transfinito" sinonimicamente a quello di "infinitario", e perciò in un senso più ampio rispetto alla nozione di transfinito di Cantor. Prendiamo in considerazione un esempio di assioma transfinito introdotto da Hilbert, che viene riportato da Mancosu. «Hilbert postula l'esistenza di un operatore τ tale che, per ogni proprietà A(x), è lecito formare un termine $\tau A(x)$, ove $\tau A(x)$ è tale che, se A(x) vale di esso, allora A(x) è vero per tutti gli x. Hilbert chiarisce il suo significato per mezzo del seguente esempio. Sia il significato di A(x) *essere corruttibile*. Allora $\tau A(x)$ è un uomo di una tale integrità morale che, se egli è corruttibile, allora tutti gli uomini sono corruttibili. Formalmente abbiamo il seguente assioma transfinito $A(\tau(A(x)) \rightarrow A(x)$» (Cfr. *ibidem*).

[45] La citazione è presa dalla traduzione di Mancosu del testo originale di Hilbert (Cfr. *ivi*, p. 509).

Da ultimo, resta da evidenziare, con Mancosu, che nel corso di pochi anni lo statuto epistemologico del formalismo hilbertiano (come abbiamo detto, mai pienamente precisato) subisce una serie di modificazioni.[46] Dapprima, Bernays sembra caratterizzare l'intuizione primitiva finitista (o, equivalentemente, finitaria) in un senso empirico. In un testo del 1922, egli scrive:

> Faremo nondimeno nostra la tendenza delle scienze esatte che tende a eliminare per quanto possibile le facoltà cognitive più raffinate e a basarsi solo sui modi più primitivi di cognizione. […] Esamineremo se sia possibile fondare quelle assunzioni trascendenti in modo tale che entrino in gioco solo delle cognizioni primitivamente intuitive. Sulla base di questa restrizione […] non possiamo, d'altro canto, richiedere che questa fondazione ci permetta di riconoscere come verità (nel senso filosofico) le assunzioni che devono essere fondate. Piuttosto saremo soddisfatti se riusciremo a dimostrare che l'aritmetica costruita a partire da quelle assunzioni è un sistema di pensiero possibile (cioè consistente).[47]

Mentre successivamente, sia Bernays che lo stesso Hilbert paiono accostarsi ad una forma di apriorismo kantiano. In un ciclo di lezioni del 1923, infatti, Hilbert afferma:

> Io credo che la conoscenza matematica si basi su una sorta di intuizione e che persino per la costruzione della teoria dei numeri è necessaria una certa concezione intuitiva e a priori. In questo modo le idee più generali di Kant mantengono la loro importanza: cioè il problema filosofico di precisare la concezione intuitiva a priori e con ciò le condizioni di possibilità di ogni possibile conoscenza concettuale […]. Questo è fondamentalmente quello che è successo nella mia teoria sui fondamenti della matematica.[48]

Da quanto detto, risulta che per il finitismo la condizione di possibilità della costruzione del sapere matematico, e in generale, di quello scientifico, risiede nell'intuizione a priori dei simboli che sono alla base del programma di formalizzazione. Tuttavia, come vedremo nel prossimo capitolo, i risultati del successivo lavoro di Gödel, decretando l'impossibilità di una prova assoluta della coerenza, non soltanto dell'aritmetica primitiva ricorsiva formalizzata, ma di una qualsiasi teoria formalizzata che ne costituisca un'estensione conservativa, determinarono con ciò un forte ridimensionamento del programma formalista, ma non il suo totale fallimento. Come rileva Mancosu, infatti, Gödel (al contrario di Von Neumann) fu molto cauto nel derivare le conseguenze implicate dai suoi teoremi per il progetto hilbertiano. Infatti, proprio a causa dell'indeterminatezza con cui viene assunta la nozione di finitarietà di una procedura dimostrativa da parte di Hilbert e Bernays, i teoremi

[46] Cfr. *ivi*, pp. 516-519.
[47] La citazione è presa dalla traduzione di Mancosu del testo originale di Bernays (Cfr. *ivi*, p. 517).
[48] La citazione è presa dalla traduzione di Mancosu del testo originale di Hilbert (Cfr. *ivi*, p. 518).

di incompletezza attestano soltanto che esiste una formula finitariamente non derivabile all'interno di PRA, ma non anche che non esiste alcuna procedura finitariamente significante di tipo metamatematico che sia in grado di attestare la completezza sintattica di PRA. La questione si gioca, come abbiamo già evidenziato con Galvan, sulla distinzione tra i due sensi di "finitario". Alla luce dei risultati dei teoremi di limitazione, dunque, Bernays concorda con Gödel nell'affermare che «ci sono enunciati finitisticamente giustificabili ma non dimostrabili nella teoria dei numeri. […] In altre parole, non tutte le considerazioni finitiste possono venire formalizzate nel sistema».[49]

[49] Cfr. *ivi*, p. 530.

Capitolo 2
I teoremi di incompletezza di Gödel

Come abbiamo visto, l'idea che muoveva la ricerca di molti matematici e logici agli inizi del Novecento, tra cui Frege, Russell e Hilbert, era quella per cui si riteneva possibile fornire una solida fondazione di ogni branca del sapere matematico mediante la sua formalizzazione in un sistema e una dimostrazione finitaria della sua coerenza. Tuttavia, come abbiamo anticipato, i teoremi di incompletezza di Gödel mostrano che la consistenza di un sistema formale non è finitariamente derivabile, ossia non è garantibile esclusivamente sulla base di procedimenti inferenziali finitari rappresentabili nel linguaggio di PRA.

Gödel formula il teorema VI e il suo corollario, ovvero quelli che vengono comunemente chiamati i due "teoremi di incompletezza", nell'articolo *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I*,[50] pubblicato nel 1931, e i cui risultati furono esposti per la prima volta in pubblico ad una conferenza tenuta a Königsberg nel 1930. Il termine "incompletezza" è da riferirsi alla questione posta dal cosiddetto "secondo problema" di Hilbert: dimostrare la coerenza di un sistema formale mediante soli metodi finitari; ad esso si collega un'assunzione-chiave del progetto formalista, che Hilbert illustra come segue: «la dimostrazione della non-contraddittorietà dell'ipotesi secondo cui ogni problema matematico è risolvibile rientra perfettamente nell'ambito della nostra teoria».[51] Nel suo lavoro, Gödel prende in considerazione principalmente i due sistemi formali per la teoria degli insiemi più avanzati all'epoca, ovvero quello dei *Principia Mathematica*, elaborato da Russell e Whitehead (a cui mi riferirò usando la sigla PM), e quello sviluppato da Zermelo e Fraenkel (ZF): due teorie che costituivano allora la formalizzazione più generale della matematica. Nella seconda parte del testo, in cui Gödel procede all'esposizione formale dei suoi risultati, viene precisato che la teoria formale generale (che egli denota con *P*) utilizzata nella dimostrazione è un'estensione conservativa di PM, ottenuta attraverso l'aggiunta degli assiomi dell'aritmetica di Peano (PA). Nell'incipit del saggio, Gödel scrive:

> Si potrebbe quindi supporre che questi assiomi e queste regole di inferenza [di PM e di ZF] siano sufficienti per decidere *tutti* i problemi matematici che si possano comunque esprimere formalmente in

[50] K. Gödel, *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I (1931)*, in *Opere*. *Volume I: 1929-1936*, trad. it. di E Ballo, S. Bozzi, G. Lolli, C. Mangione, Bollati Boringhieri, Torino 1999, pp. 113-138; tit. orig. *On formally undecidable propositions of* Principia mathematica *and related systems I (1931)*, in *Collected Works*. *Volume I: Publications 1929-1936*, Oxford University Press, New York 1986.

[51] Cfr. D. Hilbert, *Sull'infinito*, trad. it. di A. Frigo, Castelvecchi, 2013; tit. orig. *Über das Unendliche*, in «Mathematische Annalen», LVII, 1925, vol. 95, pp. 161-190.

essi. Mostreremo nel seguito come ciò non valga, che anzi nei due sistemi indicati esistono problemi relativamente semplici della teoria degli usuali numeri interi che non possono venir decisi sulla base degli assiomi. Questa situazione non dipende in alcun modo dalla particolare natura dei sistemi costruiti, ma si verifica per un'ampia classe di sistemi formali, fra i quali, in particolare, vi sono tutti quelli che si ottengono dai due sopra indicati aggiungendo loro un numero finito di assiomi.[52]

Prima di procedere con la trattazione, precisiamo che nell'illustrare i teoremi di incompletezza faremo affidamento, in particolare, alle ricostruzioni di Berto, Galvan, Nagel e Newman. Inoltre, ai fini della comprensione, è alquanto utile fornire, in prima battuta, un'esposizione, per così dire, "informale" dei teoremi, includente cioè la nozione semantica di verità. La coerenza e la completezza semantica di una teoria formale, naturalmente, in tanto in quanto alludono alla verità o alla falsità relativamente ad una data formula, implicano, perciò stesso, l'esistenza di una struttura semantica. Inoltre, è opportuno ricordare la distinzione tra dimostrabilità e verità di un enunciato (o formula), che è di massima importanza per le conclusioni che si possono trarre dalla prova di Gödel. Egli era ben consapevole dell'eventuale critica che i formalisti avrebbero potuto muovere nei confronti di una dimostrazione di coerenza semantica, e fu perciò attento nel restringere la sua argomentazione alla coerenza sintattica. In particolare, egli adoperò come requisito una condizione ancora "più debole" della semplice coerenza, che analizzeremo in seguito, ovvero "ω-coerenza". Per ora limitiamoci ad illustrare in maniera informale i due teoremi di incompletezza sulla falsariga della ricostruzione fornita da Gödel nel suo articolo del 1931, e facendo uso della nozione di coerenza "semantica".

Previa considerazione del fatto, già menzionato, che una qualunque teoria assiomatizzata può venire espressa nel linguaggio di PRA, si assuma una teoria formale T corretta. Allora è possibile formulare in T un enunciato G (detto "enunciato gödeliano" di T) di tipo "autoreferenziale", ovvero tale che, dal punto di vista dell'analisi metamatematica intorno alla sintassi di T, esso asserisca che G non è derivabile in T. Ora, se assumiamo invece che G sia derivabile in T, allora G è falso, e T risulta semanticamente incoerente, contrariamente a quanto ipotizzato. Ma nemmeno la negazione di G, ¬G, è derivabile in T. Infatti, se così fosse, allora ¬G sarebbe vero e G falso, ma abbiamo visto che la falsità di G implica la sua derivabilità; sarebbero quindi derivabili in T entrambe le formule contraddittorie, contrariamente all'assunto di partenza. Da ciò risulta il primo teorema di incompletezza, che formuliamo in modo generale come segue: se T è una teoria formale corretta, ed un'estensione conservativa di PRA, allora è formulabile nel linguaggio L(T) un enunciato G indecidibile in T.

[52] Cfr. K. Gödel, *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I (1931)*, cit., pp. 113-114.

È importante evidenziare, qui, che dall'indecidibilità di G in T segue immediatamente che: (a) G è vero e T è semanticamente incompleto; (b) T è corretto. Il punto (b), come è evidente, non è altro che una banale *petitio principii*, in quanto esprime una conclusione che coincide con la premessa. Ciononostante, esso mette in luce l'aspetto principale della strategia gödeliana: l'autoreferenzialità dell'enunciato G. La proprietà di autoreferenzialità, la quale, come è noto, sta all'origine di numerosi paradossi linguistici (vedi, ad esempio, il paradosso del mentitore), caratterizza anche il concetto di insieme intuitivamente inteso, da cui derivano i celebri paradossi della teoria degli insiemi rilevati da Russell. Dal canto suo, già Cantor, molto tempo prima, si era reso conto della paradossalità della nozione di insieme "assoluto".[53] È naturale, quindi, aspettarsi che la medesima struttura antinomica di autoreferenzialità, valida per un qualsiasi sistema inteso come una collezione contenente in atto "ogni" possibile configurazione di un certo tipo (cioè sottostante ad un determinato criterio di inclusione), si ripresenti anche all'interno del contesto della logica formale, e in particolare, in riferimento all'idea formalista della definizione di una teoria formale comprensiva di "tutte" le verità matematiche.

Abbiamo già rilevato che la coerenza semantica costituisce una condizione più forte della coerenza sintattica, nel senso che la prima è condizione sufficiente per la seconda. Ora, se denotiamo con C l'enunciato che asserisce la coerenza sintattica di T, allora per (b) vale il seguente condizionale: 'se G è indimostrabile in T, allora T è coerente ($G \Rightarrow C$)'. Il suo inverso sarà perciò: 'se T è coerente, allora G è indimostrabile in T ($C \Rightarrow G$)'. Da ciò risulta che G equivale a C ($C \Leftrightarrow G$). Dunque, dalla conclusione del primo teorema che stabilisce l'indimostrabilità di G in T, deriva il corollario, noto come secondo teorema di incompletezza, che enunciamo informalmente come segue: se T è una teoria formale corretta, ed è un'estensione conservativa di PRA, allora l'enunciato C che asserisce la consistenza di T non è dimostrabile all'interno di T. La dimostrazione è immediata: se C fosse dimostrabile in T, allora, per la regola di introduzione del *modus ponens*, seguirebbe G. Ma questo è escluso dal primo teorema che afferma che, se T è corretta, allora G non è dimostrabile al suo interno.

Possiamo già vedere in questo contesto quanto abbiamo anticipato verso la fine del capitolo precedente, ovvero il fatto che i teoremi di limitazione sanciscono la necessità di un ridimensionamento del progetto formalista. Qui, in proposito, ci limitiamo ad osservare che il secondo teorema dimostra che la consistenza sintattica (e, a maggior ragione, quella semantica) di una teoria formale non è finitariamente dimostrabile all'interno di un'estensione conservativa di PRA.

A questo punto, è necessario procedere ad analizzare nel dettaglio i due teoremi per due motivi: in primo luogo, al fine di mostrare quali siano le conclusioni sul piano formale che da essi è

[53] In proposito si veda G. Basti, *Le radici forti del pensiero debole*, Il Poligrafo, Padova 1996, pp. 181-202.

possibile trarre; in secondo luogo, per ridimensionare la portata delle loro implicazioni, spesso sopravvalutate, nel contesto della riflessione filosofica intorno al problema mente-corpo, che è al centro del nostro lavoro. Una simile analisi ci consente, peraltro, di giudicare la legittimità dell'uso dei teoremi di limitazione, da parte di pensatori come Lucas e Penrose, per avallare la posizione anti-meccanicista riguardante la natura della mente umana. Nell'illustrare la «prova di Gödel», tuttavia, riteniamo opportuno omettere alcuni passaggi di natura alquanto tecnica che appesantirebbero inutilmente il discorso, senza con ciò sacrificare il rigore richiesto in questa sede.[54]

Ritorniamo all'enunciato G: quest'ultimo, lo ricordiamo, è un enunciato del linguaggio della teoria formale T. In che modo, dunque, per mezzo di un linguaggio formale quale L(PRA), è formulabile un enunciato tale da contenere un'asserzione dotata di significato intorno alla sintassi di T, e quindi tale da riferirsi a T? E, in particolare, come è possibile formulare un enunciato tale che si riferisca proprio a se stesso? In primo luogo, si tenga a mente la già menzionata distinzione, proposta da Hilbert, tra teoria della dimostrazione e metamatematica. Il carattere di autoreferenzialità applicato al linguaggio della teoria della dimostrazione, in questo caso, implica la possibilità da parte di quest'ultimo di produrre un'asserzione semantica intorno agli elementi e alle relazioni sintattiche di cui è costituito. In secondo luogo, si consideri il risultato precedentemente acquisito circa la possibilità di esprimere il contenuto di una qualsiasi teoria formale assiomatizzata (e quindi anche del sistema formale *P*, a cui Gödel fa riferimento) nel linguaggio di PRA. Nello specifico, i termini individuali di L(PRA), sono detti "numerali", e denotano gli oggetti del dominio di interpretazione privilegiato $\mathbb{N}$ di PRA. Ora, l'autoreferenzialità del linguaggio di PRA si ottiene mediante due passaggi fondamentali: (a) l'"aritmetizzazione della sintassi" di una teoria formale, mediante una procedura ideata da Gödel, comunemente detta "gödelizzazione"; (b) la rappresentazione in PRA delle funzioni-relazioni aritmetiche. Entrambi chiamano in causa la funzione di "rappresentazione", tale per cui «una struttura astratta di relazioni valide in un dominio di "oggetti" si può dimostrare valida tra "oggetti" […] di un altro dominio».[55]

La gödelizzazione è una procedura meccanica di codificazione, e consiste in una funzione biunivoca che associa a ciascun elemento del linguaggio di PRA uno e un solo numero naturale, detto "numero di Gödel", "gödeliano", o "codice" di quel dato elemento. La biunivocità di tale funzione consente, pertanto, la codifica di L(PRA) in numeri di Gödel, e viceversa, la "decodifica" di questi ultimi (e delle loro relazioni aritmetiche) in L(PRA). È da notare che il criterio adoperabile per la

[54] Per una ricostruzione dettagliata dei teoremi di incompletezza gödeliani, rimandiamo il lettore ai seguenti testi: E. Nagel, J.R. Newman, *La prova di Gödel*, cit.; S. Galvan, *Introduzione ai teoremi di incompletezza*, cit.; F. Berto, *Tutti pazzi per Gödel!*, cit..

[55] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 75.

codifica è arbitrario o convenzionale.[56] Come evidenziano Nagel e Newman, attraverso questa procedura, «una proposizione metamatematica intorno alle espressioni [di PRA] e alle loro reciproche relazioni può essere costruita come una proposizione intorno ai numeri (di Gödel) corrispondenti e alle loro reciproche relazioni aritmetiche»[57]. Inoltre, la gödelizzazione è dimostrabilmente estendibile a qualsiasi teoria assiomatizzabile, e definisce quella che abbiamo chiamato "aritmetizzazione della sintassi".

A questo punto, per introdurre la nozione di "rappresentabilità" di una relazione, proprietà o funzione aritmetica nel linguaggio di PRA, è preliminarmente necessario, innanzitutto, ricordare che PRA consiste nel (tentativo di) assiomatizzazione di ARP (l'aritmetica ricorsiva primitiva), e che ARP contiene tutte e sole le PR-formule (formule ricorsive primitive) chiuse di LPR (linguaggio primitivo ricorsivo) che sono vere rispetto al suo modello standard. Possiamo quindi fornire le seguenti definizioni, riprendendole da Galvan:[58]

Def. 1. Sia dato […] *dice* [sic!] *il termine* $t(x_1 \dots x_n) \in$ LPR. $t(k_1 \dots k_n) = k_{n+1}$ *significhi, inoltre, che* $t(k_1 \dots k_n) = k_{n+1}$ *è vero nel modello standard di* LPR. *Allora si dice che* $t(x_1 \dots x_n)$ *è rappresentabile in PRA attraverso se stesso se e solo se* $t(k_1 \dots k_n) = k_{n+1} \Rightarrow \text{PRA} \vdash t(\overline{k}_1 \dots \overline{k}_n) = \overline{k}_{n+1}$.

Def. 2. Sia data la relazione $R(x_1 \dots x_n) \in$ LPR. $R(k_1 \dots k_n)$ *(rispettivamente* $\neg R(k_1 \dots k_n)$*) significhi, inoltre, che* $R(k_1 \dots k_n)$ *(rispettivamente* $\neg R(k_1 \dots k_n)$*) è vero nel modello standard di* LPR. *Allora si dice che* $R(x_1 \dots x_n)$ *è rappresentabile in PRA se e solo se esiste una formula* $\alpha(x_1 \dots x_n)$ *tale che* $R(k_1 \dots k_n) \Rightarrow \text{PRA} \vdash \alpha(\overline{k}_1 \dots \overline{k}_n)$ e $\neg R(k_1 \dots k_n) \Rightarrow \text{PRA} \vdash \neg\alpha(\overline{k}_1 \dots \overline{k}_n)$.

Da tali definizioni derivano i seguenti teoremi, che ci limiteremo ad enunciare:

Teorema di rappresentabilità dei termini ricorsivi primitivi in PRA*: Ogni termine ricorsivo primitivo è rappresentabile in* PRA *attraverso se stesso.*

[…] *Teorema di caratterizzazione delle* PR-*formule: Sia* $\alpha(x_1 \dots x_n)$ *una* PR-*formula. Allora esiste un termine ricorsivo primitivo* $t(x_1 \dots x_n)$ *che caratterizza in PRA* α, *vale a dire:*

$$\text{PRA} \vdash \alpha(x_1 \dots x_n) \leftrightarrow t(x_1 \dots x_n) = 0$$

[56] Per alcuni esempi di gödelizzazione rimandiamo ai testi citati nella nota 51. In questa sede, per la costruzione dell'enunciato G faremo affidamento sul criterio di codificazione usato da Nagel e Newman (Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., pp. 78-85).

[57] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 86.

[58] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 119.

[…] *Metateorema generale di rappresentabilità (in breve* MGR*): Sia data la* PR-*formula* $\alpha(x_1 \ldots x_n)$. *Come sopra, poi,* $\alpha(k_1 \ldots k_n)$ *(rispettivamente* $\neg\alpha(k_1 \ldots k_n)$*) significhi che* $\alpha(k_1 \ldots k_n)$ *(rispettivamente* $\neg\alpha(k_1 \ldots k_n)$*) è vero nel modello standard di* LPR. *Allora* $\alpha(k_1 \ldots k_n) \Rightarrow$ PRA - $\vdash \alpha(\overline{k}_1 \ldots \overline{k}_n)$ *e* $\neg\alpha(k_1 \ldots k_n) \Rightarrow$ PRA - $\vdash \neg\alpha(\overline{k}_1 \ldots \overline{k}_n)$.

[…] *Teorema di rappresentabilità in* PRA *delle relazioni ricorsive primitive: Ogni relazione ricorsiva è rappresentabile in* PRA.[59]

In particolare, MGR stabilisce che tutte le formule di ARP sono rappresentabili nel linguaggio di PRA, ed è equivalente, perciò, al teorema V dell'articolo di Gödel del 1931.

Da quanto detto, risulta che, attraverso la procedura di gödelizzazione di L(PRA) e l'aritmetizzazione della sintassi di PRA, quest'ultima diviene capace di formulare delle espressioni che si riferiscono, attraverso i numerali, a quei numeri di Gödel che codificano espressioni appartenenti alla teoria medesima. In particolare, è da notare che l'autoreferenzialità dell'enunciato gödeliano è data dalla capacità dell'enunciato di riferirsi "indirettamente" al suo stesso numero gödeliano, come dice Berto, «mediante una sorta di descrizione definita […] che, per via della stessa costruzione formale in gioco, venga soddisfatta precisamente dall'enunciato gödeliano stesso».[60] In proposito, in una nota al testo, Gödel sottolinea:

A dispetto delle apparenze, una tale proposizione non contiene circoli viziosi, poiché inizialmente essa afferma [soltanto] che una certa formula ben definita (cioè quella ottenuta […] mediante una certa sostituzione) non è dimostrabile. Solo successivamente (e in un certo senso per caso) vien fuori che questa formula è precisamente quella mediante la quale era stata espressa la proposizione stessa.[61]

Rimangono, ora, soltanto da definire alcune regole che servono alla costruzione dell'enunciato G. Nel medesimo articolo del 1931, Gödel definisce una serie di funzioni-relazioni ricorsive primitive dell'aritmetica e dimostra la loro rappresentabilità in *P* (nel nostro caso, PRA). Una di queste consiste in una relazione aritmetica ricorsiva primitiva a due posti, che denotiamo con "$\mathrm{Dim}_{\mathrm{PRA}}$", la quale[62] traduce una proposizione metamatematica intorno alla relazione sintattica tra una dimostrazione e un

[59] Cfr. *ivi*, pp. 119-123. Per la loro dimostrazione cfr. *ibidem*. Si noti che '$\overline{k}$' indica un certo numerale

[60] Cfr. F. Berto, *Tutti pazzi per Gödel!*, cit., p. 110.

[61] Cfr. K. Gödel, *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I (1931)*, cit., p. 116.

[62] Come sottolinea Berto, la ricorsività di $\mathrm{Dim}_{\mathrm{PRA}}$ deriva dalla ricorsività dell'impianto assiomatico di PRA.

teorema di PRA. In particolare, tale relazione definisce un insieme D, i cui elementi sono coppie ordinate di numeri naturali (di Gödel), tale che una coppia ordinata (<*m*, *n*>) appartiene all'insieme D se, e solo se, *m* è il gödeliano di una dimostrazione in PRA e *n* il gödeliano del teorema risultante da tale dimostrazione. Per MGR (e quindi per il Teorema V di Gödel), inoltre, $\mathrm{Dim}_{\mathrm{PRA}}$ è rappresentabile nel linguaggio di PRA. Seguendo l'indicazione di Berto, per distinguere $\mathrm{Dim}_{\mathrm{PRA}}$ dalla sua rappresentazione formale, denotiamo quest'ultima in grassetto come "$\mathbf{Dim}_{\mathbf{PRA}}$(x, y)". È importante sottolineare che $\mathbf{Dim}_{\mathbf{PRA}}$(x, y) esprime una relazione binaria di PRA tra le variabili numeriche *x* e *y*, e denota perciò una formula aperta. Data una coppia ordinata di numeri naturali che soddisfano $\mathrm{Dim}_{\mathrm{PRA}}$, <*m*, *n*>, ad esempio, sostituendo ad x e y i numerali ad essi corrispondenti, $\overline{m}$ ed $\overline{n}$, si ottiene la formula chiusa $\mathbf{Dim}_{\mathbf{PRA}}(\overline{m}, \overline{n})$.

Successivamente, consideriamo il predicato aritmetico, "$\mathrm{Teor}_{\mathrm{PRA}}$", tale che:

$$\mathrm{Teor}_{\mathrm{PRA}}(n) =_{\mathrm{df}} \exists m \mathrm{Dim}_{\mathrm{PRA}}(m, n).$$

$\mathrm{Teor}_{\mathrm{PRA}}$ afferma, cioè, che *n* è il gödeliano di un teorema di PRA, se e solo se esiste un numero *m* tale che *m* è il gödeliano di una dimostrazione del teorema il cui gödeliano è *n*. Ora, dal momento che l'insieme dei teoremi di una teoria formale non è decidibile, bensì soltanto enumerabile (o semidecidibile), $\mathrm{Teor}_{\mathrm{PRA}}$ è una proprietà soltanto ricorsivamente enumerabile, e, per questo motivo, soltanto semi-rappresentabile in PRA[63] attraverso la formula a una variabile "$\mathbf{Teor}_{\mathbf{PRA}}$(y)", tale che:

$$\mathbf{Teor}_{\mathbf{PRA}}(\mathrm{y}) =_{\mathrm{df}} \exists \mathrm{x} \mathbf{Dim}_{\mathbf{PRA}}(\mathrm{x}, \mathrm{y}).$$

Anche in questo caso, vale la precisazione per cui $\mathbf{Teor}_{\mathbf{PRA}}$(y) è una formula aperta. La sua chiusura si ottiene sostituendo alla variabile numerica *y* un termine numerale, che rappresenta in PRA un dato numero naturale (di Gödel).

Da ultimo, specifichiamo la funzione essenziale alla costruzione dell'enunciato "autoreferenziale" G, ovvero la funzione di sostituzione "Sost". In generale, in un linguaggio formale, la funzione di sostituzione consente di sostituire, all'interno di una data formula, alle occorrenze di una variabile libera, un certo termine individuale. In particolare, Sost denota una funzione aritmetica, operante sui numeri di Gödel, che traduce un'operazione sintattica sulle formule di PRA. Ora, se assumiamo che *m* sia il gödeliano della formula α di PRA, *n* il gödeliano della variabile x, e *o* un certo numero naturale, allora il valore dato dalla funzione Sost(*m*, *n*, *o*) sarà un determinato numero

[63] Cfr. F. Berto, *Tutti pazzi per Gödel!*, cit., p. 106.

di Gödel. Questo numero corrisponde (via gödelizzazione), in PRA, al valore dato dalla funzione di sostituzione operata sulla formula di partenza α. Tale valore equivale alla formula $\alpha(x/\overline{o})$, ottenuta da α per sostituzione delle occorrenze libere della variabile x con il termine individuale $\overline{o}$ (il numerale del numero naturale *o*). Anche la funzione aritmetica Sost è rappresentata in PRA, mediante numerali, dalla funzione **Sost**. Come al solito, ricordiamo che **Sost** individua una formula aperta quando opera su delle variabili numeriche libere, mentre individua una formula chiusa quando alle variabili sostituiamo i numerali che corrispondono agli argomenti della funzione aritmetica Sost.

Costruiamo, dunque, la seguente formula aperta di PRA a una variabile libera:

(G(y)) $$\neg\exists x\mathbf{Dim}_{\mathbf{PRA}}\left(x, \mathbf{Sost}(y, \overline{13}, y)\right).$$

Concentriamo la nostra attenzione sulla formula **Sost**(y, $\overline{13}$, y). Innanzitutto, **Sost** rappresenta in PRA la funzione aritmetica Sost, la quale traduce in termini aritmetici la proposizione metamatematica intorno alla funzione sintattica di sostituzione sugli elementi di PRA. In particolare, Sost opera su tre argomenti, rispettivamente: il gödeliano della formula di PRA all'interno della quale viene operata la sostituzione; il gödeliano (il numero naturale 13, il quale, nella notazione di Nagel e Newman da noi usata, codifica proprio la variabile y) che codifica la variabile libera *sostituenda* che occorre all'interno della formula di partenza; il gödeliano che codifica il termine con cui viene sostituita la variabile. In particolare, **Sost**(y, $\overline{13}$, y) opera, rispettivamente, sui seguenti argomenti: la variabile numerica *y* che rappresenta in PRA il numero di Gödel della formula di partenza all'interno della quale viene operata la sostituzione; il numerale $\overline{13}$ che rappresenta il numero di Gödel che codifica la variabile libera *sostituenda*; la variabile y identica al primo argomento. Analizziamo ora il valore dato dalla funzione **Sost**(y, $\overline{13}$, y). Esso equivale a una certa formula aperta che rappresenta in PRA il codice (o il numero di Gödel) di quella formula che si ottiene dalla formula di partenza, il cui gödeliano è rappresentato in PRA dalla variabile numerica y, attraverso l'operazione di sostituzione delle occorrenze libere in essa della variabile il cui gödeliano è il numero naturale 13 (ovvero la variabile y) con il numerale che corrisponde al gödeliano della formula di partenza. Questo risultato (un po' contorto) ci consente di riformulare, con linguaggio metamatematico, ciò che viene affermato da G(y) nel modo seguente: non esiste alcuna variabile numerica *x*, tale che *x* rappresenti in PRA il gödeliano di una dimostrazione di una formula di PRA il cui gödeliano è rappresentato in PRA dal valore della funzione **Sost**(y, $\overline{13}$, y).

La formula G(y), a sua volta, è codificabile mediante un certo numero di Gödel, diciamo *n*. Sostituendo in G(y) alla variabile y il numerale di *n*, ovvero $\overline{n}$, otteniamo la seguente formula chiusa:

(G) $$\neg\exists x\mathbf{Dim_{PRA}}\left(x, \mathbf{Sost}(\overline{n}, \overline{13}, \overline{n})\right).$$

In questo caso, il valore di **Sost**($\overline{n}$, $\overline{13}$, $\overline{n}$) è una formula chiusa che, mediante numerale, rappresenta in PRA il gödeliano di una formula ottenuta dalla formula di partenza il cui gödeliano è *n*, tramite operazione di sostituzione delle occorrenze libere in essa della variabile il cui gödeliano è 13, con il numerale del suo stesso gödeliano (ovvero $\overline{n}$). Analogamente a prima, l'enunciato G, dal punto di vista metamatematico, afferma che non esiste alcuna variabile x tale che x rappresenti in PRA il gödeliano della dimostrazione di una formula il cui gödeliano è rappresentato in PRA dalla formula equivalente al valore della funzione **Sost**($\overline{n}$, $\overline{13}$, $\overline{n}$). Ma la formula in questione è proprio G. Infatti, G è proprio quella formula che si ottiene dalla formula di partenza G(y) (il cui gödeliano è *n*) mediante sostituzione delle occorrenze libere in essa della variabile y (il cui gödeliano è 13) con il numerale $\overline{n}$ della formula stessa (il cui gödeliano è *n*). Possiamo dunque sintetizzare questo risultato dicendo che G è un enunciato che rappresenta mediante numerali una relazione aritmetica tra numeri naturali (di Gödel), la quale, a sua volta, codifica la proprietà sintattica dell'enunciato G di non essere un teorema di PRA, ovvero di non essere derivabile all'interno di PRA. Vediamo così un fatto che avevamo già anticipato, ovvero che l'autoreferenzialità di G si dà in maniera soltanto "indiretta".

Possiamo ora enunciare il Teorema VI dell'articolo di Gödel, ovvero il primo teorema di incompletezza, il quale è suddiviso in due punti:

(1) Se PRA è coerente, allora G non è derivabile in PRA: $\nvdash_{PRA} G$;

(2) Se PRA è ω-coerente, allora ¬G non è derivabile in PRA: $\nvdash_{PRA} \neg G$

Per la dimostrazione, riteniamo opportuno, per una questione di comodità e di chiarezza, seguire da vicino la ricostruzione di Berto. Il punto (1) si può dimostrare come segue: se G fosse derivabile in PRA, allora la sua dimostrazione avrebbe un certo codice (un gödeliano) *k*. Perciò, *k* e il codice di G, diciamo *g*, apparterrebbero all'insieme delle coppie ordinate per cui vale la relazione aritmetica Dim_{PRA}. Rispettivamente ai numerali di *k* e *g*, varrebbe perciò anche $\mathbf{Dim_{PRA}}(\overline{k}, \overline{g})$, per cui vale che la formula il cui gödeliano è *g* è derivabile in PRA. Per generalizzazione universale otteniamo quindi: $\vdash_{PRA} \exists x\mathbf{Dim_{PRA}}(x, \overline{g})$. Ma, dal momento che $\overline{g}$ equivale al numerale che rappresenta il gödeliano di G, e che quest'ultimo asserisce di non essere derivabile in PRA, allora la formula $\exists x\mathbf{Dim_{PRA}}(x, \overline{g})$ equivale a ¬G. La derivabilità di G implicherebbe anche quella della sua contraddittoria, e da ciò risulterebbe l'incoerenza di PRA, contro l'assunzione iniziale.

Per quanto riguarda il punto (2), si assuma che la negazione di G sia derivabile in PRA, ovvero che $\vdash_{PRA} \exists x\mathbf{Dim_{PRA}}(x, \overline{g})$. Per (1), se PRA è coerente, allora G non è derivabile al suo interno.

Dunque, per ogni numero naturale *n*, non esiste un *n* che sia il codice di una dimostrazione di G in PRA. In termini metamatematici: $\forall n \vdash_{PRA} \neg \mathbf{Dim_{PRA}}(\overline{n}, \overline{g})$ (ovvero, per esteso: $\vdash_{PRA} \neg \mathbf{Dim_{PRA}}(0, \overline{g})$, $\vdash_{PRA} \neg \mathbf{Dim_{PRA}}(\overline{1}, \overline{g})$, $\vdash_{PRA} \neg \mathbf{Dim_{PRA}}(\overline{2}, \overline{g})$, …). Tale risultato, insieme all'assunzione della derivabilità di $\exists x \mathbf{Dim_{PRA}}(x, \overline{g})$, equivale alla ω-incoerenza di PRA, e contraddice, perciò, l'ipotesi di partenza. Pertanto, se si assume l'ω-coerenza di PRA,[64] da (1) e (2) risulta: (a) che l'enunciato G non è decidibile in PRA; (b) che PRA è sintatticamente incompleta.

Ora, è possibile dimostrare che il condizionale espresso in (1) è rappresentabile all'interno di PRA mediante una formula derivabile dai suoi assiomi. In altre parole, (1) è derivabile in PRA. Inoltre, per la legge dello pseudo-Scoto,[65] vale che l'inconsistenza di una teoria implica la derivabilità di qualsiasi enunciato in essa. La consistenza di quest'ultima equivale perciò alla non-derivabilità di almeno un enunciato appartenente ad essa. È possibile, pertanto, denotare con $\mathbf{Cons_{PRA}}$ una generica formula: $\exists y \neg \exists x \mathbf{Dim_{PRA}}(x, y)$, la quale rappresenta in PRA la seguente relazione aritmetica tra numeri di Gödel: $\exists n \neg \exists m \mathrm{Dim}_{PRA}(m, n)$. Quest'ultima traduce aritmeticamente la proposizione metamatematica 'esiste almeno un enunciato di PRA (il cui gödeliano è *n*) tale che non esiste alcuna dimostrazione (il cui gödeliano è *m*) di quell'enunciato.

Tuttavia, si presti attenzione al fatto che l'enunciato di consistenza di PRA è rappresentabile anche dall'enunciato G, il quale afferma, per l'appunto, la sua stessa inderivabilità. Da ciò segue, quindi, che $\mathbf{Cons_{PRA}}$ è equivalente a G, più precisamente: $\mathbf{Cons_{PRA}} \leftrightarrow G$. A questo punto, possiamo riformulare (1) nel modo seguente:

(3) $$\mathbf{Cons_{PRA}} \rightarrow \neg \exists x \mathbf{Dim_{PRA}}(x, \overline{g}).$$

Il corollario del teorema VI, ovvero il secondo teorema di incompletezza, afferma perciò:

(4) Se PRA è consistente, allora $\mathbf{Cons_{PRA}}$ non è derivabile in PRA: $\nvdash_{PRA} \mathbf{Cons_{PRA}}$.

La dimostrazione è immediata, ed è analoga a quella usata nella formulazione semantica del teorema data sopra. Si assuma la derivabilità in PRA di $\mathbf{Cons_{PRA}}$. Per la regola del *modus ponens*, da (3) segue $\neg \exists x \mathbf{Dim_{PRA}}(x, \overline{g})$. Ma da (1) segue che $\neg \exists x \mathbf{Dim_{PRA}}(x, \overline{g})$ non è derivabile in PRA. Dunque, $\mathbf{Cons_{PRA}}$ non è derivabile in PRA.

---

[64] Ciò è legittimo, infatti, poiché l'ω-coerenza implica la coerenza.

[65] *Ex falso sequitur quodlibet* ("dal falso segue qualsiasi cosa"). Si tratta di un principio logico classico, per cui da una contraddizione è derivabile qualunque enunciato, ed è esprimibile nel calcolo proposizionale mediante la formula: $(A \wedge \neg A) \rightarrow B$.

Con le parole di Nagel e Newman, possiamo perciò concludere questa sezione dicendo che «se l'aritmetica è coerente, la sua coerenza non può essere dimostrata da un ragionamento metamatematico dotato di una rappresentazione nell'ambito del formalismo dell'aritmetica».[66] Nel prossimo capitolo analizzeremo le conseguenze formali di un simile risultato.

[66] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 103.

Capitolo 3

Le conseguenze dei teoremi di incompletezza per il programma formalista

A questo punto della discussione possiamo chiederci: se l'enunciato gödeliano G, che afferma indirettamente di se stesso di non essere derivabile in PRA, non è in effetti derivabile all'interno di quest'ultima, da ciò non deriva forse anche la verità di G? Come abbiamo già osservato, Gödel fece attenzione a non introdurre nella sua prova alcun riferimento alla nozione semantica di verità. La prova dei teoremi di incompletezza ha, infatti, sempre e solo a che fare con la nozione di dimostrabilità di un enunciato "all'interno" un sistema formale. Infatti, come si diceva, la crisi dei fondamenti della matematica è sorta in concomitanza con la scoperta delle geometrie non euclidee, che ha rimesso in discussione il metodo assiomatico classico. Per questo motivo nacque il bisogno di slegare la matematica dall'intuizione e dall'evidenza, cioè dal suo dipendere da postulati che rimandano, "verso l'esterno", ad un piano ontologico che trascende quello della teoria[67] e su cui l'intelletto non può avere presa diretta, se non attraverso metodi empirici suscettibili di falsificazione. Perciò il criterio fondante da utilizzare nella costruzione dell'edificio matematico passò da essere quello della verità di postulati autoevidenti, a quello della dimostrabilità di un teorema all'interno di un sistema formalizzato, e quindi alla coerenza di un tale sistema comprendente l'intero delle nozioni matematiche; un sistema basato cioè su di un'«evidenza simbolica»,[68] ossia su di un insieme di assiomi e regole composto solamente di segni primitivi non considerati in relazione ad alcun un referente semantico.

Come abbiamo visto, l'obiettivo formalista di Hilbert era duplice: in primo luogo, conseguire la "chiusura" della matematica, cioè dimostrare la decidibilità di ogni problema matematico all'interno di un unico sistema formale; in secondo luogo, fornire una prova assoluta, diretta, della coerenza della matematica, che prevedesse l'utilizzo esclusivo di procedimenti metamatematici finitari, non facenti cioè appello a nozioni intuitive di infinito in atto. Riguardo al primo punto, il primo teorema di incompletezza mostra precisamente che non si può definire un'unica teoria all'interno della quale non vi siano *ignorabimus*. Infatti, all'interno di un dato sistema assiomatizzato sintatticamente coerente, in grado di rappresentare le funzioni-relazioni ricorsive, esiste almeno un enunciato che non può essere deciso a partire dai suoi soli assiomi e regole. In altri termini, tale enunciato risulta vero rispetto al modello standard della teoria formalizzata, ma non è derivabile all'interno di quest'ultima. Questo, infatti, è garantito anche dal teorema di Tarski, che stabilisce

[67] In questo proposito anticipiamo qualcosa di cui più precisamente diremo nel seguito, vale a dire che non ha senso sottrarre l'ambito delle teorie formali al riferimento a una qualche forma di realtà.

[68] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 194.

l'impossibilità di definire le condizioni di verità (l'impianto semantico) per gli enunciati di una teoria all'interno di quest'ultima. A tal fine, è necessaria pertanto l'elaborazione una metateoria (o un'estensione conservativa) dotata di un apparato semantico più potente della teoria di partenza.

Il secondo teorema di incompletezza, invece, dimostra l'insostenibilità anche del secondo punto del programma formalista sopra illustrato, ovvero l'impossibilità che si dia una prova assoluta della coerenza per l'aritmetica mediante soli procedimenti finitari. In proposito, Nagel e Newman sottolineano:

> La possibilità di costruire una dimostrazione finitistica assoluta di coerenza per l'aritmetica non è esclusa dai risultati di Gödel. Gödel mostrò che non è possibile alcuna prova che sia rappresentabile nell'ambito dell'aritmetica. Il suo ragionamento non elimina la possibilità di prove strettamente finitistiche non rappresentabili nell'ambito dell'aritmetica. Oggi, però, nessuno ha un'idea chiare del probabile aspetto di una prova finitistica *non* suscettibile di una rappresentazione o formulazione aritmetica.[69]

Infatti, nonostante Hilbert non avesse definito se non in maniera vaga il concetto di finitarietà che una simile prova diretta avrebbe dovuto possedere, è tuttavia lecito sostenere che un suo requisito cardine sia la rappresentabilità dei procedimenti metamatematici all'interno dello stesso sistema formale intorno a cui vertono. La procedura di gödelizzazione, infatti, consente proprio di rappresentare aritmeticamente le proposizioni metamatematiche intorno alla sintassi del sistema formale. Le relazioni aritmetiche così ottenute, a loro volta, sono formalizzabili all'interno del sistema stesso mediante numerali. Il secondo teorema mostra esattamente che il sistema può dimostrare la "propria" coerenza se, e solo se, è in grado di fornire una dimostrazione del suo enunciato gödeliano; e dal momento che questo è indecidibile a partire soltanto dalle sue regole e dai suoi assiomi, ne deriva che il sistema non dispone di una prova finitaria, o, per così dire, "interna" della sua stessa coerenza. Da ultimo, Galvan evidenzia che, a patto che una certa teoria formale sia un'estensione conservativa di PRA, che egli sostiene rappresentare «il sistema appropriato ad esprimere l'apparato logico di una qualsiasi teoria formale»,[70] le conseguenze dei teoremi risultano valide per la matematica in generale. Ora, Galvan mette in luce che i teoremi di incompletezza ci aiutano a comprendere

> qualcosa di molto profondo a proposito della natura della matematica, da intendersi quale una forma di sapere aperto, costituito da una progressione senza limite di teorie via via più inclusive,

[69] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 105.
[70] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 195.

incapaci di autogiustificarsi prese una ad una ma tali che le teorie più deboli sono perfettamente giustificabili partendo dalla prospettiva di quelle più potenti. […] Ciò avviene perché nel passaggio a teorie più potenti ha luogo, per così dire, un cambio di prospettiva, attuato attraverso l'introduzione di nuovi concetti o regole che, descrivendo sotto una veste linguistica finita una infinità di casi o di processi illimitati, permettono di aggirare le difficoltà connesse con la non effettività di operazioni infinitarie, le sole che nella teoria più debole consentirebbero di ottenere le formule altrimenti inderivabili.[71]

In altre parole, il rafforzamento di una teoria è conseguibile in due modi: da un lato, è possibile formalizzare la nozione di verità relativa all'interpretazione privilegiata di una data teoria nel linguaggio di una metateoria di volta in volta più potente dal punto di vista espressivo; dall'altro, l'estensione della teoria si ottiene attraverso «la semplice aggiunzione delle equivalenze definitorie per il concetto di formula vera» in essa. La differenza tra i due metodi, osserva Galvan, consiste nel fatto che, rispetto al secondo, «le definizioni, alla stregua di veri e propri nuovi assiomi, svolgono un ruolo creativo, dando luogo ad estensioni non inessenziali e non conservative della teoria rafforzata».[72]

I teoremi d'incompletezza hanno esercitato un enorme influenza anche al di fuori della logica, ossia in quel campo della teoria della conoscenza noto in inglese come *Philosophy of Mind*. Essi hanno contribuito ad alimentare un forte senso di incertezza nei confronti degli innumerevoli sforzi della ricerca volta a scoprire i fondamenti non solo della matematica e della logica, ma anche del sapere scientifico in generale. «Il fallimento del programma hilbertiano implica, con la negazione di una fondazione assoluta, una falsificazione del ruolo di *onnipotenza fondativa* attribuito alla logica».[73] Se, infatti, il sapere matematico non risulta in grado di autofondarsi mediante l'uso esclusivo di inferenze logiche, allora sembrerebbe, a maggior ragione, che lo stesso debba valere anche altre branche del sapere. Da ciò emerge la necessità di procedimenti di fondazione, o di giustificazione, che non abbiano la pretesa di essere assoluti, ma che siano, al contrario, condizionati e di natura extra-logica, come ad esempio i principi della logica dell'evidenza che approfondiremo più avanti. Una tale giustificazione appartiene a quello che Galvan definisce «ordine *aletico* della spiegazione», ossia instaura un rapporto di fondazione di tipo "esplicativo". Torneremo su questi concetti più dettagliatamente nel prosieguo del lavoro.

[71] Cfr. S. Galvan, *Pensiero e Logica*, in «Rivista di Filosofia Neo-Scolastica», LXXXIX, 1987, p. 122.
[72] Cfr. *ivi*, p. 123.
[73] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 196. Il corsivo è dell'autore.

Capitolo 4

La macchina di Turing e la formalizzazione della nozione di algoritmo

Prima di prendere in considerazione le osservazioni dello stesso Gödel in merito ai risultati dei suoi teoremi, che egli espresse in occasione di una lezione seminariale da lui tenuta a Providence nel 1951, è importante fare una breve digressione per spiegare alcuni altri concetti chiave che interessano il nostro lavoro. Alan M. Turing fu un matematico britannico a cui si deve, tra le altre cose, la teorizzazione della celebre "macchina di Turing": un'idealizzazione che influenzò notevolmente una lunga tradizione di ricerca nell'ambito della computazione. Il punto di partenza della riflessione di Turing fu il famoso "problema della decisione" (*Entscheidungsproblem*) proposto da Hilbert, il quale consiste nel tentativo di individuare un insieme di procedure di calcolo meccaniche talmente potente per riuscire a tradurre nel linguaggio dell'aritmetica, e a portare a termine, tutti i ragionamenti deduttivi che la mente umana è in grado di compiere. In questo senso, quello che Hilbert aveva in mente era un algoritmo universale, capace di giungere a una decisione intorno alla determinazione vero-funzionale di ogni formula appartenente ad un sistema formale logico del primo ordine. Turing venne a conoscenza del problema di Hilbert e dei teoremi di incompletezza di Gödel frequentando il corso tenuto da Max Newman a Cambridge del 1935 intorno ai fondamenti della matematica, ed in quell'occasione cominciò ad interessarsi alla questione analizzando il modo in cui l'essere umano esegue calcoli aritmetici al fine di fornire una definizione operativa e rigorosa di procedura algoritmica. Da quest'osservazione nacque l'idea della macchina di Turing, ovvero di una macchina in grado computare gli stessi calcoli eseguibili attraverso un algoritmo.[74] Per quanto riguarda la descrizione della struttura di una macchina di Turing e del suo funzionamento mi servirò a grandi linee dell'esposizione chiara e concisa che ne dà Berto.[75] Essa è costituita da:

- un nastro di carta di lunghezza infinita suddiviso in caselle di uguale dimensione ognuna delle quali può essere vuota oppure contenere uno tra i simboli primitivi (in numero finito) di un linguaggio formale scelto.
- un dispositivo, o "testina", in grado di scorrere nei due versi lungo la direzione del nastro, leggere il contenuto delle caselle, oppure cancellarlo e riscriverlo sostituendo un nuovo simbolo al simbolo precedente.

[74] Cfr. M. Davis, *Il calcolatore universale*, trad. it. di G. Rigamonti, Adelphi, Milano 2018, pp. 190-192; tit. orig. *The Universal Computer*, 2000.

[75] Cfr. F. Berto, *Tutti pazzi per Gödel!*, pp. 136-137.

- “stati interni”, ovvero l’insieme di ciascuno dei singoli stati in cui la macchina si trova in un dato momento. Uno stato interno può coincidere con lo stato iniziale, con quello finale, oppure con uno stato intermedio. Ogni stato può venire modificato soltanto dopo l’esecuzione di un’istruzione.
- Istruzioni nella forma di “quintuple” (gruppi di cinque simboli) che specificano l’azione che la macchina deve eseguire ad ogni stato. L’insieme delle istruzioni viene detto “programma”. Indicando con R ed S due stati interni della macchina, Davis fornisce il seguente esempio: «*quando la macchina si trova nello stato* R *e legge sul nastro il simbolo* a*, sostituisce* a *con* b*, si sposta di una casella a destra e passa nello stato* S».[76] Ciò può venire formulato mediante i seguenti simboli: R a : b → S.

È evidente che l’operare di una simile macchina segue un procedimento algoritmico e deterministico, ovvero univocamente determinato ad ogni stato in cui essa si trova, e che il linguaggio formale in cui viene scritto il programma della macchina, ovvero l’insieme delle sue istruzioni, è arbitrario (i computer di oggi, ad esempio, utilizzano il sistema binario). Inoltre, una macchina di Turing può venire istruita in modo tale che, a partire da certi *input*, si arresti una volta raggiunto un certo stato, che coinciderà con lo stato finale, oppure continui a computare *ad infinitum*.

Con un procedimento analogo a quello di Gödel, Turing elaborò una codifica attraverso la quale le quintuple che istruiscono una macchina di Turing vengono associate univocamente a stringhe di numeri naturali. I codici numerici delle quintuple costituiscono nel loro insieme un certo numero naturale che codifica il programma della macchina. Ora, tenendo a mente che il comportamento di una macchina è determinato dai dati iniziali in ingresso, è possibile affermare che alcuni numeri in ingresso determinano l’arresto della macchina ad un certo stato finale, mentre altri fanno sì che il computo della macchina non termini. L’insieme dei numeri del primo tipo viene detto “insieme di fermata” di una macchina. Pertanto, attraverso il metodo di diagonalizzazione di Cantor, è possibile costruire un insieme di numeri naturali D, contenente tutti e soli i codici numerici dei programmi di macchine di Turing, tale da risultare diverso dall’insieme di fermata di qualsivoglia macchina di quel tipo. In altri termini, il numero che codifica il programma di una certa macchina di Turing non può appartenere all’insieme dei numeri che compongono l’insieme di fermata di quella macchina. A partire da questo risultato, Turing fu in grado di elaborare una dimostrazione del fatto che il problema della decisione di Hilbert non può avere una soluzione, ovvero che esiste un problema matematico non risolvibile algoritmicamente.[77] Per chiarire questo punto, riportiamo l’esempio di Davis e

[76] Cfr. M. Davis, *Il calcolatore universale*, cit., p. 197.
[77] Cfr. *ivi*, pp. 203-206.

prendiamo in esame il seguente problema: «*trovare un algoritmo che dato un numero naturale stabilisca se appartiene o no all'insieme* D».[78] Se esistesse un simile algoritmo, allora esisterebbe anche una macchina di Turing ad esso equivalente. Tale macchina, inoltre, potrebbe venire istruita in modo tale da arrestarsi qualora il numero appartenesse a D, e da continuare a computare all'infinito nel caso contrario. In questo modo, l'insieme di fermata di tale macchina verrebbe fatto coincidere con l'insieme D. Ma per la definizione di D ciò è impossibile. Dunque, l'ipotesi di partenza è errata, e il problema avanzato è indecidibile.

Da ultimo, va ricordato che, sebbene a partire da un approccio completamente diverso, ovvero a partire dalla nozione di ricorsività generale come definizione rigorosa del concetto intuitivo di algoritmo, negli stessi anni anche il matematico americano di Princeton Alonzo Church giunse alla medesima conclusione di Turing circa l'esistenza di problemi matematici insolubili. Turing, venuto a conoscenza di ciò, fu in grado di dimostrare l'equivalenza tra la nozione di ricorsività generale di Church e quella di computabilità da lui sviluppata. È dunque per questo motivo che è divenuto d'uso comune riferirsi alla tesi per cui ogni funzione effettivamente calcolabile (o algoritmo) corrisponde ad una funzione ricorsiva generale o computabile come alla tesi di Church-Turing.[79] Va sottolineato però che siamo di fronte – in entrambi gli autori – ad una congettura, e quindi ad un'ipotesi che non è verificabile in maniera definitiva, proprio a causa della vaghezza che pertiene alla nozione di algoritmo. Ciononostante, come evidenza Berto, il fatto che oltre a quella di Church esistano altre formalizzazioni della nozione di algoritmo e che queste siano state «dimostrate *equivalenti* fra loro […], avvalora ciascuna di esse singolarmente presa».[80] Da quanto detto, risulta pertanto che ad ogni sistema formale corrisponde una macchina di Turing che dimostra ("stampa") tutti e soli i teoremi che appartengono a quel sistema, e viceversa.

---

[78] Cfr. *ivi*, p. 208. Il corsivo è dell'autore.

[79] Cfr. *ivi*, pp. 213-214.

[80] Cfr. F. Berto, *Tutti pazzi per Gödel!*, p. 91. Il corsivo è dell'autore. Questo fatto è stato riconosciuto anche dallo stesso Gödel. Nell'*incipit* della Gibbs Lecture del 1951, che esamineremo nel prossimo paragrafo, a proposito dell'indagine sui fondamenti della matematica e dei risultati conseguiti in quei decenni, egli afferma che «il miglioramento più grande è stato reso possibile tramite una definizione rigorosa del concetto di procedimento finito [o algoritmico] […]». E in merito alle diverse soluzioni avanzate, egli continua dicendo: «[i]l più soddisfacente, a mio parere, è la riduzione del concetto di procedimento finito a quello di una macchina con un numero finito di componenti, come è stato fatto dal matematico inglese Turing» (Cfr. K. Gödel, *Alcuni teoremi basilari sui fondamenti della matematica e loro implicazioni filosofiche (1951)*, in *Opere. Volume 3: Saggi inediti e conferenze*, trad. it. di E. Ballo, G. Lolli, C. Mangione e P. Pagli, Bollati Boringhieri, Torino 2006, p. 268; tit. orig. *Some basic theorems on the foundations of mathematics and their implications*, in *Collected Works. Volume III: Unpublished Essays and Lectures*, Oxford University Press, New York, 1995).

Capitolo 5

La *Gibbs Lecture*: Gödel a Providence

Dopo il matematico John Von Neumann, Nagel e Newman sono stati forse tra i primi a riconoscere la portata dell'impatto che il lavoro di Gödel ha esercitato sulla riflessione filosofica intorno al problema mente-corpo da cui il nostro lavoro ha preso le mosse. A tale proposito, un punto di notevole interesse è che le «conclusioni di Gödel fanno sorgere la questione se sia possibile costruire una macchina calcolatrice che faccia concorrenza al cervello umano in fatto di intelligenza matematica».[81]. La conclusione dei due autori è che, sebbene allo stato attuale delle cose (e questo vale sia per la loro epoca che per quella presente) le macchine o i computer non siano in grado di eguagliare le capacità di ragionamento e dimostrazione della mente umana, ciò non significa che «non abbiamo il diritto di sperare di spiegare la materia vivente e la ragione umana in termini fisici e chimici». «La possibilità di una tale spiegazione» - essi continuano - «non è né preclusa né affermata dal teorema di incompletezza di Gödel».[82]

Allo scopo di far luce su quanto appena detto occorre prendere in esame la celebre Gibbs Lecture, una lezione seminariale tenuta da Gödel in occasione del congresso dell'American Mathematical Society alla Brown University il 26 dicembre 1951. Nella prima parte dell'intervento, quella che più ci interessa, Gödel si concentra sul «fenomeno della inesauribilità della matematica», il quale «è sempre presente in qualche forma, indipendentemente dal punto di vista assunto»[83]; un fenomeno a cui conducono i risultati dei suoi teoremi e della ricerca intorno ai fondamenti della matematica di quei decenni. Assumendo, come Gödel fa, la completa riducibilità della matematica in senso stretto alla teoria degli insiemi,[84] la questione centrale è quella dell'assiomatizzazione di quest'ultima. Ciò a cui si giunge in proposito, però, non è il risultato tanto atteso da Hilbert; al contrario, se ne conclude l'impossibilità di costruire un impianto assiomatico finito che comprenda tutti gli assiomi in una procedura finitaria in grado di determinarli. Il carattere di essenziale

[81] Cfr. E. Nagel, J.R. Newman, cit., p. 107.

[82] Cfr. *ivi*, p. 108.

[83] Cfr. K. Gödel, *Alcuni teoremi basilari sui fondamenti della matematica e loro implicazioni filosofiche (1951)*, in *Opere. Volume 3: Saggi inediti e conferenze*, cit., p. 269. Nelle successive citazioni tratte da quest'opera, il corsivo è dell'autore.

[84] Gödel definisce la matematica in senso stretto come «quel nucleo di proposizioni matematiche che sono valide in senso assoluto, senza alcuna ipotesi ulteriore. Proposizioni cosiffatte devono esistere, perché altrimenti non esisterebbero neppure i teoremi ipotetici». Egli afferma, poi, che «il fenomeno dell'inesauribilità della matematica, tuttavia, è sempre presente in qualche forma, indipendentemente dal punto di vista assunto. Così posso anche spiegarlo partendo dall'ottica più semplice e naturale, che prende la matematica così come è, senza critiche che la limitino. In quest'ottica tutta la matematica è riducibile alla teoria astratta degli insiemi» (Cfr. *ibidem*).

incompletezza della matematica intesa in questo senso risulta particolarmente evidente dal "secondo teorema". Quest'ultimo, infatti, dice Gödel: «*rende impossibile che qualcuno costruisca un sistema di assiomi e regole rigorosamente definito e faccia su di esso la seguente affermazione: percepisco (con certezza matematica) che questi assiomi con queste regole sono corretti e credo inoltre che comprendano tutta la matematica.* Se qualcuno formula questa affermazione si contraddice».[85] Ciò deriva dal fatto che, all'interno di un dato sistema formale, l'assunzione della correttezza degli assiomi e delle regole d'inferenza di quel sistema non può essere dimostrata. Esiste quindi almeno un enunciato indecidibile interno al sistema, tale per cui il sistema non può perciò essere completo. A questo proposito è necessario distinguere tra la matematica «in senso oggettivo e soggettivo». La prima costituisce l'insieme di tutte le proposizioni matematiche vere, mentre la seconda l'insieme di tutte le proposizioni matematiche dimostrabili o decidibili. Perciò, se anche esistesse un sistema formale completo, ovvero contenente l'intero di tutte le proposizioni matematiche dimostrabili, non saremmo in grado di stabilire se tale sistema sia coerente o meno, «cioè non potremmo mai sapere con certezza matematica che tutte le proposizioni che [tale sistema] genera sono corrette».[86] Al massimo saremmo in grado di dimostrare soltanto una per volta ogni proposizione generata dagli assiomi di un tale sistema. Ciò che segue rappresenta un punto di cardinale importanza:

> Però l'affermazione che sono vere tutte [le proposizioni generate dall'ipotetico sistema in grado di derivare tutta la matematica soggettiva], potrebbe al più essere ricavata con certezza empirica sulla base di un numero sufficiente di casi particolari o tramite altre inferenze induttive. Se fosse così, ciò significherebbe che la mente umana (nell'ambito della matematica pura) *è* equivalente a una macchina finita che però non è in grado di comprendere completamente il suo stesso funzionamento. Questa incapacità dell'uomo di comprendere se stesso potrebbe apparirgli erroneamente come una sua [(della mente)] [*sic*] illimitatezza o inesauribilità. […] Così risulta inevitabile il dilemma seguente: *o la matematica è incompletabile in questo senso: che i suoi assiomi evidenti non possono mai essere compresi in una regola finita, vale a dire la mente umana (perfino nell'ambito della matematica pura) supera infinitamente la potenza di ogni macchina finita, oppure esistono problemi diofantei del tipo indicato assolutamente insolubili* (e il caso che ambedue i termini del dilemma siano veri non è escluso, così che ci sono, a rigore, tre alternative). È questo fatto stabilito matematicamente che mi sembra di grande interesse filosofico.[87]

L'espressione "assolutamente insolubili" si riferisce a problemi non dimostrabili da nessuna procedura concepibile dall'uomo e non soltanto a partire da qualche determinato sistema

[85] Cfr. *ivi*, p. 272.
[86] Cfr. *ivi*, p. 273.
[87] Cfr. *ibidem*.

assiomatizzato, mentre un problema diofanteo[88] consiste in un'equazione a coefficienti interi, la cui soluzione è richiesta nella forma di numeri interi.[89]

In una nota al testo, in riferimento agli eventuali metodi empirici adoperabili per dimostrare che tutte le proposizioni generate dall'ipotetico sistema sono vere, Gödel ammette la possibilità che la ricerca neurobiologica giunga a comprendere in maniera talmente esatta i meccanismi che regolano i processi cerebrali da ridurre il fenomeno della coscienza a quegli stessi processi, e quindi a spiegare il funzionamento della mente umana nei termini di una macchina di Turing. Tuttavia, George Boolos fa giustamente notare che l'interesse filosofico del dilemma è quantomeno dubitabile, dal momento che dall'esposizione di Gödel non risulta in modo chiaro in che modo la mente umana possa essere equiparata ad una macchina finita, e in particolare ad una macchina di Turing. A nostro avviso, la vaghezza imputata da Boolos all'ipotesi di Gödel riguarda principalmente la considerazione per la quale non è immediatamente ovvio che la mente umana, nel dimostrare e comprendere la verità di proposizioni matematiche, operi in maniera analoga ad una macchina di Turing, e non faccia uso di procedimenti alternativi non catturabili dalla potenza computativa di quest'ultima.

Secondo il matematico e filosofo americano Solomon Feferman, la questione centrale posta da Gödel è se la matematica soggettiva coincida o meno con la matematica oggettiva. Se si tralascia il problema riguardante il significato di un'equiparazione della mente a una macchina, il dilemma si riduce alle seguenti alternative: se le due coincidono, allora la mente umana è in grado di produrre dimostrazioni che superano le capacità di qualsiasi macchina di Turing; se non coincidono, allora esistono verità oggettive che non possono essere dimostrate in senso assoluto. Inoltre, Feferman evidenzia tre presupposti di fondo dell'argomentazione di Gödel che stanno alla base del suo dilemma:

a) Per dimostrare proposizioni matematiche, la mente umana procede soltanto per via di assiomi evidenti e regole d'inferenza che ne conservano il valore di verità.
b) Tali assiomi e regole comprendono quelli dell'Aritmetica di Peano (PA).
c) La macchina finita a cui la mente umana viene equiparata corrisponde ad una macchina di Turing che stampa soltanto i teoremi che essa è in grado di dimostrare.

[88] L'aggettivo deriva dal nome del matematico Diofanto di Alessandria.

[89] Nello specifico, un'equazione diofantea (o diofantina) consiste nello stabilire se un certo tipo di prodotto polinomiale a coefficienti interi a due variabili (tra cui le x_i sono le incognite e le y_i i parametri) si annulli per valori interi delle incognite a prescindere o meno dalla scelta di valori interi per i parametri. Ad esempio, ($ax + by = c$) è un'equazione diofantea che ammette infinite soluzioni intere, e soltanto una soluzione intera positiva ($x = 3$, $y = 4$).

Un primo problema fondamentale riguarda la prima assunzione. Come si diceva, non è immediatamente chiaro come funzioni la mente umana in generale, né come operi nel dimostrare nello specifico teoremi matematici. Inoltre, sembra evidente che Gödel assume una nozione ideale, non-finita, della mente umana, ovvero la intende come capace di estendere il proprio operare nel tempo indefinitamente.[90] Ma, come accade per la mente, nulla vieta di idealizzare anche la nozione di macchina finita. Come fa notare Feferman (in una lezione risalente al 2003), riguardo le cosiddette "macchine connessioniste" (*connectionist machines*) o "reti neurali" (*neural networks*) esiste soltanto un'idea molto generale di come esse funzionino, ma «non esiste alcuna spiegazione matematica definitiva di che cosa costituisca una macchina connessionista»[91]. Infine, ad essere problematica è l'assunzione dell'evidenza degli assiomi di cui la mente, secondo Gödel, si servirebbe nel dimostrare teoremi matematici. Il concetto di evidenza, in questo caso, è quantomeno vago. Inoltre, non ci sono prove definitive del fatto che, anche qualora il funzionamento della mente si adeguasse ad una struttura di tipo logico, un certo impianto assiomatico sarebbe sufficiente a rappresentare tale struttura.

David Hilbert, come sappiamo, non avrebbe avuto dubbi su quale dei due corni del dilemma di Gödel fosse quello vero. Egli, infatti, credeva che non esistessero problemi insolubili (*ignorabimus*) in linea di principio a tal punto da considerare questa convinzione addirittura un assioma.[92] Anche Gödel era d'accordo con Hilbert su questo punto, e nella seconda parte della Gibbs Lecture egli avanza numerosi argomenti di natura filosofica a favore del realismo matematico che, dal suo punto di vista, contribuirebbero a mostrare l'insostenibilità della tesi meccanicista o fisicalista.[93] Gödel, di fatto, non credeva che la mente fosse riducibile al funzionamento di una macchina di Turing, o che esistessero verità matematiche indimostrabili in linea di principio. Una

---

[90] In proposito si veda H. Wang, *A Logical Journey. From Gödel to Philosophy*, MIT Press, 1996, p. 189.

[91] Cfr. S. Feferman, *Are There Absolutely Unsolvable Problems? Gödel's Dichotomy*, in «Philosophia Mathematica», XXXV, 2006, vol. 14, no. 2, p. 8.

[92] Cfr. *ivi*, p. 11.

[93] «In corrispondenza con la formulazione in termini di dilemma del teorema principale sull'incompletabilità della matematica, le implicazioni filosofiche *prima facie* avranno anch'esse l'aspetto di una disgiunzione; però ambedue le alternative sono decisamente in contrasto con la filosofia materialistica. Precisamente, se vale la prima questo sembra implicare che il funzionamento della mente umana non può essere ridotto al funzionamento del cervello che secondo ogni apparenza è una macchina finita con un numero finito di componenti […]. Così manifestamente si è indotti ad assumere una qualche posizione vitalistica. D'altra parte la seconda alternativa, in cui esistono proposizioni matematiche assolutamente indecidibili, sembra confutare l'idea che la matematica sia soltanto una nostra creazione; perché il creatore conosce necessariamente tutte le proprietà delle sue creature dal momento che esse non possono averne altre eccetto quelle che lui ha dato loro. Così questa alternativa sembra implicare che gli enti e i fatti matematici (o almeno *qualcosa* in essi) esistano oggettivamente e indipendentemente dai nostri atti e dalle decisioni mentali, vale a dire […] una qualche forma di platonismo o "realismo" relativo agli enti matematici» (Cfr. K. Gödel, *Alcuni teoremi basilari sui fondamenti della matematica e loro implicazioni filosofiche (1951)*, in *Opere. Volume 3: Saggi inediti e conferenze*, cit., p. 275).

simile convinzione costituisce ciò che Hao Wang definisce "ottimismo razionalistico".[94] Tuttavia, dal modo dilemmatico in cui pose il problema e dalla sua ammissione della possibilità di una prova empirica dell'equivalenza tra mente e macchina di Turing, risulta evidente che Gödel non ritenne mai di poter confutare il meccanicismo in maniera definitiva.

Come abbiamo già accennato, due autori in particolare, Lucas e Penrose, sono noti per aver affermato che proprio a partire dai teoremi di incompletezza è possibile avanzare ragioni assai valide a favore della tesi per cui il funzionamento di una macchina di Turing non è assimilabile (né come duplicato né come simulazione) a quello della mente umana. I loro argomenti, ormai noti come argomenti gödeliani, hanno suscitato un grande interesse nel campo della filosofia della mente, ma sono per altro stati oggetto di dure critiche da parte di vari pensatori. Nel prossimo capitolo prenderemo in considerazione soltanto quelle obiezioni che risultano a nostro avviso più rilevanti: quelle avanzate da autori quali Hilary Putnam, Paul Benacerraf, Stewart Shapiro, John R. Searle, nonché le relative contro obiezioni di Lucas e Penrose.

[94] «Se essa [la seconda alternativa] fosse vera significherebbe che la ragione umana è assolutamente irrazionale a porre questioni cui non può rispondere, proprio mentre asserisce enfaticamente che solo la ragione può trovarne risposta. La ragione umana sarebbe allora davvero imperfetta e in un certo senso perfino incoerente, in contraddizione stridente col fatto che quelle parti della matematica che sono state sviluppate sistematicamente e completamente […] mostrano un grado stupefacente di bella e perfezione […]. Questi fatti sembrano giustificare quello che si potrebbe chiamare "ottimismo razionalistico" (Cfr. H. Wang, *Dalla matematica alla filosofia*, trad. it. di A. Giacomelli, Bollati Boringhieri, Torino 2016, p. 341; tit. orig., *From Mathematics to Philosophy*, Routledge & Kegan Paul, Londra 1974).

Capitolo 6

Gli argomenti gödeliani contro la tesi meccanicistica dell'identità tra mente umana e macchina di Turing

6.1 L'argomento di Lucas

Il lungo ed intenso dibattito della seconda metà del Novecento sulle implicazioni dei teoremi di Gödel in merito al progetto dell'intelligenza artificiale ed alla possibilità di stabilire l'irriducibilità della mente umana a una macchina finita ha inizio in seguito all'uscita del saggio intitolato *Minds, Machines and Gödel*, che Lucas lesse alla Oxford Philosophical Society il 30 ottobre del 1959. Nell'*incipit* egli afferma che il suo intento è quello di spiegare nei dettagli l'argomentazione per la quale risulta che i teoremi di incompletezza «provano che il meccanicismo è falso».[95] Occorre notare fin da subito che l'uso del termine "prova" qui non è affatto corretto. Infatti, ciò che i teoremi di Gödel dimostrano non va oltre la verità di un condizionale, ovvero che *se* un dato sistema assiomatico formalizzato, in grado di rappresentare una certa porzione di aritmetica, è coerente, allora quel sistema non è completo, in quanto non è capace di decidere l'enunciato che afferma la sua stessa coerenza. Questo, però, non significa immediatamente che sia dimostrabile l'antecedente di quel condizionale. Un'eventuale confutazione della posizione meccanicista può, tuttavia, ragionevolmente svolgersi sulla base di addizionali osservazioni di tipo filosofico. Come vedremo, infatti, Lucas è consapevole della reale portata dei teoremi di Gödel, ma intende spingere la sua argomentazione oltre l'ambito della logica formale. Come cercheremo di mostrare, si tratta di una direzione, a nostro avviso, corretta ed obbligata nel tentativo di sciogliere quello che Chalmers etichetta come il «duro problema della coscienza»:[96] una direzione, però, da Lucas soltanto indicata e non anche adeguatamente approfondita.

Innanzitutto, riteniamo che il testo di Lucas possa essere suddiviso in cinque punti fondamentali:

---

[95] «[*Gödel's Theorem seems to me*] *to prove that Mechanism is false*» (Cfr. J.R. Lucas, *Minds, Machines and Gödel*, in «Philosophy», XXX, 1961, vol. 36, no. 137, p. 112). Qui, come nelle seguenti citazioni tratte da questo saggio, la traduzione dall'inglese è ad opera nostra ed il corsivo è dell'autore.

[96] Cfr. D.J. Chalmers, *Facing Up to the Problem of Consciousness*, in «Journal of Consciousness Studies», II, 1995, vol. 2, no. 3, p. 201. I temi trattati da Chalmers in questo saggio sono esposti in modo approfondito in D.J. Chalmers, *The Conscious Mind. In Search of a Fundamental Theory*, Oxford University Press, 1996.

1) Il “gioco della gödelizzazione”, o la competizione tra macchina e mente.
2) Le inferenze non-deduttive.
3) L’inconsistenza della mente umana.
4) Gli argomenti informali a sostegno dell’irriducibilità della mente.
5) La complessità della mente e la sua emergenza.

1) Lucas afferma che sulla base dei risultati di Gödel è possibile asserire che «in ogni sistema coerente che è abbastanza potente da rappresentare l’aritmetica elementare [l’aritmetica di Peano] esistono formule che non possono essere dimostrate all’interno del sistema, ma che noi siamo capaci di vedere che sono vere».[97] Inoltre, prendendo in considerazione la tesi di Church-Turing da cui, come abbiamo visto, risulta possibile tradurre un sistema formale nei termini di una macchina di Turing, egli conclude dicendo: «segue che nessuna macchina può rappresentare un modello della mente completo o adeguato, che le menti sono essenzialmente differenti dalle macchine».[98] Inoltre, Lucas ritiene che «per quanto complessa [sia] una macchina che costruiamo, essa, se è una macchina, corrisponderà ad un sistema formale, il quale a sua volta sarà sottoponibile alla procedura di Gödel per scoprire una formula indimostrabile in quel sistema». «Questa formula» – continua Lucas – «la macchina non sarà in grado di produrla come vera, nonostante una mente possa vedere che è vera».[99] Egli pone perciò la questione nei termini di un gioco, in cui ad ogni stadio spetta la mossa a ciascuno dei due concorrenti: da un lato il meccanicista, che comincia il turno avanzando un modello meccanico definito della mente; dall’altro la mente umana, la quale, «grazie al teorema di Gödel, […] ha sempre l’ultima parola».[100] Un’importante questione che Lucas prende in considerazione riguarda il fatto che la procedura di gödelizzazione e di costruzione dell’enunciato gödeliano di un sistema formale S è interamente meccanica e iterabile ed è, pertanto, inseribile in una macchina di Turing – T – sottoforma di una specifica regola di inferenza aggiuntiva. Tale macchina sarebbe perciò in grado di produrre l’enunciato gödeliano di S ed aggiungerlo ad S come nuovo assioma, producendo così un nuovo sistema S’. Analogamente, T produrrebbe il gödeliano di S’ e lo aggiungerebbe all’impianto

[97] «*In any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true*» (Cfr. J.R. Lucas, *Minds, Machines and Gödel*, cit., p. 112).

[98] «*It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines*» (Cfr. *ivi*, p. 113).

[99] «*However complicated a machine we construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-the-system. This formula the machine will be unable to produce as being true, although a mind can see that it is true*» (Cfr. *ivi*, p. 116).

[100] «*Thanks to Gödel’s theorem,* […] *always has the last word*» (Cfr. *ibidem*).

assiomatico di quest'ultimo, producendo un nuovo sistema S'' e così via. Il problema sta nel fatto che anche rispetto a T esiste un enunciato gödeliano indecidibile in T. In altri termini, a T sarebbe precluso operare la procedura di gödelizzazione rispetto a se stessa. Al contrario, una mente, sostiene Lucas, potrebbe produrre il gödeliano di T e riconoscerlo come vero. Ciò è quanto dovrebbe bastare per rispondere alla nota osservazione di Turing secondo la quale, allo stesso modo delle macchine, anche noi esseri umani saremmo limitati nel nostro operare. Tuttavia, si tratta qui non di una questione di superiorità tra mente e macchina, bensì di essenziale differenza: "vedere" la verità di un enunciato che una macchina avanzata come modello della nostra mente non è in grado di dimostrare, è sufficiente a stabilire tale differenza.[101]

2) Un'obiezione che l'autore prende in considerazione riguarda l'impossibilità di rigettare con certezza l'eventualità che la mente umana funzioni in modo inconsistente (o incoerente), e di conseguenza, di rigettare un modello meccanico della mente, come quello proposto da Hartley Rogers, che «dovrebbe consentire la possibilità di compiere inferenze non-deduttive», le quali «potrebbero fornire un modo per sfuggire al risultato di Gödel».[102] Nello specifico, l'idea di Rogers è quella di una macchina dotata di una sorta di capacità critica, che accetti in maniera provvisoria determinate proposizioni non ancora dimostrate, e che, qualora si rivelasse necessario, le rigetti oppure le aggiunga definitivamente alla lista dei suoi assiomi. Tuttavia, fa notare Lucas, la costruzione di una macchina siffatta andrebbe incontro a diverse difficoltà. Innanzitutto, essa dovrebbe essere progettata in modo tale da non risultare banalmente incoerente e quindi da non poter dimostrare come vera qualunque formula. In secondo luogo, posta di fronte ad un enunciato indecidibile, essa non dovrebbe accettare come vera sia l'affermazione che la negazione di esso, bensì dovrebbe attenersi a determinati criteri di selezione. Questi ultimi sarebbero sì specificabili, ma dipenderebbero, in ultima istanza, da una scelta arbitraria da parte del programmatore, ed inoltre non consentirebbero alla macchina di decidere intorno all'enunciato gödeliano suo proprio. Questo costituisce un sotto-caso analogo a quello esposto nel punto precedente relativamente alla macchina T, programmata per iterare la procedura di gödelizzazione.

3) Lucas riconosce che dai teoremi di Gödel è possibile dimostrare formalmente soltanto che *se* un sistema è coerente,[103] allora il suo enunciato gödeliano è indimostrabile in esso. Egli ammette anche che non esiste una prova assoluta della coerenza di quel sistema, e che «al massimo possiamo dire che la macchina è coerente, posto che noi lo siamo». «Ma con quale diritto» - si chiede -

[101] Cfr. *ivi*, pp. 116-117.

[102] «*Would have to allow for the possibility of making non-deductive inferences, and these might provide a way of escaping the Gödel result*» (Cfr. *ivi*, p. 119).

[103] E non "completo" come scrive invece Lucas (Cfr. *ivi*, p. 120).

«possiamo farlo?».[104] Gödel stesso, come abbiamo visto, non esclude che la mente umana possa essere una macchina di Turing talmente complessa da non comprendere il proprio funzionamento e quindi da non essere in grado di stabilire la propria coerenza. Lucas espone dunque una serie di argomenti con cui tenta di mostrare che l'incoerenza umana non è programmatica, ma soltanto occasionale.[105] Il modo in cui il pensiero dell'uomo si istanzia nella ricerca scientifica e filosofica, infatti, non ammette la dimostrabilità di qualunque proposizione, ma procede selettivamente per prove ed errori e, se necessario, è pronto a revisionare quanto ottenuto qualora, si scoprissero delle contraddizioni. Perciò, sebbene l'eventualità dell'attestazione dell'incoerenza del nostro modo di ragionare non possa escludersi a priori, ciononostante, se fosse il caso che il sistema del nostro sapere (che comprende l'aritmetica elementare sulla base della quale viene dimostrato il teorema di incompletezza) non può essere né coerente né corretto, allora, secondo Lucas, «dovremmo abbandonare non solo tutta la matematica […] ma la totalità del pensiero».[106]

4) Nell'ipotesi che la mente umana sia una complessa macchina di Turing, essa non può effettivamente provare la propria coerenza. Tuttavia, se partiamo dalla posizione per cui la mente non è riducibile a una macchina, allora nulla vieta, afferma Lucas, di «uscire dal sistema e […] produrre argomenti informali per la coerenza o di un sistema formale o di qualcosa di meno formale e sistematizzato».[107] A titolo di esempio, si consideri il carattere di autoreferenzialità dell'enunciato gödeliano. Esso determina un problema insolubile per una macchina, dal momento che implica, per sua natura, un riferimento a sé ed al proprio operare. Riprendendo la distinzione di Searle tra sintassi e semantica, potremmo dire che per la prospettiva anti-meccanicista una macchina o un sistema formale non hanno al loro interno alcun orizzonte di significato, alcuna semantica, ma solo un insieme finito di simboli primitivi la cui scelta è del tutto arbitraria. Secondo Lucas, la macchina, perciò, non avrebbe alcuna comprensione del senso delle proprie computazioni, né sarebbe in grado di abbracciare, per così dire, in un solo sguardo la propria totalità per trascenderla, e quindi per "uscire dal sistema" di cui è un'istanziazione. A proposito poi di quella tipologia di questioni per le quali ci si chiede che cosa si pensa quando si pensa qualcosa, o chi è che pensa quando il pensiero si rivolge a se stesso, Lucas ritiene che si tratti di domande insensate, dal momento che è insito nel concetto stesso di coscienza un rimando alla consapevolezza di sé e quindi ad un continuo ed infinito

---

[104] «*At best we can say that the machine is consistent, provided we are. But by what right can we do this?*» (Cfr. *ivi*, p. 120).

[105] Come vedremo, simili argomenti vengono ripresi anche da Penrose (Cfr. R. Penrose, *Shadows of the Mind. A Search for the Missing Science of Consciousness* (1994), Vintage, Londra 2005).

[106] «*We shall have to abandon not merely the whole of mathematics* […] *but the whole of thought*» (Cfr. J.R. Lucas, *Minds, Machines and Gödel*, cit., p. 123).

[107] «[*But there is no objection to*] *going outside the system and* [*no objection to*] *producing informal arguments for the consistency either of a formal system or of something less formal and less systematized*» (Cfr. *ivi*, p. 124).

sdoppiamento in soggetto e oggetto dell'atto del pensare. Egli, infatti, sostiene: «c'è, lo riconosciamo, un'infinità qui, ma non rappresenta un regresso all'infinito in cattivo senso […]. Al contrario, insistiamo nel dire che un essere cosciente è un'unità, e che nonostante parliamo di parti della mente, lo facciamo solo per via di metafora».[108] Tuttavia, in questo testo Lucas non sviluppa adeguatamente questa linea argomentativa. Pertanto, nel capitolo ottavo ci serviremo delle riflessioni di Galvan e di Pagani per mettere in luce ed analizzare il carattere di infinità e di unitarietà dell'"anima", intendendo con ciò la dimensione caratteristica dell'uomo, più ampia rispetto a quella della *mens* o della ragione calcolante. In questo modo, vediamo che Lucas, per difendere l'irriducibilità della mente rispetto a una macchina, ritiene necessario oltrepassare il campo della logica formale e produrre argomentazioni che traggono forza da considerazioni "informali", pertinenti alla disciplina filosofica in senso proprio.

5) L'ultimo punto riguarda un argomento avanzato da Turing[109] che concerne l'ipotesi dell'emergenza di un duplicato del fenomeno della coscienza a partire da macchine artificiali la cui complessità, in futuro, giunga a superare una certa soglia "critica". Per quanto poco plausibile, secondo Lucas, non si tratta di un'eventualità impossibile. Tuttavia, una macchina in cui venga ad emergere un comportamento autenticamente intelligente, non potrebbe più dirsi propriamente "macchina", almeno non nel senso specifico in cui Lucas intende tale termine, ovvero come un artefatto il cui comportamento sia interamente prevedibile a partire dal funzionamento delle sue singole parti.

6.2 Alcune obiezioni alla tesi di Lucas

Il saggio di Putnam, intitolato *Minds and machines*,[110] contiene una critica indiretta – divenuta ormai classica – all'argomento di Lucas; indiretta perché, in realtà, essa è rivolta alla seguente affermazione di Nagel e Newman nel loro testo del 1958, che riportiamo per intero:

> Le inerenti limitazioni delle macchine calcolatrici implicano che non abbiamo il diritto di sperare di spiegare la materia vivente e la ragione umana in termini fisici e chimici. La possibilità di una tale spiegazione non è preclusa né affermata dal teorema di incompletezza di Gödel. Il teorema in questione indica

[108] «*There is, we recognize, an infinity here, but it is not an infinite regress in the bad sense* […]. *Rather, we insist that a conscious being is a unity, and though we talk about parts of the mind, we do so only as a metaphor*» (Cfr. *ivi*, p. 125).

[109] Cfr. A.M. Turing, *Computing Machinery and Intelligence*, in «Mind», 1950, vol. LIX, no. 236, p. 454.

[110] Cfr. H. Putnam, *Minds and Machines*, in *Dimensions of Mind*, ed. S. Hook, New York 1960, e in *Philosophical Papers. Volume 2: Mind, Language and Reality*, Cambridge University Press, 1975.

che la struttura e la potenza della mente umana sono di gran lunga più complesse e sottili di qualunque macchina non vivente finora immaginata.[111]

La conclusione di questi due autori, per la quale una macchina di Turing non può rappresentare un modello adeguato della mente umana, come rileva giustamente Putnam, è scorretta. Sebbene Nagel e Newman non forniscano un'argomentazione esplicita per giungere a tale conclusione, Putnam immagina che una possa essere la seguente. Sia T una macchina di Turing che si suppone costituire un modello adeguato della mia mente, ovvero tale che derivi tutti gli enunciati matematici che io sono in grado di dimostrare. Allora, per i teoremi di incompletezza, io potrei venire a conoscenza di un enunciato di T che T non è in grado di derivare ma che io sono capace di dimostrare come vero, e con ciò dimostrerei di non essere identico a T. Ma un simile risultato, rileva Putnam, non è legittimamente derivabile dai teoremi di incompletezza:

> Data una macchina *T* arbitraria, tutto ciò che posso fare è trovare una proposizione *U* tale che *io* posso dimostrare:
>
> [C] Se *T* è consistente, *U* è vero,
>
> dove *U* è indecidibile in *T* se *T* è in effetti consistente. Tuttavia, anche *T* è perfettamente capace di dimostrare [C]! E l'enunciato *U*, che *T* non può derivare […], non posso dimostrarlo nemmeno *io* (a meno che io non possa dimostrare che *T* è consistente, il che è improbabile se *T* è molto complicata)![112]

Notiamo che un simile ragionamento, naturalmente, è valido anche contro l'argomento di Lucas. Sulla base di quanto appena detto, e «non tenendo conto di eventuali ragioni specifiche possedute dalla mente a sostegno della consistenza di T e non condivise dalla macchina» – afferma Galvan – «la mente non si trova in una situazione di superiorità rispetto alla macchina per il fatto di conoscere il valore di verità di [*U*]».[113]

---

[111] Cfr. E. Nagel, J.R. Newman, *La prova di Gödel*, cit., p. 108.

[112] «*Given an arbitrary machine* T*, all I can do is find a proposition* U *such that I can prove:* [C] *If* T *is consistent,* U *is true, where* U *is undecidable by* T *if* T *is in fact consistent. However,* T *can perfectly well prove* [C] *too! And the statement* U*, which* T cannot *prove* […], I *cannot prove either (unless I can prove that* T *is consistent, which is unlikely if* T *is very complicated)!*» (Cfr. H. Putnam, *Minds and Machines*, cit., p. 366). Il corsivo è dell'autore. La traduzione dall'inglese è ad opera nostra.

[113] Cfr. S. Galvan, *Gödel e il modello computazionale della mente*, in «Rivista di Filosofia Neo-Scolastica», XCVI, 2004, no. 1, p. 149.

Anche Paul Benacerraf, nel saggio del 1967, intitolato *God, the Devil, and Gödel*,[114] sostiene che l'uso dei teoremi di incompletezza da parte di Lucas non è legittimo, in quanto poggia su delle considerazioni metamatematiche, oltre che di tipo filosofico, e perché trae da essi una conclusione a suo avviso scorretta. Innanzitutto, egli sottolinea, anche se solo *en passant*, un punto importante su cui ci soffermeremo più avanti quando parleremo di macchine connessioniste e reti neurali, ovvero che rimane «una questione aperta se certe cose che non soddisfano le specificazioni di Turing possano contare anch'esse come macchine (ai fini del meccanicismo)».[115] Se la risposta a questo quesito fosse positiva, la tesi di Lucas perderebbe la sua efficacia, in quanto il suo attacco è diretto contro le macchine di Turing, che egli considera, giustamente, come implementazioni operative dei sistemi formali. Vedremo più avanti che la risposta che Penrose fornisce circa trent'anni dopo è negativa, ovvero che qualsiasi macchina connessionista è, almeno in linea di principio, traducibile nei termini di una macchina di Turing.

Secondo Benacerraf, poi, Lucas usa una nozione equivoca e vaga di "prova" quando afferma di essere in grado di "vedere" che l'enunciato indecidibile di un qualunque sistema formale è vero. Naturalmente, Lucas non ha in mente un concetto di prova come derivazione formale a partire da determinati assiomi e regole di inferenza, poiché è chiaro che, altrimenti, cadrebbe in una banale contraddizione. Al contrario, sebbene non sviluppi esplicitamente questo aspetto, riteniamo che egli si riferisca ad un senso lato di tale nozione che si rifà alla distinzione tra finito e infinito e all'idea di "evidenza veritativa". Approfondiremo più avanti questi concetti, quando cercheremo di mostrare in che modo la nozione di infinito trascendentale, introdotta in questo contesto da Pagani, costituisca un presupposto fondamentale per l'esercizio del pensiero, e come la mente, nel suo procedere, faccia appello a metodi basati su una forma particolare di evidenza o di intuizione.

Inoltre, Benacerraf sostiene che, anche ammesso che Lucas, o una mente umana idealizzata, fosse in grado di dimostrare l'incompletezza di qualunque macchina proposta da un sostenitore del meccanicismo come suo duplicato intelligente, ciò non confuterebbe affatto il meccanicismo, poiché in questo modo Lucas sposterebbe la discussione dall'essere in linea di principio all'essere fattuale. È chiaro, infatti, che ogni eventuale macchina di Turing che il meccanicista fosse in grado di realizzare, essa, in quanto costruita da una mente "finita", sarebbe per sua natura incompleta. Tuttavia, è pur sempre concepibile una macchina che superi in potenza ognuna delle macchine concretamente realizzabili dal meccanicista. Evidentemente la questione si gioca sulla distinzione tra finitarietà e infinitarietà, ovvero sulla distinzione che intercorre tra la successione potenzialmente infinita di tutte

[114] Cfr. P. Benacerraf, *God, the Devil, and Gödel*, in «The Monist», LIII, 1967, vol. 51, no. 1, pp. 9-32. Nelle successive citazioni prese da questo saggio, la traduzione dall'inglese è ad opera nostra.

[115] «*It is an open question whether certain things which do not satisfy Turing's specifications might also count as machines (for the purposes of Mechanism)*» (Cfr. *ivi*, p. 14).

le macchine finite specificabili come candidate a modello della mente umana e il concetto limite di una macchina che idealmente sia in grado di superare ciascuna di esse. È importante notare che, in relazione al dibattito intorno all'utilizzo dei risultati dei teoremi gödeliani per la refutazione della tesi meccanicista, come spiega anche Wang, «il paragone fra uomo e macchina di Turing […] fa ricorso a varie idealizzazioni».[116] Una di queste è l'assunzione della capacità della mente di operare illimitatamente nel tempo, dotata di una memoria senza limiti, e senza essere soggetta a logoramento ed ostacoli di tipo fisico.

Da ultimo, l'informalità degli argomenti a sostegno della coerenza della mente umana che Lucas propone alla fine del suo saggio, secondo Benacerraf, non sarebbe logicamente stringente. Per questo motivo, egli tenta di riformulare la tesi di Lucas in termini più rigorosi e formali.[117] La sua conclusione è che i teoremi di incompletezza non rigettano affatto, come abbiamo visto, la possibilità che un certo sistema formale, o la corrispondente macchina di Turing, rappresentino quell'insieme della matematica soggettiva a cui si riferiva Gödel nella Gibbs Lecture. Al contrario, essi stabiliscono che:

> data una qualsiasi macchina di Turing W_j, o io non posso provare che W_j è adeguata per [rappresentare] l'aritmetica, o, se io costituisco un sottoinsieme di W_j, allora non posso provare che sono in grado di dimostrare tutto quello che W_j dimostra. *Sembra essere consistente con tutto questo che io sono proprio una macchina di Turing, ma una con un programma talmente complesso che non posso determinare quale esso sia.* […] Certamente, potrei essere una macchina di Turing non-coerente [*inconsistent*].[118]

Un punto-chiave dell'argomentazione di Lucas è il fatto che, posto che il meccanicista avanzi uno specifico modello algoritmico della mente, sarebbe idealmente possibile analizzare il programma di quella macchina, risalire così al suo enunciato gödeliano, e mostrare quindi che la mente è capace di dimostrare più di quanto quella macchina non sia in grado di fare. Tuttavia, controbatte Benacerraf, questo dipende proprio dal modo in cui l'algoritmo viene presentato. Se al posto di una macchina di Turing dotata di un programma "trasparente", ovvero decodificabile, venisse presentata una «scatola nera (*black box*)» non sarebbe possibile costruire l'enunciato indecidibile del sistema formale rappresentato da quella macchina. A nostro avviso, tale osservazione, sebbene non errata se presa in

[116] Cfr. H. Wang, *Dalla matematica alla filosofia*, cit., p. 333.

[117] Per la ricostruzione dettagliata dell'argomento cfr. P. Benacerraf, *God, the Devil, and Gödel*, cit., pp. 23-30.

[118] «*Given any Turing machine W_j, either I cannot prove that W_j is adequate for arithmetic, or if I am a subset of W_j, then I cannot prove that I can prove everything W_j can.* It seems to be consistent with all this that I am indeed a Turing machine, but one with such a complex machine table (program) that I cannot ascertain what it is. *[…] Of course, I might be an inconsistent Turing machine*» (Cfr. *ivi*, p. 29). Il corsivo è dell'autore.

se stessa, non inficerebbe affatto il ragionamento di Lucas per due motivi principali. In primo luogo, se con "scatola nera" si intende un certo tipo di "intelligenza artificiale", cioè una certa macchina connessionista capace di "apprendere" nel corso del tempo e quindi di mutare il proprio comportamento in base all'apprendimento e ai fattori causali e casuali legati all'interazione con l'ambiente, allora – stando a quanto afferma Penrose – l'algoritmo che sta alla base di questa evoluzione e l'insieme dei fattori fisici concomitanti sarebbero di natura interamente computazionale e, in linea di principio, conoscibile dall'uomo. Illustreremo più nel dettaglio questo punto in seguito, quando esamineremo l'argomentazione di Penrose. Per il momento basti considerare che anche il funzionamento di una macchina connessionista è riconducibile, in ultima istanza, a quello di una macchina di Turing, ed è quindi completamente computazionale. In secondo luogo, se anche ammettessimo l'inconoscibilità di un tale algoritmo, con ciò la tesi meccanicista si svuoterebbe, in un certo senso, di contenuto, e si ridurrebbe ad una mera ipotesi non falsificabile a priori.

In sintesi, Benacerraf rimprovera a Lucas due cose: primo, di avere rappresentato quello che ho chiamato "gioco della gödelizzazione" come se avvenisse tra due menti fallibili, cioè quella di Lucas e quella del sostenitore del meccanicismo; secondo, di aver sorvolato sul fatto che la sua capacità di cogliere la verità dell'enunciato gödeliano di un sistema dipende principalmente dal fatto che tale sistema è trasparente, ovvero che il programma della macchina di Turing che lo implementa è deterministico. Riteniamo che tali obiezioni, oltre a non essere corrette, non colgono neppure la vera natura dialettica dell'argomentazione di Lucas. Per quanto riguarda il primo punto, Lucas assume esplicitamente l'idealità sia della mente umana che della macchina, il che vuol dire che è pronto ad ammettere una successione infinita di sistemi formali, e quindi di macchine di Turing, di potenza dimostrativa sempre maggiore, ciascuno dei quali in grado di dimostrare problemi indecidibili a partire dagli assiomi e dalle regole di un sistema meno potente. Per ciò che riguarda il secondo punto, invece, Benacerraf sembra fraintendere, o non tener debitamente in conto, ciò che è l'assunto principale di una tesi meccanicista che intenda porsi come "non vuota" di contenuto, ovvero che la mente sia effettivamente "una" certa macchina che funzioni secondo un "determinato" algoritmo, per quanto complesso quest'ultimo possa essere. Ora, sebbene non sia possibile escludere a priori la sua inconoscibilità, tuttavia ci sono ottime ragioni per credere che un simile algoritmo, se esistesse, sarebbe idealmente conoscibile, e che i suoi assiomi e le sue regole, qualora venissero specificati, sarebbero analizzabili e dimostrabilmente incompleti, e perciò inadeguati a rappresentare un modello completo della mente. Una tra tali ragioni, che trovano fondamento sul piano razionale-filosofico, è quella che abbiamo già incontrato nel capitolo secondo quando abbiamo esposto l'obiezione di Popper al fisicalismo. Essa riguarda la difficoltà di sostenere in modo razionale, ovvero con convinzione, una posizione secondo la quale non è da escludersi che il nostro pensiero, alla base, sia

inconsistente e per la quale quella dimensione che noi identifichiamo con il concetto di verità coincide invece con quello di utilità. Come sostiene anche Lucas, rinunciando a quella visione che viene chiamata "ottimismo razionalistico", e quindi rinunciando a credere nella capacità ideale della nostra mente di eradicare ogni errore logico dall'edificio del sapere umano, l'esercizio del pensiero verrebbe privato di senso, della sua stessa condizione di possibilità che, come abbiamo già detto, Pagani definisce "infinità trascendentale". Per questo motivo, se si vuole render giustizia all'argomentazione di Lucas, è pertanto importante tenere a mente il senso, a cui abbiamo accennato sopra, in cui dev'essere intesa la nozione di prova in riferimento alle capacità della mente umana; capacità che, come vedremo, sembrano andare ben oltre quelle della logica formale che sono le uniche cioè a cui una macchina ha, per così dire, accesso.

In ultima istanza, la questione centrale della posizione di Lucas che Benacerraf non sembra cogliere, a nostro avviso, è la seguente: di fronte a ciò che rappresenta l'evidenza dell'esistenza di qualcosa come l'insieme dei fenomeni coscienti, e del loro essere parte integrante della realtà, la visione meccanicista – la quale, come abbiamo visto, si identifica con la tesi dell'intelligenza artificiale forte, per cui la mente è interamente riducibile in termini computazionali – può porsi in due modi: o come una teoria determinata, o come un'ipotesi ultimativamente non confutabile. Nel primo caso, se accettiamo, sulla base di quanto detto sopra, l'essenziale correttezza della nostra capacità di comprensione e di ragionamento, oltre che alcune idealizzazioni relativamente alle capacità della mente umana, allora ogni specifica determinazione di un modello assunto come esaustiva e adeguata riproduzione della mente è rigettabile in base ai risultati di Gödel. Nel secondo caso, invece, il meccanicismo perde ogni forza coercitiva dal punto di vista dialettico, rifugiandosi in un'ipotesi, sebbene non logicamente confutabile, cionondimeno assai poco plausibile, o comunque inconsistente con i suoi stessi presupposti di fondo; presupposti vòlti ad una spiegazione della totalità della realtà dal punto di vista fisico-computazionale. In questo senso, tale posizione potrebbe venire accostata a quella dello scettico. Infatti, – e questo è un punto su cui è utile insistere – se l'ultima roccaforte del meccanicista fosse costituita dall'impossibilità di escludere a priori l'inconsistenza del pensiero umano, ciò equivarrebbe all'ammissione dell'intrinseca limitatezza della portata esplicativa della teoria meccanicista stessa; limitatezza che cozza con gli scopi della ricerca scientifica *tout court*, e non soltanto di quella mirata alla progettazione dell'intelligenza artificiale. Questo è quanto rinveniamo stare al fondo della posizione di Lucas e rispetto a cui, in ultima analisi, concordiamo. Come vedremo in seguito, sebbene a partire da presupposti diversi e mosso da ben altri intenti, anche Shapiro giunge ad una conclusione simile, relativamente al secondo caso appena esposto. Egli, infatti, afferma che la tesi meccanicista non è sufficientemente determinata per essere adeguatamente

confutabile a partire dalle implicazioni dei teoremi gödeliani. Rinviamo l'esposizione di questo punto al paragrafo §6.5.

6.3 Il carattere dialettico dell'argomento di Lucas

In un saggio del 1968, dal titolo *Satan Stultified: A Rejoinder to Paul Benacerraf*,[119] Lucas controbatte alle critiche mosse da Putnam e Benacerraf mettendo in luce un aspetto cardine del ragionamento che aveva esposto nel suo scritto originale anni prima, ovvero il carattere "dialettico" della sua confutazione della tesi meccanicista. Egli ritiene che l'argomento gödeliano possa essere inteso sia come una dimostrazione formale sia come una tipologia di argomentazione adattabile a differenti casi. La dialetticità della sua argomentazione dipende dal fatto che non si tratta tanto di una «prova diretta che la mente è qualcosa di più di una macchina», quanto di «uno schema di confutazione per ogni versione particolare del meccanicismo».[120] Lucas evidenzia che la forza della sua tesi risiede nel fatto che spetta al meccanicista l'onere della prova dell'equivalenza tra mente e macchina e che una volta che egli abbia fatto la "prima mossa" allora è possibile, secondo lui, mostrare come tale tentativo sia destinato al fallimento.[121] Possiamo già cogliere qui una certa somiglianza con il procedimento di *élenchos* che sarà oggetto d'esame nel proseguio del lavoro. Infatti, come è noto, *élenchos* è quella determinata forma argomentativa per cui la verità di una tesi, anziché essere dimostrata, risulta dal processo attraverso il quale viene mostrato che la sua negazione implica una contraddizione e con ciò l'affermazione degli stessi principi che sono sostenuti nella tesi. In questo senso, vedremo come la procedura elenctica, applicata alla fondazione della logica, non costituisca una prova assoluta di essa, bensì una forma di "eterofondazione".

---

[119] Cfr. J.R. Lucas, *Satan Stultified: A Rejoinder to Paul Benacerraf*, in «The Monist», LIV, 1968, vol. 52, no. 1, pp. 145-158. Anche in questo caso, nelle successive citazioni la traduzione dall'inglese è ad opera nostra ed il corsivo è dell'autore.

[120] «*A direct proof that the mind is something more than a machine, but a schema of disproof for any particular version of mechanism*» (Cfr. *ivi*, p. 145). In questo senso, possiamo dire che tale schema di confutazione ricalchi quello della confutazione dell'ω-coerenza di un sistema formale. Infatti, data una macchina di Turing T, e indicando con U l'insieme delle verità matematiche dimostrabili dalla mente umana, la teoria meccanicista M sostiene la tesi per cui $\exists T$ tale che $T = U$. Ma per ogni specifica macchina, vale $T_1 \neq U$, $T_2 \neq U$, …, $T_k \neq U$, …, ecc. (Cfr. *ivi*, p. 151).

[121] In proposito, Lucas scrive: «Una persona, o mente, idealizzata, potrebbe essere capace di fare più rispetto a quello che tutte le macchine logicamente possibili, tra di loro, sono in grado di fare come no: ma per ogni macchina logicamente possibile c'è qualcosa che essa può fare mentre questa no; e quindi essa non può essere identica a nessuna macchina logicamente possibile» («*An idealized person, or mind, may not be able to do more than all logically possible machines can, between them, do: but for each logically possible machine there is something which he can do and it cannot; and therefore he cannot be the same as any logically possible machine*») (Cfr. *ivi*, p. 146).

Per quanto riguarda invece la vaghezza che Benacerraf evidenzia nella nozione di prova usata da Lucas nel saggio del 1959, è utile ricordare la distinzione tra dimostrabilità e verità impostasi sulla riflessione intorno ai fondamenti della matematica, di cui abbiamo parlato nella prima Parte del secondo capitolo. È chiaro che, al pari di una macchina, neanche la mente possa fornire una prova dell'enunciato gödeliano del sistema formale, a cui quella macchina corrisponde, a partire dagli assiomi e dalle regole d'inferenza del sistema stesso. Il tipo di prova in questione, perciò, non va inteso né in senso formale, né in senso assoluto. In proposito, Lucas afferma:

> Per una macchina, e secondo la tesi meccanicista anche per gli uomini, 'produrre come vero' è semplicemente ed interamente la conseguenza di certe condizioni antecedenti, e perciò può essere rappresentato come l'esser-provato-in-un-sistema-formale di qualche tipo. […] Il meccanicista, considerando gli uomini come qualcosa di inferiore agli uomini, ovvero come macchine, considera il loro concetto di verità di nuovo come qualcosa di inferiore, ossia come dimostrabilità-in-un-dato-sistema. Ma non è plausibile ricostruire la verità come dimostrabilità-in-un-dato-sistema, e il teorema di Gödel mostra questo (o, perlomeno, che il prezzo filosofico da pagare per identificare la verità con la dimostrabilità-in-un-sistema sarebbe molto alto).[122]

Un risultato importante dei teoremi di incompletezza, messo peraltro in evidenza dallo stesso Gödel nella sua Gibbs Lecture, è in effetti proprio quello di aver stabilito che la dimensione delle proposizioni matematiche vere (matematica oggettiva) eccede quella delle formule matematiche dimostrabili (matematica soggettiva). Il ridimensionamento del programma hilbertiano, pertanto, stabilisce l'impossibilità di identificare verità e dimostrabilità. Inoltre, come abbiamo visto, il teorema di Tarski vieta che si possano esprimere le condizioni di verità di un linguaggio prescindendo dall'uso di un metalinguaggio. A partire da questo risultato, in una nota al testo, Lucas avanza un argomento, proposto anche da Popper in maniera analoga. Dal momento che tale teorema stabilisce l'impossibilità di fornire una definizione formale di "verità" a partire dai soli mezzi espressivi del linguaggio-oggetto, «se al meccanicista è precluso dire che il meccanicismo è vero sulla base dei suoi stessi principi» – scrive Lucas – «allora non c'è nulla che egli può dire che possa preoccuparci».[123]

---

[122] «*For a machine, and on the mechanist thesis for men too, 'producing as true' is simply and entirely the consequence of certain antecedent conditions, and therefore can be represented as being proved-in-a-formal-system of some sort. […] The mechanist, regarding men as something less than men, namely machines, regards their concept of truth as again something less, namely provability-in-a-given-system. But it is implausible to reconstrue truth as provability-in-a-given-system, and Gödel's theorem shows this (or, at least, that the philosophical price of identifying truth with provability-in-a-system would be very high)*» (Cfr. *ivi*, p. 148).

[123] «*If the mechanist is by his own principles precluded from saying that mechanism is true, there is nothing he can say that could worry us*» (Cfr. *ibidem*).

Questo è un punto già sottolineato più volte: se il meccanicismo si riduce alla constatazione della non-confutabilità della sua tesi, allora rinuncia anche a difendere la verità di questa. Anche qui è particolarmente evidente il piano dialettico su cui è posta la riflessione di Lucas. Inoltre, è assai importante ricordare che l'assunto di fondo dell'intera posizione di Lucas, e che egli ribadisce anche nella risposta a Benacerraf, consiste nell'accettazione di quell'ottimismo razionalistico che abbiamo già visto costituire il tratto caratteristico della visione di Gödel, e che attribuisce alla mente una capacità creativa idealmente illimitata.

Lucas poi contro obietta alla riformulazione del suo argomento originario proposta da Benacerraf. Secondo quest'ultimo, tale ricostruzione è atta a mostrare che una premessa fondamentale per Lucas consiste nella possibilità di conoscere il programma della macchina di Turing avanzata come modello adeguato della mente, specie se restiamo all'interno del "gioco della gödelizzazione" proposto per la mente da Lucas. Ora, tale premessa non solo non viene concessa, ma non viene neppure ammessa da Benacerraf. Pertanto, come sappiamo, la conclusione di quest'ultimo si riduce ad una duplice constatazione: da un lato, è possibile che, nel caso in cui fossimo equivalenti ad una macchina di Turing, non ci sarebbe dato sapere esattamente di quale macchina si tratti; dall'altro, se anche tale macchina fosse, per così dire, "ispezionabile", non potremmo dimostrarne la consistenza e, di conseguenza, non potremmo "uscire dal sistema" e stabilire la differenza tra noi ed essa. La prima conclusione non è di grande importanza, come abbiamo più volte evidenziato, in quanto implicherebbe la traduzione in termini di teoria della computazione del problema epistemologico per eccellenza: la conoscenza di sé; senza con ciò fornire alcun nuovo elemento a supporto di un'ipotetica soluzione di esso. La seconda, invece, merita un'analisi più attenta. Data una macchina T, a proposito del condizionale C, 'se T è consistente, allora U [che nella notazione di Putnam denota l'enunciato gödeliano di T] è vero', Lucas afferma:

> da dove […] ottengo la premessa per la quale T è consistente? Io rispondo *"dal meccanicista stesso"*. […] Se il meccanicista avanza la pretesa che una particolare macchina […] sia equivalente a me, allora io posso chiedere "Dunque, sostieni che T sia consistente?" ed egli deve concederlo, o altrimenti abbandonare la sua pretesa che T sia equivalente a me. Ma non appena egli garantisca la consistenza di T, io posso produrre la formula gödeliana di T come vera […].[124]

[124] «*Whence* […] *do I obtain the premiss that T is consistent? I answer "*From the mechanist himself*"* […]. *If the mechanist advances the claim that a particular machine* […] *is equivalent to me, then I can say "So you say T is consistent?" and he must concede it, or else abandon his claim that T is equivalent to me. But once he has granted that T is consistent, I can produce the Gödelian formula of T as true* […]» (Cfr. *ivi*, p. 154).

Come sottolinea lo stesso Lucas, la strategia dialettica qui usata è quella della *reductio ad absurdum*. Egli cioè ammette l'ipotesi meccanicista per la quale T costituirebbe un modello adeguato della mente. Stando alla visione da lui abbracciata, tale modello dovrebbe essere consistente, così come consistente è il pensiero dell'uomo. Da tale premessa, Lucas trae dunque una conclusione contraddittoria, ossia la differenza tra T e la mente. È chiaro, tuttavia, che affinché una simile dimostrazione per assurdo abbia successo, occorrono due cose: primo, che la tesi meccanicista abbia un contenuto determinato, ovvero che riguardi una certa specificazione formale quale modello della mente; secondo, che valga l'assunzione di consistenza della ragione umana.

Di seguito riportiamo la formulazione dell'argomentazione per assurdo di Lucas che Putnam ha suggerito a quest'ultimo in una discussione del 1967, dove i termini "Vedo_I" e "Vedo_T" indicano l'operatore che esprime la capacità di "vedere" la verità di una formula, rispettivamente da parte della mente ("Io") e della macchina T:[125]

(1) Se T è consistente, allora T ≠ me
(2) Vedo_I(Se T è consistente, allora U)
(3) Se T è consistente, allora Vedo_I(U)
(4) Se T è consistente, allora ¬Vedo_T(U)
(5) Se T è consistente, allora T ≠ me.

Come abbiamo già visto, Putnam sostiene che, esattamente al pari della mente, anche la macchina T può provare il condizionale C. Ma da ciò non segue logicamente, data soltanto l'ipotesi che T sia consistente, che la mente possa "vedere" la verità di U. A tale scopo sarebbe necessario che la mente "vedesse" (o riconoscesse come evidente) anche la consistenza di T. Tuttavia, – e questo è un punto centrale – la mente potrebbe non riconoscere, o non poter riconoscere, che T è consistente. Del resto, tale riconoscimento può avvenire sulla base di due riferimenti differenti: da un lato, il contenuto d'evidenza; dall'altro, la dimostrazione formale di U in un'estensione conservativa più potente del sistema di partenza. Riguardo al primo, come ammette lo stesso Lucas, il riconoscimento della consistenza di T fa appello ad un contenuto d'evidenza non rigorosamente formalizzabile. Per quanto riguarda, invece, il secondo, si tratta di un «formalismo computazionale capace di crescere su di sé e, dunque, di autocorreggersi».[126] Ciò a cui Galvan si riferisce qui, lo spiegheremo meglio in seguito. Basti per ora ricordare quello che abbiamo detto riguardo la possibilità di costruire metateorie di volta in volta più potenti all'interno delle quali decidere l'enunciato gödeliano di una data teoria,

[125] Cfr. *ibidem*.
[126] Cfr. S. Galvan, *Gödel e il modello computazionale della mente*, cit., p. 152.

e ricordare che tale costruzione può essere svolta meccanicamente ed è ipoteticamente eseguibile da una certa macchina di Turing.

Lucas, perciò, propone una modificazione della formulazione della critica di Putnam, e applica in modo generale l'operatore "Vedo" a tutte le formule precedenti:[127]

(1)' $Vedo_I$(Se T è inconsistente, allora T ≠ me)
(2) $Vedo_I$(Se T è consistente, allora U)
(3)' $Vedo_I$(Se T è consistente, allora $Vedo_I$(U))
(4)' $Vedo_I$(Se T è consistente, allora ¬$Vedo_T$(U))
(5)' $Vedo_I$(Se T è consistente, allora T ≠ me).

A questo punto, egli sostiene che (3)' segue da (2) in modo naturale. Ma l'errore, a nostro avviso, non viene così eliminato, poiché appare immediatamente evidente che il fatto di riconoscere la verità del condizionale C ('se T è consistente, allora U'), non può garantire la capacità di riconoscere che, se T è consistente, allora U è riconoscibilmente vero. In altre parole, riteniamo che l'uso dell'operatore "$Vedo_I$" in (3)' non sia da Lucas adeguatamente giustificato dal punto di vista strettamente logico.

Un'altra questione degna di nota sollevata da Lucas riguarda la specifica prospettiva da cui la critica di Putnam può essere svolta: in prima od in terza persona. Questo è un punto che viene ripreso estesamente anche da Penrose, e perciò riteniamo sia utile illustrarlo. Lucas pone l'attenzione sul soggetto che assume l'ipotesi di consistenza di T. Svolto in terza persona, l'argomento di Putnam è il seguente. Consideriamo un certo soggetto, diciamo *s*, che per la tesi meccanicista è equivalente alla macchina T, ma a cui non è dato sapere se T è consistente o meno. Tuttavia, se ci poniamo dal punto di vista di un osservatore esterno a *s*, allora ci è concesso assumere la consistenza di T, e affermare, come fa Putnam, che, sebbene T "sia" consistente, *s* potrebbe comunque non vedere la verità di U, e quindi la consistenza di T. Secondo Lucas, invece, l'argomento svolto in prima persona sarebbe significativamente differente. Se assumiamo il punto di vista di *s*, ovvero ci identifichiamo con questi, allora dobbiamo limitarci ad un'assunzione più debole di quella precedente, ovvero ad assumere non che T "sia" consistente, ma che T "possa essere" consistente. In base a ciò, allora, possiamo affermare che, sebbene T possa essere consistente, noi potremmo non essere in grado di vedere la verità di U. Come spiega Lucas, infatti, «se non mi è concesso assumere come vero che T sia consistente, allora non posso assumere che lo *sia* – anche se potrei assumere che potrebbe esserlo». Perciò, «*se* sto facendo l'assunzione categorica espressa dalla clausola 'sebbene T *sia* consistente'» – continua Lucas

[127] Cfr. J.R. Lucas, *Satan Stultified: A Rejoinder to Paul Benacerraf*, cit., p. 155.

– «*allora* io *posso* vedere che U è vero».[128] Tuttavia, analogamente a prima, anche in questo caso riteniamo che la risposta di Lucas non riesca davvero a dimostrarsi valida. Ricordiamo che l'obiezione di Putnam ruota intorno al fatto che, anche nel caso che T sia consistente, esiste la possibilità che la mente non sia capace di riconoscerne la consistenza, e quindi di vedere la verità dell'enunciato indecidibile U in T. Ciò significa che per vedere la verità di U, alla mente non è sufficiente soltanto assumere che T sia consistente, bensì anche vedere la consistenza di T. Questo è il punto che la risposta di Lucas a Putnam non tiene debitamente in considerazione, e su cui è imperniata anche l'analisi di Shapiro.

Ora, il nodo centrale di tutta la presente questione consiste proprio nel postulato della consistenza della mente umana, cioè della sua infallibilità in linea di principio. Naturalmente, questa è un'assunzione non immediatamente verificabile, e poggia su di una convinzione che esula dall'ambito della discussione di stretta pertinenza della logica formale, e che è stata pertanto oggetto di critiche da parte di diversi autori, i quali sostengono che non è possibile escludere a priori l'identificazione della mente con un algoritmo "non riconoscibilmente" consistente. Del resto, come aveva già rilevato Gödel, sono proprio i teoremi di limitazione a impedire di dimostrare formalmente la consistenza dell'insieme della matematica soggettiva. In ogni caso, l'argomento di Lucas, come anche quello di Penrose, risulta vincente soltanto a patto di dare per scontato tale postulato. È su di esso, infatti, che si gioca, per così dire, il vero scontro tra meccanicismo ed anti-meccanicismo.

Da quanto detto, possiamo perciò affermare che Lucas sembra derivare dai teoremi gödeliani più di quanto la loro effettiva formulazione non consenta di fare. Ciò viene riconosciuto dallo stesso Lucas nella conclusione del suo saggio, in cui afferma:

> Non potrei credere di essere una macchina di Turing inconsistente senza abbandonare la mia convinzione che esiste una differenza tra verità e falsità […]. Si tratta, nel linguaggio della teologia matematica, di un atto di fede. Nessuna dimostrazione formale è possibile: ma l'alternativa mortifica se stessa. Sebbene la possibilità del mio essere essenzialmente inconsistente non possa essere esclusa da argomenti formali, ed è forse almeno concepibile, è una posizione di completo nichilismo matematico […].[129]

---

[128] «*If I am not entitled to assume the T is consistent, I must not assume that it* is *– though I may assume that it may be.* […] *for* if *I am making the categorical assumption expressed by the clause 'Although T* is *consistent',* then *I* can *see that U is true*» (Cfr. *ivi*, p. 156).

[129] «*I could not believe I was an inconsistent Turing machine without abandoning my conviction that there was a difference between truth and falsehood* […]. *It is, in the language of mathematical theology, an act of faith. No formal proof is possible: but the alternative is self-stultifying. Although the possibility of my being essentially inconsistent cannot be ruled out by formal arguments and is perhaps just conceivable, it is a position of complete mathematical nihilism*» (Cfr. *ivi*, p. 158).

Inoltre, come abbiamo visto, nel suo testo egli si limita a sostenere la necessità, implicata dal teorema tarskiano, di tenere distinti i concetti di verità e di dimostrabilità formale. Noi riteniamo, invece, che alle sue considerazioni vada aggiunta una precisazione fondamentale, che si rifà a quanto abbiamo detto nel capitolo terzo. In quel luogo dicevamo che ogni modello interpretativo di una teoria formalizzata possiede un contenuto semantico astratto, che fornisce le condizioni di verità delle proposizioni della teoria, ma che non è però possibile esplicitare in termini finitari all'interno della teoria formale stessa. Ora, Galvan evidenzia che la nozione di verità non può coincidere, a meno di cadere in un *regressus ad infinitum*, con quella di dimostrabilità in una metateoria di volta in volta più potente rispetto a quella immediatamente precedente. La costruzione di questa serie di metateorie, dunque, non può che presupporre quello che egli definisce un orizzonte dell'«evidenza veritativa»: tale orizzonte va inteso come una totalità della struttura semantica, che fa da condizione di possibilità della dimensione linguistica formale in cui le varie teorie vengono espresse. Quella successione, invece, se considerata semplicemente in se stessa, darebbe luogo ad un'infinità nel senso cattivo del termine, ovvero ad un'infinità "potenziale". Come vedremo nel capitolo ottavo, Pagani sostiene che affinché quest'ultima si dia è necessario postulare un altro tipo di infinità che la contenga e che ne sia condizione di possibilità: l'infinità "trascendentale", intesa come una sorta di "apertura" comprensiva di ogni determinazione della nostra esperienza, all'interno della quale soltanto, cioè, può costituirsi ogni attività del pensiero.

6.4 L'argomento nuovo di Penrose

La tesi principale dell'opera del fisico Roger Penrose, intitolata *Shadows of the Mind*,[130] è che «la comprensione (*understanding*) *matematica* è qualcosa che non può essere ridotto alla computazione».[131] Come Lucas, anche Penrose ritiene che i teoremi di incompletezza di Gödel forniscano una prova della natura non-computazionale del ragionamento matematico, e quindi della differenza essenziale tra mente umana ed intelligenza artificiale. Egli si spinge sino ad affermare che l'intero nostro modo di comprendere la realtà si basa su processi non-computazionali. Tuttavia, mentre la natura della nostra comprensione sarebbe difficilmente determinabile relativamente ad esperienze di tipo comune, nel caso invece della nostra capacità di cogliere il carattere infinito dell'insieme dei numeri naturali, spiega Penrose, ciò sarebbe possibile. Fin dalle prime battute del

[130] Cfr. R. Penrose, *Shadows of the Mind*, cit. Nelle successive citazioni tratte da quest'opera, la traduzione dall'inglese è ad opera nostra ed ogni corsivo è dell'autore.

[131] «Mathematical *understanding is something that cannot be reduced to computation*» (Cfr. *ivi*, p. 48).

testo, Penrose si professa platonista matematico, e ritiene che la peculiare capacità di visualizzazione della mente non sia catturabile nei termini di un sistema formale o di una macchina di Turing. A questo proposito egli afferma:

> Alcune relazioni matematiche richiedono lunghe catene di ragionamento prima di essere percepite con certezza. Ma lo scopo dalla dimostrazione matematica è, in effetti, quello di fornire quelle catene […] dove ogni *passo* è appunto qualcosa che può essere percepito come ovvio. Di conseguenza, la fine del ragionamento è qualcosa che deve essere accettato come *vero*, nonostante possa non essere affatto ovvio in sé. Uno potrebbe immaginare che sia possibile produrre una lista di tutti i possibili passi 'ovvi' del ragionamento una volta per tutte, di modo che in seguito tutto possa essere ridotto a computazione […]. Ciò che l'argomento di Gödel mostra […] è che questo non è possibile. Non c'è alcun modo per eliminare il bisogno di *nuove* 'ovvie' concezioni.[132]

Come abbiamo già detto nel capitolo quarto, certe computazioni possono terminare o meno, nel senso che una macchina istruita per eseguire tali operazioni può arrestarsi raggiunto un certo risultato (stato finale), oppure continuare indefinitamente a cercare la soluzione richiesta. A questo proposito, Penrose propone alcuni interessanti esempi per mostrare come certi metodi dimostrativi in matematica siano corretti, sebbene non siano propriamente formalizzati sotto forma di operazioni logiche computabili.[133] E sebbene alcuni di questi metodi siano formalizzabili mediante il principio di induzione, secondo l'autore, i teoremi gödeliani dimostrerebbero che tale principio, come del resto un qualunque sistema formale traducibile in termini computazionali, non è sempre sufficiente a determinare se una computazione termina o meno.

Penrose avanza due argomenti contro la tesi meccanicista. Il primo riprende, a grandi linee, quello originario del testo di Lucas del 1961, e la conclusione che da esso deriva non è necessariamente inconsistente con la tesi meccanicistica. Il secondo, comunemente detto il "nuovo argomento", è atto a rafforzare il primo (superandone alcuni aspetti problematici già rilevati) e a mostrare l'implausibilità del meccanicismo, dimostrando che, se per assurdo esistesse un insieme ricorsivamente enumerabile di tutte le procedure inferenziali adoperate da matematici umani idealizzati riunite in un sistema assiomatizzato, tale insieme non potrebbe essere fermamente creduto

---

[132] «*Some mathematical relations require long chains of reasoning before they can be perceived with certainty. But the object of mathematical proof is, in effect, to provide such chains* […] *where each* step *is indeed something that can be perceived as obvious. Consequently, the endpoint of the reasoning is something that must be accepted as* true, *even though it may not, in itself, be at all obvious. One might imagine that it would be possible to list all possible 'obvious' steps of reasoning once and for all, so that from then on everything could be reduced to computation* […]. *What Gödel's argument shows* […] *is that this is not possible. There is no way of eliminating the need for* new *'obvious' understandings*» (Cfr. *ivi*, p. 56).

[133] Cfr. *ivi*, pp. 68-72.

(o riconosciuto) come tale. Dal momento che entrambi gli argomenti (specialmente il secondo) sono assai estesi, per non appesantire l'esposizione riteniamo opportuno fornirne una sintesi che sia in grado di mettere in risalto le loro implicazioni più importanti.

Il primo argomento è riassumibile come segue.[134] Assumiamo una procedura algoritmica A, eseguibile da una macchina di Turing, tale che A termini in un numero finito di passi qualora ottenga la dimostrazione dell'enunciato E 'la computazione $C(n)$ su un certo numero naturale n non termina'.[135] Assumiamo, inoltre, che A rappresenti l'insieme di «*tutte* le procedure disponibili ai matematici per dimostrare in maniera convincente che certe computazioni non terminano».[136] Secondo la visione di Penrose, ciò equivale alla clausola di conoscibilità e di correttezza (*soundness*) di A. In altre parole, è richiesto che A termini soltanto qualora abbia dimostrato come vero l'enunciato E, ovvero soltanto qualora dimostri che la computazione $C(n)$ effettivamente non termina. In generale, $C(n)$ denota la collezione ordinata di tutte le possibili computazioni su di un unico numero naturale n, ed è rappresentata dalla successione $C_0(n)$, $C_1(n)$, $C_2(n)$, …, $C_q(n)$, …. La procedura A, pertanto, è esprimibile nella forma di una funzione a due argomenti $A(q, n)$ (dove q è un numero naturale ed indica la q-esima computazione appartenente alla collezione C), per la quale, dati q e n, se $A(q, n)$ termina, allora $C_q(n)$ non termina, mentre se A non termina, allora $C_q(n)$ termina. Ora, per ipotesi possiamo affermare quindi che:

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

Se poniamo $q = n$, otteniamo:

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

Notiamo che $A(n, n)$ opera su un unico numero, ovvero n, e deve, perciò, essere una tra le diverse computazioni su n che formano la collezione ordinata $C(n)$. Poniamo che $A(n, n)$ sia la k-esima computazione su n, e quindi che $n = k$. Da ciò si ottiene:

(3) $A(k, k) = C_k(k)$,

e da (2), con $n = k$:

---

[134] Cfr. *ivi*, pp. 72-75.

[135] Come sottolinea Penrose, tale enunciato è una Π_1-formula.

[136] «All *the procedures available to human mathematicians for convincingly demonstrating that computations do not stop*» (Cfr. *ivi*, p. 73).

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

Sostituendo (3) in (4) si ottiene:

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

Da ciò risulta "evidente" che $C_k(k)$ non può terminare. Inoltre, dall'equivalenza $C_k(k) = A(k, k)$ risulta anche che A non è in grado di stabilire in un numero finito di passi se $C_k(k)$ termina oppure no. Pertanto, l'ipotesi di partenza, per la quale A equivale ad un sistema formale rappresentante l'insieme di tutte le procedure matematiche conoscibili e corrette, è falsa, perché, come abbiamo appena visto, noi siamo in grado di stabilire che $C_k(k)$ non termina. In altre parole, «*A* non è capace di incapsulare la nostra comprensione […] *a patto che sappiamo che è corretta*».[137] Posto in questi termini, notiamo che l'argomento poggia sulla medesima assunzione che sta al fondo di quello avanzato da Lucas: la convinzione per cui il nostro pensiero è, in linea di principio, non soltanto consistente e corretto, ovvero in grado di dimostrare con assoluta certezza solo verità matematiche, ma anche completo, ovvero tale da poter conoscere "tutte" le verità matematiche. La conclusione che Penrose trae è la seguente:

(*G*) I matematici non usano un algoritmo riconoscibilmente corretto (*a knowably sound algorithm*) per dimostrare verità matematiche.[138]

Precisiamo, qui, che "riconoscibilmente corretto" implica due cose: primo, la conoscibilità in linea di principio dell'algoritmo *A* da parte di una mente ideale; secondo, la capacità da parte quest'ultima di "vedere" che l'algoritmo in questione è corretto. Anche in questo caso, la discussione in merito alla natura di questo "vedere" intuitivo necessita di considerazioni ulteriori che affronteremo sia analizzando la critica di Shapiro all'ottimismo razionalistico gödeliano sia illustrando la posizione di Galvan in merito alla logica dell'evidenza. Va notato, inoltre, che *G* è coerente con il dilemma posto da Gödel; esso non rappresenta cioè una reale confutazione del meccanicismo, poiché lascia aperta l'ipotesi– ammessa anche da Putnam e da Benacerraf – che la nostra mente funzioni secondo un algoritmo non conoscibile in linea di principio né riconoscibile come corretto.

[137] «A cannot *encapsulate our understanding* […] so long as we know it to be sound» (Cfr. *ivi*, p. 75).
[138] «*Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth*» (Cfr. *ivi*, p. 76).

Prima di passare all'esposizione del secondo argomento, Penrose risponde a diverse possibili obiezioni poste al primo (alcune delle quali analizzeremo più avanti). Una di queste riguarda la vaghezza della nozione di verità applicata alle proposizioni matematiche. Abbiamo già visto come il progetto formalista abbia rappresentato, relativamente alla questione dei fondamenti della logica e della matematica, il culmine di una corrente di pensiero che, rifiutando la sostenibilità di una definizione definitiva e assoluta di verità assiomatica, ha posto la coerenza sintattica come criterio fondante la validità (o verità) degli assiomi di una teoria formale. A questo proposito, Penrose ritiene che il problema dell'appello alla nozione di verità in senso lato, implicata dall'assunzione della coerenza semantica di un sistema formale F (che si suppone equivalente a una macchina di Turing in grado di eseguire la procedura algoritmica *A*), sia superabile restringendo il campo alle sole Π_1-formule di F.[139] Nel capitolo secondo della Parte II, abbiamo detto che le Π_1-formule sono formule ricorsive primitive chiuse (mediante l'operatore universale), e che l'insieme delle Π_1-formule vere di LPR costituisce l'aritmetica reale di Hilbert. Quest'ultima, perciò, come dimostrano i teoremi di limitazione, non è propriamente formalizzabile, in quanto implica un riferimento al criterio infinitario della verità. La definizione di Π_1-formula fornita da Penrose è equivalente a quella da noi usata, sebbene sia espressa in termini computazionali come un enunciato che asserisce che una certa computazione eseguibile da macchina di Turing non termina. Si noti, in proposito, che l'enunciato gödeliano di un sistema formale equivale precisamente ad una Π_1-formula, in quanto indecidibile in un numero finito di passi sulla base delle regole di tale sistema. Ora, Penrose sostiene che la questione della verità in senso assoluto delle Π_1-formule di F non sia problematica, se non per alcune questioni di tipo pratico, le quali però esulano dall'argomentazione in linea di principio da egli avanzata, riguardante l'idealizzazione sia della macchina di Turing sia della mente.[140] In altre parole, egli ritiene che, una volta definita una Π_1-formula, non sia davvero ragionevole pensare che sia impossibile da parte della comunità matematica raggiungere un accordo comune e universale intorno alla verità di essa. Ancora una volta, va evidenziato che una simile convinzione dipende dalla particolare posizione

[139] Penrose propone, perciò, una nuova formulazione della conclusione *G*, ovvero *G***: «I matematici umani non usano un algoritmo riconoscibilmente corretto al fine di accertare la verità delle Π_1-formule» («*Human mathematicians are not using a knowably sound algorithm in order to ascertain the truth of Π_1-sentences*») (Cfr. *ivi*, p. 100).

[140] «L'unico ragionevole problema riguardo la natura assoluta della verità di tali enunciati [Π_1-formule] potrebbe essere che alcune computazioni che terminano possono richiedere così tanto tempo da non poter essere concepibilmente completabili nella pratica […]. Inoltre, il concetto (assai limitato) di verità mathematica di cui necessito […] è non meno ben definito dei concetti di *vero*, *falso* e *indecidibile*, per ogni sistema formale *F*» («T*he only reasonable issue with regard to the absolute nature of the truth of such statements might be that some terminating computations can take so inordinately long that they might not conceivably be completed in practice* […]. *As a further point, the (very limited) concept of mathematical truth that I need […] is actually no less well defined that the concepts of* true, false, *and* undecidable, *for any formal system* F») (Cfr. *ivi*, p. 96).

platonista in ambito matematico che Penrose esplicitamente assume, oltre che da un forte ottimismo – da lui condiviso con Gödel e Lucas – circa le capacità mentali dell'uomo.

Il secondo argomento viene sviluppato da Penrose al fine di mostrare l'implausibilità dell'ipotesi sopra menzionata, per la quale potrebbe esistere un algoritmo assolutamente inconoscibile da parte dell'uomo come modello adeguato del funzionamento della mente. Si noti che è lo stesso Penrose, in un certo senso, ad ammettere che il suo argomento è lungi dal confutare in maniera logicamente conclusiva la tesi meccanicistica dell'IA forte.

Egli ammette le seguenti quattro possibilità relativamente al sistema formale F:

a) che F sia conoscibile e riconoscibile come corretto, ma che in realtà non sia tale;
b) che F sia conoscibile e corretto, ma non riconoscibile come tale;
c) che F sia conoscibile ma non corretto né riconoscibile come tale;
d) che F non sia conoscibile né riconoscibilmente corretto.

Innanzitutto, è bene sottolineare che è lecito assumere che la conoscibilità di F (da intendersi in linea di principio) equivalga al suo essere già conosciuto. Per quanto riguarda l'ipotesi (a), se si assume che F sia conoscibile e riconoscibile come corretto, allora dalla conclusione *G* ("noi non dimostriamo teoremi matematici sulla base di una procedura algoritmica riconoscibilmente corretta") risulta che F non può essere corretto, dal momento che *G* stabilisce che F non può essere riconoscibilmente corretto. Ma ciò, spiega Penrose, è irrilevante, poiché se, per ipotesi, F è "creduto fermamente" come corretto, allora, dal secondo teorema di incompletezza, siamo in grado di comprendere che l'enunciato gödeliano di F è vero, anche se non è derivabile in F, contrariamente all'assunto iniziale per cui il ragionamento di un matematico ideale rispetta le regole di F. L'ipotesi (a) perciò va rigettata.

L'ipotesi (b), secondo l'autore, è logicamente possibile, per quanto sia assai implausibile. Questa coincide con l'ipotesi (ammessa anche da Gödel) secondo la quale è ammissibile l'esistenza di una regola che dimostri tutte e sole le proposizioni della matematica soggettiva, sebbene essa non possa essere conoscibile. Per ipotesi, F è analizzabile come insieme finito di assiomi e regole ma non è riconoscibile come corretto. Per questo motivo, il nostro matematico ideale potrebbe giungere a riconoscere la correttezza soltanto di tutti i suoi assiomi ma non anche di tutte le sue regole d'inferenza. In altri termini, dovrebbe esistere almeno una regola inferenziale di F non riconoscibilmente corretta dalla mente. In caso contrario, ciò implicherebbe il riconoscimento della correttezza di F nel suo insieme e quindi della verità dell'enunciato gödeliano indimostrabile in F. Perciò: o F non è corretto (il che rimanda all'ipotesi (c)), oppure è corretto ma almeno una delle sue

regole non è riconoscibile come tale. Secondo Penrose, non è possibile escludere la concepibilità di un sistema formale quale F, sebbene, tuttavia, sembra alquanto implausibile l'esistenza di una tale regola «che sembra miracolosamente in grado di generare precisamente [tutti gli assiomi di F] incontrovertibilmente percepibili dall'uomo».[141]

Anche l'ipotesi (c), non è escludibile a priori, per quanto sia, secondo Penrose, inaccettabile qualora si volesse continuare ad attribuire validità al sapere matematico e scientifico in generale. A questo proposito, ci limitiamo ad osservare quanto abbiamo già detto anche in merito all'ottimismo razionalistico gödeliano. Anche Penrose crede nell'emendabilità in linea di principio di ogni errore logico sulla base di un criterio di verità incontrovertibile come ideale normativo del sapere umano.

L'ultima ipotesi (d) è riconducibile, in ultima istanza, alla prima. Penrose, infatti, dedica diversi paragrafi alla spiegazione del fatto che un sistema formale che comprendesse l'insieme di tutte le proposizioni aritmetiche dimostrabili dall'essere umano non potrebbe davvero essere inconoscibile in linea di principio, bensì soltanto da un punto di vista pratico. La non-specificabilità di un tale sistema dipenderebbe da un ordine di complessità tale da eccedere le capacità di comprensione di una mente umana soltanto nel caso in cui questa venisse supposta come finita.

Ora, argomenta Penrose, i sostenitori della tesi meccanicista sono disposti ad ammettere l'esistenza di un algoritmo talmente complesso da non poter essere, non soltanto riconosciuto come modello adeguato del funzionamento della mente umana, ma nemmeno specificabile dall'uomo nelle sue componenti di tipo strutturale e logico. Un simile programma, stando alla prospettiva meccanicista, non potrebbe essere realizzato grazie alle scoperte della comunità scientifica "attuale", ma si costituirebbe "emergendo" in maniera graduale. Dapprima, a partire da macchine connessioniste, le quali, mediante certe procedure algoritmiche codificabili direttamente dall'uomo nel loro linguaggio-macchina (processi *top-down*) insieme a processi di apprendimento (processi *bottom-up*), imparerebbero da sé a realizzare macchine di nuova generazione. Queste, a loro volta, creerebbero nuove classi di macchine di potenza sempre maggiore, e così via. Alla fine, in un certo tempo nel futuro, si raggiungerebbe così una progenie robotica talmente avanzata che la complessità del programma che ne sta alla base sarebbe tale da oltrepassare le capacità di comprensione della mente. Penrose rigetta fermamente questa ipotesi, in quanto dimostra che, se assumiamo che il cominciamento di questo sviluppo dipende dall'azione dell'uomo e dall'implementazioni di algoritmi "sorgente" da esso conoscibili, allora, seguendo un ragionamento "per gradi", non possiamo prospettare una situazione in cui la mente umana non sia più in grado di risalire alle cause che determinano per intero il funzionamento di un simile programma. Dunque, si può mostrare che, se si

[141] «*Which seems miraculously to be able to generate precisely [all the axioms of S] that can be unassailably humanly perceived*» (Cfr. *ivi*, p. 136).

assume la congettura meccanicista appena esposta, allora l'ipotesi (d) viene a coincidere, in ultima analisi, ad (a), a (b) oppure a c). Il ragionamento è sintetizzabile come segue.

Innanzitutto, Penrose esclude che la visione dell'IA possa sostenere la necessità di «intervento divino» per la realizzazione di un simile programma. Semmai, più coerentemente, quest'ultimo potrebbe venire inteso come il risultato di pressioni selettive e condizionamenti genetici che hanno condotto allo sviluppo dell'intelligenza e della comprensione nella specie umana. Ma nemmeno questa possibilità sembra convincente. Infatti, in origine, questo programma avrebbe dovuto contenere in sé, "in potenza", tutte le capacità deduttive del pensiero umano acquisite nel corso dell'evoluzione. Ma, se assumiamo il criterio dell'utilità selettiva come spiegazione della validità dei processi inferenziali, non risulta affatto evidente il motivo per il quale questi ultimi avrebbero dovuto originarsi, od essere prodotti, a partire dal meccanismo della selezione naturale.

In secondo luogo, l'apprendimento di una rete neurale (*machine learning*) si basa su due tipologie di fattori: esterni ed interni. I fattori esterni comprendono gli stimoli e le risposte provenienti dall'ambiente esterno (caso naturale) o da un insegnante (*trainer* o *teacher*). Nel primo caso, stando alle conoscenze attuali, sembrerebbe non esservi alcunché che faccia parte dell'ambiente fisico circostante e che non sia potenzialmente simulabile in termini computazionali (e di certo ciò è accettato dalla tesi dell'intelligenza artificiale sia forte che debole).[142] Per quanto riguarda, invece, i *feedback* provenienti dall'insegnante, essi sono codificabili nel linguaggio della macchina e consentono a quest'ultima di modificare la propria struttura interna per migliorare la propria prestazione, cioè le soluzioni ai problemi che le si presentano. Anche in questo caso, i processi *bottom-up* attraverso cui ciò avviene rispettano delle regole interamente algoritmiche. Inoltre, Penrose spiega che è sempre possibile tradurre i processi paralleli di apprendimento di una macchina connessionista nei termini di processi seriali di una macchina di Turing. Da quanto appena detto, risulta pertanto che tutti gli elementi che compongono una simile macchina (o *robot*) dovrebbero essere di natura computazionale, e quindi conoscibili in linea di principio dall'uomo, contrariamente a quanto viene assunto nell'ipotesi d).

A questo punto, possiamo illustrare il nucleo del secondo argomento di Penrose, il quale, alla pari di quello di Lucas, è mirato alla confutazione della tesi meccanicista attraverso la sua riduzione all'assurdo. Se un robot, dotato di processi *top-down* e *bottom-up*, venisse proposto come modello capace di eguagliare (se non addirittura superare) le capacità di comprensione della mente umana, allora esso dovrebbe anche possedere, proprio come noi, un criterio, diciamo "◊", di verità matematica inattaccabile (*unassailable*). Non è necessario che tale criterio corrisponda ad un concetto

[142] Penrose sostiene, infatti, che anche quei sistemi fisici definiti "caotici", per quanto complessi ed effettivamente non predicibili, siano pur sempre deterministici ed analizzabili algoritmicamente.

semantico; è sufficiente che esso sia traducibile "funzionalmente", ovvero in termini di un comportamento esteriore rispetto ad esso coerente se analizzato dal nostro punto di vista di esseri umani. Denotiamo quindi le asserzioni che sottostanno a questo criterio come ◊-asserzioni. Ora, potendo noi conoscere quei meccanismi **M** che sommati costituiscono il programma del robot, saremmo perciò in grado di definire un sistema formale *Q* tale che i suoi assiomi e teoremi comprendano tutte e sole le ◊-asserzioni ritenute vere dal robot, oltre a quelle formule ottenute da quest'ultime mediante opportune regole d'inferenza.

È qui che diviene evidente quella che Penrose chiama una «contraddizione di base». In primo luogo, notiamo che, se assumiamo il punto di vista del robot, *Q* coincide con il sistema formale *A* precedentemente considerato rispettivamente al "nostro" punto di vista. In secondo luogo, per il secondo teorema di incompletezza, il robot non può, a meno di cadere in contraddizione, «credere fermamente che il sistema formale [*Q*] *sia* equivalente alla sua stessa nozione di credenza matematica inattaccabile. Questo nonostante il fatto che *noi* […] potremmo ben sapere che i meccanismi **M** stanno *davvero* alla base del sistema di credenze matematiche del robot, e dunque che il suo sistema di credenze inattaccabile *è* equivalente a [*Q*]».[143] Infatti, se il robot potesse giungere a riconoscere la correttezza di *Q*, allora dovrebbe riconoscere anche la verità dell'enunciato gödeliano di *Q*, insieme al fatto che tale enunciato non sarebbe invece derivabile in *Q*. Di conseguenza, il robot non può fermamente credere di equivalere a *Q*.

Ora, come giustamente fa notare Penrose, un simile argomento va esaminato non dal nostro punto di vista (ovvero di coloro che conoscono sia l'insieme **M** che il sistema *Q*), bensì da quello del robot, il quale, per ipotesi, non conosce l'insieme dei propri meccanismi **M** né può riconoscere come corretto l'insieme *Q* delle proprie credenze. Se ipotizziamo che il robot sia disposto ad accettare la tesi per cui esso sarebbe costituito da meccanismi e processi computazionali, e se ipotizziamo che ad esso vengano presentati (o specificati) **M** e *Q*, conclude Penrose, esso sarebbe costretto «rigettare la possibilità che **M** stia alla base della sua comprensione matematica – *a prescindere* che lo sia davvero o meno».[144] Vediamo perché:

Si assuma, dunque, l'ipotesi *M* che asserisce che l'insieme **M** dei meccanismi del robot determina le capacità matematiche di quest'ultimo. A partire da *M*, il robot può derivare degli enunciati che ritiene veri sulla base del proprio criterio ◊. Tra questi enunciati (teoremi di *Q*) dovrebbe

[143] «[*The robot is incapable of*] *firmly believing that the formal system* [*Q*] is *equivalent to its own notion of unassailable mathematical belief. This is despite the fact that* we […] *might well know that the mechanisms* **M** do *underlie the robot's mathematical belief system, and hence that its unassailable belief system* is *equivalent to* [*Q*]» (Cfr. *ivi*, p. 163).

[144] «[*The robot*] *would have to reject the possibility that* **M** *actually could underlie its mathematical understanding* – irrespective *of whether or not it actually does*» (Cfr. *ivi*, p. 164).

comparire quello che asserisce la correttezza del sistema *Q* e, di conseguenza, anche l'enunciato gödeliano di *Q*. Ancora una volta, è fondamentale qui notare che l'assunzione della correttezza di *Q* sulla base dell'ipotesi *M* ha senso soltanto a patto che il robot presupponga la correttezza del proprio sistema di conoscenze matematiche.[145] A questo punto, è possibile costruire il sistema formale Q_1, i cui assiomi siano le ◊-asserzioni che il robot deriva da *M*. Il nostro robot, sempre sulla base dell'assunto della propria correttezza, sarebbe in grado di riconoscere, o di credere fermamente, nella correttezza di Q_1. Posto di fronte all'enunciato gödeliano di Q_1, perciò, il robot sarebbe certo anche della verità di esso. Ma questo, stando all'ipotesi iniziale, implicherebbe anche che tale enunciato fosse un teorema di Q_1, il che come sappiamo è vietato dai teoremi di limitazione. Come evidenza lo stesso Penrose, è precisamente il fatto che il robot abbia accesso alla specificazione del proprio algoritmo che determina la contraddizione appena vista.[146] La conclusione, pertanto, è la seguente:

> Nessun essere cosciente […] capace di autentica comprensione matematica […] può operare secondo un qualunque insieme di meccanismi che sia in grado di apprezzare, indipendentemente dal fatto che esso *conosca* per certo che quei meccanismi siano quelli che si suppone governino i suoi metodi [*its own routes*] per giungere ad una verità matematica inattaccabile [*unassailable*].[147]

L'argomento che abbiamo appena esposto, secondo Penrose, dovrebbe servire a rigettare la tesi dell'IA, non solo nella sua versione forte ma anche in quella debole. Infatti, egli ritiene che esso mostri come noi disponiamo di un certo tipo di intuizione (evidente specialmente in relazione alle implicazioni che siamo in grado di trarre dai teoremi di limitazione) che non sarebbe ultimamente "ripresentabile", o simulabile, computazionalmente.

Galvan rileva che gli argomenti di Lucas e Penrose hanno un importante aspetto in comune: essi non mirano a dimostrare che la mente è superiore alla macchina in quanto capace di derivare "di più" rispetto a questa, bensì vogliono mettere in luce che la differenza tra le due risiede nella capacità della mente di avere accesso a contenuti, preclusi alla macchina, attraverso una particolare forma di evidenza, o intuizione, di natura extra-logica. Uno di questi contenuti, sottolinea Galvan, è la credenza

---

[145] Questo punto costituisce il nodo centrale della critica mossa da Shapiro all'argomento di Penrose.

[146] Al nostro robot rimangono altre due possibili scappatoie dalla contraddizione. Penrose, tuttavia, previa loro analisi ne trae la medesima conclusione. Un aspetto importante riguarda la natura delle asserzioni dimostrabili come vere dal robot e la possibilità che esse possano essere scorrette e correggibili di principio. Tali asserzioni sono nella forma di Π_1-enunciati. Tuttavia, i dettagli dell'argomentazione a questo proposito non sono necessari alla nostra trattazione. Per un loro approfondimento rimandiamo ai paragrafi §§ 3-17-3.20 dell'opera (Cfr. *ivi*, pp. 167-176).

[147] «*No mathematical aware conscious being* […] *can operate according to any set of mechanisms that it is able to appreciate, irrespective of whether it actually* knows *that those mechanisms are supposed to be the ones governing its own routes to unassailable mathematical truth*» (Cfr. *ivi*, p. 167).

propria della mente umana di ragionare in maniera consistente e corretta: tale convinzione, come vedremo, è alquanto problematica se intesa come elemento a supporto di una confutazione logico-formale della tesi meccanicistica.

Da ultimo, è opportuno esporre brevemente alcuni aspetti-chiave della visione di Penrose circa il problema mente-corpo. Egli crede che la realtà contenga elementi di natura non-computazionale, rispetto ai quali, cioè, non è possibile realizzare una simulazione in termini algoritmici né fornire una spiegazione dal punto di vista della fisica classica. L'esistenza di tali elementi consentirebbe di rendere conto del fenomeno della coscienza, il quale appare immediatamente come al di fuori della chiusura causale del mondo fisico. Una simile caratteristica di non-computabilità, congettura l'autore, sarebbe rinvenibile nella struttura del citoscheletro delle cellule neuronali e, in particolare, nel funzionamento a livello quantistico dei "microtubuli" responsabili dei collegamenti sinaptici. Riprendendo le parole di Searle, «la descrizione al livello dei neuroni potrebbe essere "una semplice *ombra*" del livello più profondo, dove dobbiamo cercare le basi fisiche della *mente*».[148] In sintesi, la prima parte della corposa opera di Penrose è centrata sulla dimostrazione dell'impossibilità di elaborare un modello adeguato e completo della mente in termini computazionali; mentre la seconda parte è interamente dedicata all'esposizione dettagliata della congettura per la quale la coscienza emergerebbe dall'interazione tra il macro-livello della fisica classica e il micro-livello della meccanica quantistica. Per studiare e comprendere tale interazione, naturalmente, è necessaria l'elaborazione di una teoria fisica alternativa (ma non radicalmente diversa) rispetto a quella odierna, che consenta di gettare luce su molti elementi ancora oscuri inerenti al rapporto tra la mente e la realtà fisica.

6.5 La critica di Shapiro e Lindström al nuovo argomento di Penrose

Anche il secondo argomento avanzato da Penrose è stato oggetto di numerose critiche. In un articolo pubblicato sulla rivista *Psyche*, dal titolo *Beyond the Doubting of a Shadow*,[149] egli ribatte in maniera precisa a diverse obiezioni e riformula il suo argomento in una nuova veste, in cui viene reso ancora più esplicito il fatto che, a suo avviso, per rigettare il meccanicismo non sarebbe necessario da parte di un matematico ideale assumere di poter riconoscere la consistenza del sistema formale F

[148] Cfr. J.R. Searle, *Il mistero della coscienza*, trad. it. di E. Carli, Raffaello Cortina Editore, Milano 1998, p. 66; tit. orig. *The Mistery of Consciousness*, New York Review of Books, New York 1997.

[149] Cfr. R. Penrose, *Beyond the Doubting of a Shadow*. *A Reply to Commentaries on Shadows of the Mind*, in «Psyche», II, 1996.

(che si suppone contenere – lo ricordiamo – l'intero delle procedure dimostrative matematiche accessibili all'uomo). Di seguito riportiamo la riformulazione di Penrose:

> (A) Nonostante io non sappia se sono necessariamente F, posso concludere che se lo fossi, il sistema F dovrebbe essere corretto e, in aggiunta, F' sarebbe corretto, dove F' equivale ad F con a supplemento l'ulteriore asserzione "Io sono F". Percepisco che dall'assunzione che sono F segue che l'enunciato gödeliano G(F') deve essere vero e, inoltre, che non è una conseguenza di F'. Ma ho appena percepito che "nel caso io fossi F, allora G(F') dovrebbe essere vero", e le percezioni di una tale natura sono proprio ciò che si suppone che F dimostri. Dal momento, quindi, che io sono capace di percepire qualcosa oltre i poteri di F, deduco che, dopotutto, non posso essere F.[150]

Nel saggio *Penrose's New Argument*,[151] Per Lindström, ha fornito una ricostruzione precisa del nuovo argomento di Penrose e ne ha mostrato l'inconclusività dal punto di vista formale. Sia Sd(F) l'espressione con cui si indica correttezza (*soundness*) di F, e HC(F) l'espressione con cui si indica che F è "umanamente completo" (*humanly complete*), nel senso illustrato poco sopra. Dunque, Lindström ipotizza che l'affermazione "Io sono F" equivalga alla congiunzione $Sd(F) \wedge HC(F)$. Inoltre, sia $F^+ = F + Sd(F)$. Consideriamo l'argomento seguente ((B) nella notazione di Lindström):[152]

(1) $Sd(F) \Rightarrow Sd(F^+)$,
(2) $Sd(F^+) \Rightarrow G(F^+)$,
(3) $Sd(F^+) \Rightarrow F^+ \nvdash G(F^+)$,
(4) $Sd(F) \Rightarrow G(F^+)$,
(5) $Sd(F) \Rightarrow F^+ \nvdash G(F^+)$,
(6) $HC(F) \Rightarrow F \vdash Sd(F) \rightarrow G(F^+)$,
(7) $\neg(Sd(F)\ [\wedge]\ HC(F))$,
(8) ¬"I sono F".

---

[150] «*(A) Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F". I perceive that it follows from the assumption that I am F that the Gödel statement G(F') would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "if I happened to be F, then G(F') would have to be true", and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that, I cannot be F after all*» (Cfr. *ivi*, p. 5).

[151] Cfr. P. Lindstrom, *Penrose's New Argument*, in «Journal of Philosophical Logic», XXX, 2001, vol. 30, no. 3, pp. 241-250. La traduzione dall'inglese, in questa e nelle successive citazioni, è ad opera nostra.

[152] Cfr. *ivi*, p. 244.

(1) afferma che se F è corretto, allora è corretta anche il sistema F^+ ottenuto dall'aggiunta a F dell'affermazione della sua correttezza. (2) afferma semplicemente che se F^+ è corretto (e perciò anche consistente) allora è vero il suo enunciato gödeliano $G(F^+)$. (4) deriva per concatenazione da (1) e (2), mentre (5) deriva per concatenazione da (1) e (3). (6) esplicita l'ipotesi tale per cui vale HC(F), ovvero tale per cui dall'insieme F di tutte le procedure dimostrative umanamente accessibili deriva che, se F è corretto, allora è vero l'enunciato gödeliano dell'insieme F^+ ottenuto per aggiunta di Sd(F) a F. (7) risulta da (6), per cui F è in grado di derivare $G(F^+)$, e da (5), per cui F^+ non può derivare $G(F^+)$, e afferma che F non può essere insieme corretto e umanamente completo. (8) deriva da (7) e conclude che F non equivale a me. La conclusione di (B), come riconosce lo stesso Penrose, non è dimostrata matematicamente, ma dipende dalla credenza circa la correttezza di F, nonché dal preciso significato di "correttezza" riferito ad F. Resta, perciò, da definire in maniera precisa l'operatore di correttezza "Sd" «in maniera tale che diventi chiaro che (1) […] è vero in modo inattaccabile».[153] Penrose, come abbiamo detto, usa il criterio della Π_1-correttezza, tale per cui F è Π_1-corretto, se tutte le Π_1-formule derivabili in F sono vere. Tale criterio, naturalmente, equivale alla richiesta più forte della consistenza di F. Si noti, perciò, che al posto di (1) è sufficiente il condizionale più debole (9): $Sd(F) \Rightarrow Cons(F \wedge Sd(F)) \Rightarrow Cons(F^+)$. Ma, se F è consistente, allora, per il secondo teorema di incompletezza, risulta che $F + \neg Cons(F)$ è consistente. Dunque, è evidente che $F + \neg Cons(F) + Cons(F + \neg Cons(F))$ non è consistente. Perciò, $F + \neg Cons(F)$ rappresenta un controesempio di (9), e stabilisce che il criterio di Π_1-correttezza non è valido.

Come sottolinea Galvan, per T = F:

> per evitare questo esito, si può intendere la proprietà di correttezza in un senso più forte, per esempio come predicato definibile nel linguaggio di T che vale di T se T è vera. In tal caso però si incorre in una serie di altre difficoltà, evitabili solo se Sd è inteso nel senso pieno del predicato di verità. Questo, tuttavia, per il teorema di indefinibilità di Tarski non può essere definito nel linguaggio di T, il che mina dalle fondamenta la possibilità di formulare correttamente passaggi cruciali dell'argomento […].[154]

Pertanto, data l'impossibilità di definire il predicato di verità per le formule di un sistema formale all'interno del linguaggio del sistema stesso, l'argomento di Penrose, nella formalizzazione (B), non può risultare conclusivo. Lindström conclude affermando che

[153] «*In such a way that it becomes clear that (1)* […] *is unassailably true*» (Cfr. *ivi*, p. 245).
[154] Cfr. S. Galvan, *Gödel e il modello computazionale della mente*, cit., p. 154.

se uno di questi argomenti possa o meno essere modificato in qualche (altro) modo per derivarne un argomento conclusivo rimane, almeno in teoria, una questione aperta. Dunque, rimane ancora fondamentalmente aperta la questione se il ragionamento (matematico) dell'uomo possa o meno essere catturato da un sistema formale (computabile) […]. Infatti, potrebbe risultare che la questione non sia ben definita e che per questo motivo non abbia una ben definita risposta.[155]

Anche Shapiro, dal canto suo, concorda nel ritenere che il problema centrale della presente questione sia l'"indeterminatezza" della tesi meccanicista a cui abbiamo accennato sopra. Egli, infatti, sostiene che «non c'è alcuna tesi meccanicista plausibile da offrire che sia sufficientemente precisa da essere confutata dai teoremi di incompletezza».[156]

Il dibattito tra meccanicista ed anti-meccanicista viene posto sul piano dell'idealizzazione, per quanto concerne sia la macchina che la mente. Rispetto alla prima, sappiamo già che in che cosa consiste l'idealizzazione della macchina di Turing. Ciò che, invece, può essere fonte di incomprensione e disaccordo riguarda il modo in cui una mente ideale deve essere intesa. Quest'ultima è concepibile alla maniera prospettata dai sostenitori della corrente di pensiero del "transumanesimo", ovvero come propria di esseri umani capaci di vivere indefinitamente nel tempo senza essere soggetti a logoramenti di tipo fisico o a limitazioni di altro genere. Sulla linea dell'ottimismo razionalistico che caratterizza la visione platonista di Gödel, Lucas e Penrose intendono l'idealizzazione della mente in un senso ancora più forte, ovvero in senso normativo. Si pensi al modo di operare della comunità scientifica e matematica. Eventuali inconsistenze ed errori nella costruzione dell'edificio del sapere scientifico altro non sarebbero che il risultato di temporanei fallimenti sperimentali o di incomprensioni passibili di revisione, e non invece qualcosa che apparterrebbe intrinsecamente al modo di procedere della ragione umana. Al contrario, quest'ultima andrebbe intesa come essenzialmente consistente e capace di conoscere ogni verità matematica. È fondamentale, quindi, anche se niente affatto banale, stabilire una comune e non-ambigua definizione di che cosa dovrebbe denotare il concetto di "mente ideale", affinché il dibattito tra sostenitori e oppositori del meccanicismo trovi un terreno d'intesa sul quale avanzare argomentazioni e posizioni che abbiano la pretesa di essere conclusive. Ed è pertanto evidente che un tale terreno consti di

[155] «*Whether or not one of these arguments can be modified in some (other) way to yield a conclusive argument remains, at least in theory, an open question. Thus, it is still fundamentally open whether or not human (mathematical) reasoning can be captured by a (computable) formal system* […]. *In fact, it may turn out that this question is not well-defined and so has no well-defined answer*» (Cfr. P. Lindstrom, *Penrose's New Argument*, cit., p. 249).

[156] «*There is no plausible mechanist thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems*» (Cfr. S. Shapiro, *Incompleteness, Mechanism, and Optimism*, in «The Bulletin of Symbolic Logic», IV, 1998, vol. 4, no. 3, p. 275). Nelle successive citazioni tratte dall'opera, la traduzione dall'inglese è ad opera nostra.

elementi teorici la cui accettazione da parte di entrambe le parti richiede considerazioni che travalicano le implicazioni puramente formali dei teoremi di limitazione gödeliani.

A questo punto, si consideri l'insieme **K** (che corrisponde alla matematica soggettiva di Gödel, o al sistema *Q* dell'argomento di Penrose) di tutte e sole le proposizioni aritmetiche dimostrabili con certezza dall'uomo, e l'insieme **T** contenente tutte e sole le proposizioni aritmetiche vere (la matematica oggettiva di Gödel). Ora, supponendo che valga il principio di bivalenza, come rileva Shapiro, esistono soltanto le due seguenti possibilità:

1) **K** = **T**. Tutte le proposizioni aritmetiche vere sono anche idealmente dimostrabili dall'uomo (primo corno del dilemma gödeliano). Per il teorema di Tarski, e per il secondo teorema di incompletezza, sappiamo che **T** non può essere formalizzabile. Perciò, se **K** = **T**, risulta che nemmeno **K** è formalizzabile e che, pertanto, non può esistere una macchina di Turing in grado di enumerare ricorsivamente tutti i teoremi contenuti in **K**.
2) **K** ≠ **T**. Esistono proposizioni aritmetiche vere assolutamente indimostrabili (secondo corno del dilemma), e **K** è ricorsivamente enumerabile da parte di una macchina di Turing (W_e nella notazione di Shapiro, dove *e* indica il gödeliano di *W*), ovvero **K** = W_e.

Se assumiamo l'ipotesi 2), l'enunciato di consistenza di W_e, ovvero Cons_e, non è derivabile in **K**, e costituisce una verità aritmetica assolutamente indimostrabile. In altre parole, nessun essere umano ideale sarebbe in grado di riconoscere che W_e è consistente, e quindi che W_e = **K**. Del resto, questo risultato è coerente con quanto afferma Gödel in un importante passo della Gibbs Lecture:

> Evidentemente nessun sistema rigorosamente definito di assiomi corretti può comprendere tutta la matematica oggettiva, dal momento che la proposizione che afferma la coerenza del sistema è vera ma non dimostrabile nel sistema stesso. Però, per la matematica soggettiva [ovvero l'insieme **K** a cui si riferisce Shapiro], non è escluso che possa esistere una regola finita che fornisce tutti i suoi assiomi evidenti. Tuttavia, se esiste una regola di questo genere, noi con la nostra comprensione umana non potremmo certamente mai riconoscerla come tale, cioè non potremmo mai sapere con certezza matematica che tutte le proposizioni che genera sono corrette.[157]

Per quanto riguarda le implicazioni dei teoremi di incompletezza, dunque, Shapiro mette in luce un punto essenziale:

[157] Cfr. K. Gödel, *Alcuni teoremi basilari sui fondamenti della matematica e loro implicazioni filosofiche (1951)*, in *Opere. Volume 3: Saggi inediti e conferenze*, cit., p. 273.

Si supponga che **K** = W_e […]. Tutto ciò che segue dagli […] argomenti di Gödel (e di Lucas e Penrose) è che né i nostri soggetti idealizzati, né noi umani mortali possiamo sapere (*de re*) che W_e è corretto o almeno consistente *con la stessa assoluta certezza matematica*. Nulla ci vieta (ad ora) di concludere la correttezza di W_e sulla base di qualcosa di meno di un'assoluta e certa evidenza. Il meccanicista è libero di argomentare la sua tesi su basi empiriche o addirittura sulla base di qualche sorta di terreno metafisico *a priori* scevro da dimostrazioni matematiche».[158]

Come abbiamo visto, un simile esito viene alla fine riconosciuto anche da Lucas e Penrose. Infatti, entrambi sono disposti ad ammettere che se anche fossimo equivalenti ad una macchina di Turing, non potremmo riconoscere questo fatto con certezza matematica assoluta, e quindi che non è possibile confutare il meccanicismo a priori. Shapiro, perciò, giudica errata la pretesa di Lucas per la quale spetterebbe al meccanicista l'onere della prova della consistenza della macchina di Turing che si suppone essere modello della mente umana. Come egli fa notare, «"la parola del meccanicista" non fornisce a Lucas una premessa che egli possa usare […] semplicemente perché questa parola non corrisponde ad una certezza matematica».[159] L'autore, poi, analizzando l'argomento "nuovo" di Penrose, sottolinea, come fa anche Lindström, che il nodo centrale, e al contempo il punto debole, di esso consiste nel fatto che Penrose (alla pari di Lucas) non possiede, né può definire (per il teorema di Tarski), un «predicato generale di verità» che determini la correttezza dei metodi di ragionamento matematico dell'uomo; metodi che, nell'ipotesi meccanicistica, sono racchiusi all'interno dell'insieme **K**. Perciò, dal momento che non è possibile dimostrare formalmente la consistenza di una presunta macchina W_e che si ipotizza essere equivalente a noi (cioè a **K**), all'anti-meccanicista non rimane che assumere un criterio di correttezza più debole rispetto alla certezza matematica assoluta; criterio che, come abbiamo visto sopra, non consente all'argomentazione di Penrose di essere conclusiva. Dal punto di vista formale, dunque, Shapiro ha buon gioco nell'avanzare la sua obiezione. Tuttavia, teniamo qui a ribadire che la mira teorica, potremmo dire implicita, delle argomentazioni di Lucas e Penrose non è tanto quella di "dimostrare" la falsità del meccanicismo, quanto di "mostrare", o far emergere, attraverso la sua *reductio ad absurdum*, l'implausibilità di una simile posizione sul piano dialettico anziché su quello del formalismo logico.

---

[158] «*Suppose that* **K** = W_e […]. *All that follows from the* […] *Gödel (and Lucas-Penrose) arguments is that neither our idealized subjects nor we mortal humans can know (*de re*) that* W_e *is sound or even consistent* with that same absolute mathematical certainty. *There is nothing (so far) preventing us from concluding the soundness of* W_e *on less than absolutely certain evidence. The mechanist is free to argue for his thesis on empirical grounds or even on some sort of* a priori *metaphysical grounds short of mathematical proof*» (Cfr. S. Shapiro, *Incompleteness, Mechanism, and Optimism*, p. 281).

[159] «*This "word of the mechanist" does not give Lucas a premise he can use* […] *simply because this "word" does not amount to mathematical certainty*» (Cfr. *ivi*, p. 283).

Un altro punto importante della critica di Shapiro riguarda un noto argomento avanzato da Turing,[160] e generalizzato in seguito da Feferman.[161] Abbiamo già detto che la costruzione mediante diagonalizzazione dell'enunciato gödeliano di un dato sistema formale S, come ad esempio l'aritmetica di Peano oppure PRA, e l'aggiunzione di tale enunciato all'impianto assiomatico di quest'ultimo (la quale implica un'estensione non-conservativa di tale sistema di partenza), se prese insieme, consistono in una procedura del tutto meccanica ed eseguibile da una macchina di Turing. L'argomento di Turing prevede l'iterazione di tale procedura lungo la serie degli ordinali e degli ordinali transfiniti.[162] In questo modo, una macchina programmata per fare questo, sarebbe in grado, per così dire, di crescere su se stessa e di imitare il processo di estensione non-conservativa di una teoria, e quindi di stare al passo con l'anti-meccanicista nel "gioco della gödelizzazione" di Lucas.

In primo luogo, è necessario introdurre il cosiddetto "principio di riflessione" (*reflection principle*), che Feferman definisce come segue:

> Per principio di riflessione intendiamo la descrizione di una procedura per aggiungere a un qualunque insieme di assiomi *A* certi nuovi assiomi la cui validità segue dalla validità degli assiomi *A* e che esprimono formalmente, all'interno del linguaggio di *A*, conseguenze evidenti dell'assunzione che tutti i teoremi di *A* sono validi.[163]

Si noti che tale principio equivale all'aggiunta dell'enunciato gödeliano γ del sistema formale S al suo impianto assiomatico. Infatti, la consistenza di S implica la verità di γ. Di conseguenza, dal momento che se "crediamo" nella consistenza di S, crediamo anche nella verità di γ, allora per il principio di riflessione (che, lo ricordiamo, non è conservativo rispetto alla consistenza né alla coerenza) anche $S + \gamma$ sarà consistente.

Se aggiungiamo (meccanicamente) γ agli assiomi di S otteniamo un nuovo sistema S_1. Aggiungendo poi γ_1 agli assiomi di S_1 otteniamo un ulteriore sistema S_2, e così via. Ora, la successione dei sistemi formali di volta in volta costruiti, applicando la procedura appena descritta, costituisce un insieme ben ordinato $\{S_0, S_1, S_2, \ldots, S_n, \ldots\}$ (con $S = S_0$), in cui ad ogni S_n appartenente

---

[160] Cfr. A.M. Turing, *Computing Machinery and Intelligence*, in «Mind», LIX, 1950, vol. 59, no. 236.

[161] Cfr. S. Feferman, *Transfinite Recursive Progressions of Axiomatic Theories*, in «The Journal of Symbolic Logic», XXVII, 1962, vol. 27, no. 3, pp. 259-316.

[162] Dato l'insieme infinito di numeri ordinali finiti $\mathbb{N}(\prec)$, equipotente all'insieme $\mathbb{N}$ dei numeri naturali, Cantor definì l'elemento limite ω della serie $\mathbb{N}(\prec)$ come il primo ordinale transfinito. Si veda in proposito G. Basti, *Le radici forti del pensiero debole*, cit., pp. 192-193.

[163] «*By a* reflection principle *we understand a description of a procedure for adding to any set of axioms* A *certain new axioms whose validity follow from the validity of the axioms* A *and which formally express, within the language of* A, *evident consequences of the assumption that all the theorems of* A *are valid*» (Cfr. S. Feferman, *Transfinite Recursive Progressions of Axiomatic Theories*, cit., p. 274).

alla successione corrisponde l'n-esimo ordinale appartenente all'insieme $\mathbb{N}(\prec)$ dei numeri ordinali finiti. Ora, l'ordinale di $\mathbb{N}(\prec)$ si denota con ω, e rappresenta il primo ordinale transfinito, ovvero l'ordinale limite della serie degli ordinali che appartengono a $\mathbb{N}(\prec)$, tale per cui, per ogni $n \in \mathbb{N}(\prec)$, $n < \omega$. Indichiamo allora con S_ω l'insieme $\{S_0, S_1, S_2, \ldots, S_n, \ldots\}$, il quale specifica un sistema formale con un insieme infinito di assiomi dotato di un preciso ordinamento, ed equivale ad una macchina di Turing istruita per iterare all'infinito il principio di riflessione a partire dal sistema di base S_0 lungo la successione degli ordinali finiti. Per tale principio, inoltre, si noti che se S_0 è ricorsivamente enumerabile, allora lo è anche S_1, e che se per ogni $n < \omega$, S_n è ricorsivamente enumerabile, allora lo è anche S_ω. Naturalmente, segue anche che se crediamo che S_0 sia consistente, allora, per il principio di riflessione, possiamo credere che anche S_ω sia consistente. Inoltre, in quanto costituisce un sistema formale ben definito, anche rispetto a S_ω è possibile costruire un enunciato indecidibile, sia esso γ_ω. In modo analogo a prima, aggiungendo γ_ω agli assiomi di S_ω si ottiene il sistema $S_{\omega+1}$, e aggiungendo $\gamma_{\omega+1}$ a $S_{\omega+1}$ si ottiene $S_{\omega+2}$, ecc. Iterando il principio di riflessione lungo gli ordinali successori di ω otteniamo una nuova successione ordinata di estensioni non-conservative (indicate dalla successione degli ordinali ω, $\omega + 1$, $\omega + 2$, …), la cui serie infinita si può denotare mediante un sistema formale indicato dall'ordinale limite 2ω ($\omega + \omega$), ovvero $S_{2\omega}$. Oltre agli ordinali successori dell'ordinale limite 2ω (2ω, $2\omega + 2$, $2\omega + 3$, …), esiste poi la successione ordinata degli ordinali limite 3ω, 4ω, …, $\omega\times\omega = \omega^2$, ω^3, ω^4, …, ω^ω, …, ecc. Come spiega Shapiro, che si rifà alla lezione di Feferman, se si indica con O l'insieme infinito dei numeri naturali che denotano gli ordinali, è dimostrabile che O non costituisce un insieme ricorsivamente enumerabile.[164] Feferman, a questo riguardo, afferma che un matematico

> sfortunatamente, come avanza sempre più in là nella collezione di sistemi […], potrebbe non essere in grado di acquisire la conoscenza necessaria per decidere, per ogni dato d_0, se valga o meno $d_0 \in O$. In altre parole, per procedere, egli potrebbe doversi rivolgere ad un 'oracolo'.[165]

Ciò che Feferman intende dire è che più ci si inoltra lungo l'infinita successione degli ordinali limite, più la base sulla quale poter applicare il principio di riflessione si fa incerta. Abbiamo visto, infatti, che, per tale principio, la validità di un nuovo assioma aggiunto a quelli di un certo sistema formale di partenza dipende dalla validità del sistema stesso, la quale potrebbe non risultare affatto

[164] Cfr. S. Shapiro, *Incompleteness, Mechanism, and Optimism*, cit., pp. 287-288.

[165] «*Unfortunately, even as he advances farther and farther out into the collection of systems* […], *he may not be able to gain the knowledge necessary to decide, of any given* d_0, *whether or not* $d_0 \in O$. *In other words, in order to proceed, he may have to appeal to an 'oracle'*» (Cfr. S. Feferman, *Transfinite Recursive Progressions of Axiomatic Theories*, cit., p. 279).

"evidente" in caso di sistemi di un ordine di complessità assai elevato. Inoltre, come riporta Shapiro, il teorema di Church-Kleene afferma che «non esiste alcuna enumerazione *ricorsiva* di tutti gli ordinali ricorsivi», e quindi che «nessuna macchina può iterare la procedura attraverso tutti e soli gli ordinali ricorsivi».[166] Tuttavia, il punto fondamentale è che non si dispone di una prova certa che, al contrario, la mente umana sia in grado di farlo, oltrepassando così il limite che segna fin dove una macchina di Turing può giungere. In base a quanto abbiamo detto, perciò, sembra che la "competizione" tra mente e macchina, relativamente a questo ristretto contesto formale, debba risolversi in uno stallo.

La convinzione espressa da Lucas,[167] secondo la quale la mente umana sarebbe capace di enumerare tutti gli ordinali, o quanto meno di generare un ordinale in più rispetto a una qualsiasi macchina data, mediante una presunta procedura non-algoritmica e creativa, è certamente molto forte. Infatti, una simile capacità consentirebbe di determinare la verità o la falsità di ogni enunciato aritmetico, e di conseguenza implicherebbe l'equivalenza tra gli insiemi della matematica soggettiva e di quella oggettiva.[168] Tuttavia, come conclude giustamente Shapiro, i teoremi di limitazione, presi in se stessi, non forniscono una prova formale della falsità del meccanicismo. In altri termini, ad essere problematica, per Shapiro, è l'idealizzazione normativa della mente umana. Quest'ultima – egli afferma – «è consonante con un'epistemologia matematica di lunga data. […] Cominciamo da assiomi auto-evidenti per procedere attraverso una deduzione scevra da ogni falla. Si definisca questo il modello *euclideo* della matematica».[169] Gödel, Lucas e Penrose sembrano decisamente assumere tale modello, il quale, tuttavia, è stato messo in seria discussione, in generale, dalla riflessione filosofica post-moderna, con la quale si è abbandonata ogni pretesa di accesso ad una verità assoluta come fondamento del sapere. Se si tentasse, poi, di confutare la tesi meccanicistica sulla base di un criterio epistemologico alternativo a quello euclideo, come mostra Shapiro,[170] si incorrerebbe nella medesima difficoltà che abbiamo visto applicarsi al caso della Π_1-correttezza analizzato di Lindström.

---

[166] Cfr. S. Shapiro, *Incompleteness, Mechanism, and Optimism*, cit., pp. 288-289.

[167] Cfr. *ivi*, 289-291. In proposito, si vedano i passi riportarti da Shapiro dal saggio di Lucas del 1996 (Cfr. J.R. Lucas, *Minds, Machines, and Gödel: A Retrospect*, in *Machines and Thought: The Legacy of Alan Turing, Volume 1*, ed. P.J.R. Millican e A. Clark, Oxford University Press, Oxford 1996).

[168] «Il risultato di Feferman implica che se un essere umano è in grado di iterare i membri di *O* (o se egli può decidere intorno all'appartenenza in *O*), allora egli ha la capacità di determinare il valore di verità di *ogni* enunciato matematico. […] Per quanto riguarda la verità aritmetica, noi saremmo non solo infallibili, ma onniscienti» («*The Feferman result entails that if a human can iterate the members of* O *(or if she can decide membership in* O*) then she has the wherewithal to determine the truth value of* every *arithmetic sentence*. […] *When it comes to arithmetic truth, we would be not only infallible, but omniscient*» (Cfr. S. Shapiro, *Incompleteness, Mechanism, and Optimism*, cit., p. 290).

[169] «*The normative idealization is consonant with a longstanding epistemology for mathematics. […] We start with self-evident axioms and proceed by gap free deduction. Call this the* Euclidean *model of mathematics*» (Cfr. *ivi*, p. 293).

[170] Cfr. *ivi*, p. 295 e seguenti.

L’anti-meccanicista, cioè, non sarebbe in grado di dimostrare “definitivamente”, sulla base di questo nuovo presunto criterio di validità, la correttezza (o la consistenza) di un sistema formale assunto come modello computazionale della mente.

Parte III
PROSPETTIVE ULTERIORI

Capitolo 7
Dalla logica formale all'ontologia: logica della dimostrazione e logica dell'evidenza

Nei paragrafi precedenti abbiamo preso in esame, relativamente al dibattito intorno alle implicazioni filosofiche dei teoremi di Gödel, gli argomenti dei due esponenti principali della posizione anti-meccanicistica, oltre che alcune tra le principali e più interessanti critiche che sono state mosse loro. Ciò che ne è emerso è che la tesi del meccanicismo non è sufficientemente determinata per poter essere confutata, come vorrebbe invece l'anti-meccanicista, a partire dai soli teoremi di incompletezza. Naturalmente, la determinatezza, qui, è misurata in relazione alla formulazione di tale tesi nei termini rigorosi della logica formale. Abbiamo anche visto come Gödel, nonostante abbia sempre sostenuto una posizione filosofica mentalista, o platonista, riguardo la natura della mente umana e le sue capacità matematiche abbia comunque usato un'adeguata cautela nel tirare le somme dal risultato del suo lavoro, in particolare per quanto concerne il programma formalista di Hilbert, e più in generale per quanto riguarda la logica formale e la teoria degli insiemi. Questo tipo di cautela è l'aspetto del dibattito su cui vogliamo concentrare l'attenzione nel presente paragrafo. Come abbiamo detto, gli autori presi in esame concordano, più o meno esplicitamente, su un punto in particolare: a supportare la visione meccanicistica piuttosto che quella mentalista o platonista, non sono tanto i teoremi di incompletezza presi in se stessi, cioè considerati nella loro veste logica, quanto degli specifici presupposti di natura filosofica e, in particolare, ontologica.

Prima di affrontare tale questione, però, è opportuno tornare a riflettere sulla nozione di verità relativamente all'ambito della logica formale e, in particolare, riguardo le conseguenze dei teoremi gödeliani. Come nota Galvan, la verità di un enunciato indecidibile γ in un dato sistema formale, poniamo P, possiede uno strutturale «significato infinitario». Si noti, a proposito, che l'indecidibilità di γ in P, e quindi la sua infinitarietà, non è da intendersi in senso assoluto, bensì è relativa soltanto allo specifico sistema assiomatico di P, ovvero non è tale da non poter essere risolta mediante una qualche procedura finitaria in una certa estensione di P. A tale scopo, come abbiamo visto, esistono due vie a disposizione: da un lato, è possibile formalizzare l'impianto semantico della teoria P, e quindi derivare l'esistenza del suo modello standard all'interno di una metateoria più potente che

rappresenta un'estensione conservativa della prima; dall'altro, mediante il principio di riflessione è possibile semplicemente aggiungere γ agli assiomi di P[171] per ottenere un secondo sistema P_1 in cui lo stesso γ è decidibile attraverso una procedura finitaria. Ripetendo la procedura di gödelizzazione su P_1, è possibile formulare un nuovo enunciato indecidibile γ_1. Aggiungendo poi γ_1 agli assiomi di P_1 si ottiene un terzo sistema P_2 in cui γ_1 è decidibile. Anche relativamente a P_2 è possibile formulare un enunciato γ_2 in esso indecidibile…; e così via. Un simile procedimento, iterabile *ad infinitum*, fa notare Galvan, non è conservativo rispetto alla teoria di partenza ma, al contrario, svolge «un ruolo creativo», dove la creatività, naturalmente, è da imputare all'azione del matematico. A questo proposito, egli scrive:

> [Si] potrebbe recepire l'impressione che, essendo la verità identificabile con la deducibilità in teorie superiori, la distinzione tra finito ed infinito sia a tal punto relativizzata da essere completamente dissolta e in tal modo il finito venga a costituire in definitiva la categoria gnoseologicamente originaria. Tale impressione è però fallace, in quanto, per essere giustificata, dovrebbe sottendere la possibilità d'esecuzione d'un regresso all'infinito esaustivo. […] Non ci sarebbe data la possibilità di collocarci in una teoria finale […] in cui i teoremi coprano tutte le verità. La conclusione è che, se la verità di certe proposizioni in una teoria si può concepire come il precipitato della loro dimostrabilità in altre, il momento della *evidenza veritativa* è imprescindibile nella costruzione e giustificazione delle teorie, pena l'impossibilità di conferire un significato alla stessa nozione di dimostrazione […].[172]

In altri termini, il processo di derivazione della verità degli enunciati indecidibili, lungo la scala delle estensioni di una teoria, deve presupporre come data una dimensione comprensiva della totalità della successione infinita, per sua natura esterna alla successione stessa, che fondi il processo medesimo, ossia che ne costituisca la condizione ultima di possibilità.[173] La nozione di verità, perciò, non può venire identificata con quella di dimostrabilità: quella costituisce «una sorta di ideale regolativo al di là della portata delle procedure finitarie»,[174] e, in questo senso, «si dà come qualcosa di *autonomo* rispetto alle nostre capacità di dimostrazione».[175] Per questo motivo, «per quanto la verità sia analizzabile come deducibilità in una teoria più potente, essa deve essere assunta, almeno

[171] Galvan si riferisce, più rigorosamente, all'«aggiunzione delle equivalenze definitorie per il concetto di formula vera in [*P*]». Cfr. S. Galvan, *Pensiero e Logica*, cit., p. 122.

[172] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 190. Il corsivo è dell'autore.

[173] Si noti, qui, l'analogia con l'argomentazione di Popper (vedi capitolo terzo) circa l'imprescindibilità, per il fisicalismo, della considerazione della totalità del sistema della logica al fine di render conto dell'utilità di ogni singola inferenza logica.

[174] Cfr. *ivi*, p. 188.

[175] Cfr. S. Galvan, *Pensiero e Logica*, cit., p. 127. Il corsivo è dell'autore.

al livello più alto di formalizzazione raggiunto, come *verità rispetto al modello*»[176], ovvero a prescindere dalla sua dimostrabilità in una metateoria di volta in volta superiore.

La nozione di modello, spiega Galvan, si caratterizzare attraverso due categorie: l'astrattezza e l'intenzionalità. Per comprendere questo punto occorre ricordare che, per il teorema di Tarski sopra menzionato, la formalizzazione delle condizioni di verità degli enunciati di una teoria eccede la capacità espressiva della teoria stessa. Da ciò segue che il modello di una teoria formale quale *P*, che resta pur sempre definibile all'interno degli assiomi di una metateoria più forte, possiede un contenuto semantico astratto rispetto all'insieme finito delle relazioni sintattiche che intercorrono tra gli elementi di *P* (ovvero non formalizzabile da quest'ultime). Tuttavia, come evidenzia Galvan, l'esistenza stessa di una struttura semantica di una teoria non può dipendere, pena la ricaduta in un *regressus ad infinitum*, dalla sua derivabilità in una teoria superiore; bensì deve trovare giustificazione sul piano dell'intenzionalità, cioè nel suo «darsi come contenuto *inteso*»,[177] che appartiene ad una dimensione che trascende quella linguistica della teoria di cui tale struttura è modello. Quanto detto mette in luce la profonda differenza, almeno per quanto riguarda il campo della matematica, tra il modo di funzionare di una intelligenza artificiale e quello della mente umana. Pare cioè che un aspetto chiave del procedere "creativo" di quest'ultima sia quello dell'evidenza, ovvero di una sorta di capacità di "visualizzazione"[178] o "intuizione", di cui la prima risulta mancante, e tale per cui essa è in grado di comprendere determinate strutture astratte, come ad esempio quella costituita dall'insieme dei numeri naturali.

A questo proposito, un tentativo di risposta al problema della differenza tra mente e macchina di Turing proviene dallo sviluppo di una branca della logica intensionale, ovvero la logica epistemica.[179] Di quest'ultima fanno parte la logica dell'evidenza e quella della dimostrabilità. Galvan intende mostrare, attraverso il loro confronto, come i principi della prima non siano affatto riducibili a quelli della seconda; evidenziando, così, come i due tipi di logica stiano alla base della spiegazione dei due diversi modi di operare da parte, rispettivamente, della mente e dell'intelligenza artificiale.[180]

---

[176] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 191. Il corsivo è dell'autore.

[177] Cfr. *ivi*, p. 192. Il corsivo è dell'autore.

[178] «Anche se la verità che si vede si può esplicitare come qualcosa di dimostrabile, non è facile convincersi che alla fin fine il vedere sia una *illusione*, generata dal fatto che, in ultima istanza, il pensare è una sorta di *computazione contratta*.» (Cfr. Galvan, *Gödel e il modello computazionale della mente*, cit., p. 155). Il corsivo è dell'autore.

[179] Secondo la definizione che ne dà Giordani: «La logica epistemica è indirizzata allo studio dei concetti fondamentali della teoria della conoscenza ed è quindi finalizzata al presentare, in modo univoco ed esplicito, la struttura delle disposizioni intenzionali di tipo speculativo. In questo senso, la logica epistemica si è sviluppata innanzitutto come logica della credenza e della conoscenza, dal momento che il credere e il conoscere identificano i modi primari di questo tipo di disposizioni.» (Cfr. A. Giordani, *Logica dell'evidenza*, in "Rivista di Filosofia Neo-Scolastica", XCII, 2000, vol. 92, no. 3/4, p. 582).

[180] Per l'analisi dettagliata rimandiamo a S. Galvan, *Gödel e il modello computazionale della mente*, cit., pp. 156-164.

In questa sede, ci limiteremo ad illustrare, riprendendo l'esposizione di Galvan, soltanto il caso per il principio di ω-completezza associato all'operatore d'evidenza *E* ("è evidente che…").

Data la proprietà A sui numeri naturali standard, il principio di ω-completezza per la logica dell'evidenza (ω*E*) si può esprimere attraverso le tre formulazioni seguenti:[181]

1. $E(\mathrm{A}(0), E(\mathrm{A}(\bar{1})), \ldots \rightarrow E\forall \mathrm{xA(x)}$
2. $\forall \mathrm{n}E\mathrm{A}(\bar{\mathrm{n}}) \rightarrow E\forall \mathrm{xA(x)}$
3. $\forall \mathrm{x}E\mathrm{A(x)} \rightarrow E\forall \mathrm{xA(x)}$

Notiamo che le prime due formulazioni del principio, a differenza della terza, hanno un carattere infinitario. Infatti, l'antecedente della seconda (il quale non è altro che la forma contratta di quello della prima) è una formula infinita in cui la proprietà A viene applicata a tutti i numeri naturali n. L'antecedente della terza, invece, è una formula finitariamente significante.[182] Tuttavia, tutte e tre le formulazioni, spiega Galvan, hanno il medesimo significato, «in quanto il dominio di interpretazione della variabile x è l'insieme dei numeri naturali standard, vale a dire l'insieme degli oggetti a cui corrispondono esattamente i termini numerali».[183] Il principio ω*E* afferma dunque che se per ogni numero naturale n è evidente che vale A(n), allora è evidente che ∀xA(x), cioè che per tutti i numeri naturali vale A. Se escludiamo la possibilità per l'intelletto di cogliere in atto l'infinità potenziale dell'insieme dei numeri naturali, la capacità di cogliere come evidente il fatto che per ogni n vale A(n), secondo Galvan, non è data se non «in forza della *struttura generale e astratta* di numero naturale standard (la sua intensione)».[184] Ora, continua l'autore, «per quanto sia immediato il principio di ω-completezza nella logica dell'evidenza, esso non vale nella logica della dimostrabilità. Anzi è un teorema della metamatematica post-gödeliana che l'ω-completezza è addirittura equivalente alla inconsistenza».[185] La causa della divergenza tra questi due tipi di logica, relativamente al principio di ω-completezza, risiede perciò nel carattere di evidenza ed intuitività del modello astratto dei numeri naturali e nell'impossibilità di realizzarne una formalizzazione completa all'interno di una teoria.[186]

---

[181] Cfr. *ivi*, p. 158.

[182] Si veda quanto detto a proposito della nozione di "finitariamente significante" nel capitolo primo della Parte prima.

[183] Cfr. *ivi*, p. 158.

[184] Cfr. *ibidem*. Il corsivo è dell'autore.

[185] Cfr. *ivi*, p. 159. Per la dimostrazione, cfr. *ivi*, pp. 159-160.

[186] «Non esiste nessuna teoria formale capace di definire formalmente il concetto di numero naturale. Vale a dire, non esiste nessuna teoria formale, per quanto questa possa essere potente, in cui sia presente una formula N(x) interpretabile categoricamente (cioè a meno di isomorfismi) sull'insieme dei numeri naturali. Il resto è consequenziale: se la teoria formale non ha presa sul concetto di numero, non può garantire il passaggio dalla dimostrabilità di A($\bar{\mathrm{n}}$) per ogni singolo numerale alla dimostrabilità della formula universale; non lo può

Un punto di estremo interesse per il nostro lavoro viene messo in luce da Galvan, dove egli mostra che, applicando il principio della logica dell'evidenza ω*E* alla formula che esprime la consistenza di una teoria formale quale PRA ($\mathrm{Cons}_{\mathrm{PRA}}$), è possibile «chiarire […] in che senso l'evidenza (o intuizione) costituisca un superamento delle prestazioni di cui è capace una macchina, senza che questo significhi che la mente è capace di cogliere la verità di proposizioni inaccessibili alla macchina». «Attraverso tale applicazione» – continua l'autore – «diventa così possibile intravvedere, negli argomenti di Lucas e Penrose, un'anima di verità in mezzo a un corpo caduco».[187] Tuttavia, ciò che in questo modo si ottiene, si noti, è soltanto l'"evidenza" della consistenza di PRA ($E(\mathrm{Cons}_{\mathrm{PRA}})$), ma non anche la sua "verità".[188] A questo riguardo, Galvan afferma che

> per ottenere la verità della consistenza occorre impegnarsi ad accettare che l'evidenza di $E(\mathrm{Cons}_{\mathrm{PRA}})$ – che non è una evidenza finitista – è tuttavia riflessiva, ossia tale da implicare la verità dei suoi contenuti. Questa conclusione è altamente giustificata – perché prende le mosse da evidenze finitiste, sfrutta il principio di ω-completezza per *E*, a sua volta fondato sulla evidenza del modello standard –, ma non si tratta in ogni caso di conclusione giustificata in modo incondizionato, perché non si tratta di proposizione oggetto di evidenza puramente finitista.[189]

Risulta perciò chiaro che i due tipi di logica non sono riducibili l'uno all'altro, e quindi che la mente umana sembra in grado di trascendere l'ambito della dimostrabilità formale. Ma che cosa giustifica la diversità della mente rispetto alla macchina? Si tratta di un problema, questo, non ancora risolto sulla base delle odierne conoscenze in ambito scientifico e matematico, e che chiama in causa il futuro sviluppo dell'indagine sperimentale nel campo della neurobiologia e dell'intelligenza artificiale.

Abbiamo già esposto il dilemma avanzato da Gödel: o la mente è infinitamente diversa (e superiore) ad una macchina di Turing, oppure esistono verità matematiche assolutamente indimostrabili. I due corni del dilemma rappresentano due posizioni "filosofiche" ben diverse: il primo dipinge quello che Wang definisce "ottimismo razionalistico", che abbiamo detto attribuire alla mente capacità non formalizzabili all'interno di un sistema, e per il quale ogni proposizione della matematica oggettiva è conoscibile dalla mente; il secondo, invece, presuppone l'esistenza di un

---

fare perché la teoria formale non sa, non riuscendo a definirlo, che il concetto di naturale coincide con quello di immagine di un numerale, ovvero che l'insieme dei naturali è esaurito da quello delle immagini dei numerali» (Cfr. *ivi*, p. 161).

[187] Cfr. *ivi*, p. 162.

[188] Per i dettagli dell'applicazione del principio di ω-completezza (ω*E*) all'enunciato di consistenza di PRA, cfr. *ivi*, pp. 162-163.

[189] Cfr. *ivi*, p. 163.

modello computazionale della mente, e perciò presuppone l'esistenza di verità matematiche assolutamente inconoscibili in quanto assolutamente indimostrabili. Per la logica della dimostrabilità, la nozione di verità equivale a quella di derivabilità in una teoria superiore. Su questa base, spiega Galvan, i sostenitori della visione meccanicistica sperano di colmare il divario tra il funzionamento della mente, che comprende l'appello a contenuti d'evidenza infinitari, e quello di una macchina, che si rifà esclusivamente ad evidenze finitarie, cercando di mostrare come una simile differenza sia in realtà soltanto apparente.[190] Tuttavia, come abbiamo mostrato nel paragrafo §6.5, l'identificazione di verità e dimostrabilità formale è legittimata soltanto fino ad un certo punto critico del processo di costruzione dei formalismi attraverso gli ordinali transfiniti e gli ordinali limite. A mano a mano che si procede lungo la successione, perdiamo sempre più confidenza circa la consistenza (o, il che è equivalente, la verità degli assiomi) della teoria sulla base della quale poter dimostrare la consistenza della sotto-teoria. Inoltre, come abbiamo visto, esiste un limite, dettato dal teorema di Church-Kleene, oltre il quale non è possibile spingere il processo di estensione definitoria. Perciò, il contenuto semantico astratto di una teoria formale, ovvero il suo dominio di interpretazione, è giustificabile, da ultimo, non in base ad una formalizzazione in una teoria superiore, bensì sul piano dell'intenzionalità e, quindi, dell'evidenza. Ma ciò è proprio quel tipo di giustificazione che il meccanicista non è disposto ad accettare. Egli deve dunque rifarsi a ragioni di tipo induttivo o di utilità selettiva, le quali, afferma Galvan, vanno incontro a difficoltà gravi, simili a quelle già messe in risalto, a modo loro, da Popper e Penrose.[191] Pertanto, «è […] richiesta» – conclude Galvan – «una forma di conoscenza *a*

[190] «Secondo tali autori la verità matematica è in realtà qualcosa di costruito dalla mente umana, in quanto quest'ultima – pur essendo identica ad una macchina – è capace di oltrepassare sempre se stessa – senza modificare la propria struttura finitaria –, esattamente come un sistema formale può essere via via esteso attraverso l'aggiunta di nuovi assiomi. Naturalmente i medesimi autori riconoscono che Gödel abbia ragione nel dire che, se la mente è una macchina, essa non può avere accesso a verità matematiche nella misura in cui queste si presentano dotate di un contenuto infinitario. Però, se tali verità si presentano alla mente in veste finitaria – il che accade quando diventano teoremi formali di teorie superiori – le cose cambiano» (Cfr. *ivi*, p. 168).

[191] «Nel primo caso, se ad esempio l'uso di una teoria non ha mai portato alla contraddizione, allora è plausibile che l'estensione avvenga nella direzione di aggiungere come ulteriore assioma l'espressione della sua consistenza. Nel secondo, viene selezionato l'assioma (o teoria) che assicura un adattamento migliore. Entrambe le ipotesi presentano tuttavia serie difficoltà. La soluzione induttiva è affetti da quattro ordini di difficoltà: (i) l'induzione presuppone forme d'evidenza astratta, di tipo intensionale […]: non è possibile, ad esempio, operare induzioni corrette se non si ammette la capacità di cogliere un insieme adeguato di predicati proiettabili […]; (ii) l'induzione fornisce connessioni di tipo empirico (dunque legate alla fattualità dei mondi sperimentati e perciò contingenti) e non necessario di tipo apriorico; (iii) l'induzione ha normalmente un valore statistico, che non si addice all'ambito matematico […]; (iv) […] si può intendere l'induzione come un metodo che consente il passaggio da un caso a tutti […]. Ebbene in tale interpretazione l'induzione si basa sull'intuizione dell'oggetto in questione e coincide con la generalizzazione a tutti gli oggetti di una proprietà […] rilevata in un oggetto singolo […]. La soluzione selezionistica va incontro a due difficoltà particolari: (i) Innanzitutto come è possibile dire che il meccanismo di selezione privilegia le verità rispetto alle falsità? Si può rispondere: perché la verità matematica è più adattiva. Tuttavia questo vale solo per certe verità

priori, anche se *contenutistica*, di natura *astrattiva*».[192] In altre parole, il processo di fondazione di una teoria non può basarsi esclusivamente su quel tipo di intuizione sensibile legata alla formalizzazione di un sistema mediante segni primitivi, e quindi non può davvero prescindere dalla considerazione di un modello astratto di carattere infinitario concepibile dalla mente in maniera intenzionale. Per riprendere un'affermazione di Galvan che abbiamo già citato nel capitolo terzo, risulta evidente che «il fallimento del programma hilbertiano implica […] una falsificazione del ruolo di *onnipotenza fondativa* attribuito alla logica»,[193] e che pertanto si rende necessaria una forma alternativa di fondazione teorica che non sia assoluta, come voleva Hilbert, ma, al contrario, condizionata; non un "autofondazione", dunque, bensì una forma di "eterofondazione" che rimandi alla dimensione extra-logica dell'ontologia.

Nell'ultima sezione del suo saggio introduttivo ai teoremi di incompletezza, Galvan prende in considerazione un'importante obiezione secondo la quale, in un approccio fondativo condizionato quale quello appena prospettato, il fondamento verrebbe ad essere meno certo rispetto a ciò che deve essere fondato: ciò implicherebbe, pertanto, un circolo vizioso. Tuttavia, la soluzione viene dalla distinzione tra due tipologie di rapporto fondativo che risale già al pensiero antico:

- l'«ordine epistemico della conoscenza», per il quale ciò che deve essere fondato possiede un grado di certezza minore rispetto a ciò che ne deve costituire il fondamento conoscitivo, ovvero la ragion sufficiente.
- l'«ordine aletico della spiegazione», per il quale ciò che deve essere fondato può essere già noto con un grado di certezza maggiore rispetto a ciò che ne deve costituire il fondamento esplicativo.[194]

Riprendendo un ragionamento più volte illustrato, si consideri che il dominio di interpretazione, o modello, che soddisfa gli assiomi di una teoria rappresenta un contenuto astratto rispetto a quest'ultima, sebbene formalizzabile all'interno di un'estensione teorica più potente.

matematiche molto elementari, non per altre astratte e ideale; (ii) il meccanismo di selezione assicura verità empiriche e non matematiche […].» (Cfr. *ivi*, pp. 169-171).

[192] Cfr. S. Galvan, *Introduzione ai teoremi di incompletezza*, cit., p. 193. Il corsivo è dell'autore.

[193] Cfr. *ivi*, p. 196. Il corsivo è dell'autore.

[194] Come sottolinea Galvan, «è noto, poi, che secondo il modello antico tra i due momenti esiste una tendenziale armonia, almeno nel senso che il fondamento esplicativo è, in forza della sua evidenza (garantita dalla nostra capacità di intuizione delle essenze), anche sempre fondamento epistemico. È noto, altresì, che nel corso della storia moderna ha luogo la rottura di tale simmetria, ad opera, prima, del nuovo metodo ipotetico-deduttivo delle scienze sperimentali e del nuovo metodo assiomatico, poi. La rivoluzione scientifica moderna non ha però comportato la perdita della distinzione dei due momenti, anzi ha potentemente favorito il processo di autonomizzazione dell'uno rispetto all'altro» (Cfr. *ivi*, pp. 199-200).

Mediante tale formalizzazione, la consistenza della teoria viene dimostrata nella metateoria. Ma per dimostrare la consistenza di quest'ultima, è necessario esplicitarne il corrispettivo modello in una metateoria superiore. Seguendo la catena dei potenziamenti, vediamo dunque che ciò che deve essere fondato, ovvero il modello di una data teoria, richiede l'elaborazione di una metateoria il cui modello è più complesso e meno noto rispetto al primo. Ora, è chiaro che, se si abbandona la prospettiva conoscitiva (epistemica, o giustificazionista), l'obiezione per la quale una fondazione condizionata sarebbe circolare perde di forza. Infatti, se ci si pone dal punto di vista esplicativo (o aletico), non risulta più necessario che ciò che è deputato a costituire il fondamento del nostro sapere debba essere più semplice e noto rispetto a quest'ultimo. Al contrario, come suggerisce la lezione heideggeriana sull' etimologia greca della parola "verità" (*alétheia*), secondo la quale quest'ultima contiene in se stessa l'idea di un "dis-velamento", e quindi di un processo continuo attraverso il quale la verità si manifesta e si nasconde, l'ordine aletico della spiegazione richiede che il fondamento non sia né immediato (come dato in sé e per sé evidente) né mediato (ché, altrimenti, non sarebbe davvero fondamento), bensì costituisca una dimensione che trascende questi due momenti e ne costituisca la condizione di possibilità: una dimensione all'interno della quale il pensiero è libero di agire. Queste considerazioni ci consentono di sostenere, fino a prova contraria, una posizione secondo la quale il funzionamento della mente, almeno per quanto riguarda la scoperta matematica, anziché essere chiuso, risulta essenzialmente aperto e creativo.

Capitolo 8
Élenchos e l'infinito dell'anima

8.1 *Élenchos* e il realismo ontologico

Relativamente al saggio di Lucas del 1968 con cui risponde a Benacerraf, abbiamo già rilevato l'importanza del carattere dialettico della sua argomentazione. In quella sede, Lucas fa ricadere l'onere della prova di consistenza sul meccanicista, il quale, potremmo dire, avanza la sfida. In realtà, a noi pare chiaro che, da un punto di vista non soltanto mentalista, ma anche del senso comune, la posizione meccanicista costituisce l'antitesi rispetto alla tesi secondo la quale la coscienza esiste e costituisce un fenomeno dal quale non possiamo davvero prescindere nel tentativo di fornire un quadro esplicativo completo della realtà in cui viviamo. E il punto focale della posizione di Lucas, per quanto contenga alcuni errori logici, e nonostante assuma implicitamente dei presupposti non strettamente giustificabili dal punto di vista formale, consiste nel mettere in luce come sia proprio a partire dall'antitesi posta dal meccanicista che emerge, in ultima analisi, la verità della tesi: la differenza tra la mente umana e una macchina. Lì abbiamo anche già sottolineato una seppur lieve analogia tra lo schema argomentativo di Lucas e quello che, invece, struttura la figura di *élenchos*, oggetto centrale di questo capitolo.

Nel saggio intitolato *Sulla attualità del concetto di "anima"*,[195] Paolo Pagani evidenzia che l'assunto implicito nelle argomentazioni anti-meccaniciste di Lucas e di Penrose sia quello della «capacità umana di sopportare incompletezza semantica, e quindi di istituire procedimenti non-finitari o non-procedurali».[196] Dal punto di vista della logica formale, la nozione di incompletezza semantica va riferita al carattere aperto del processo di costruzione dei modelli, e quindi di dimostrazione della consistenza delle teorie formali. In proposito, l'autore afferma:

> Qui il punto è il trascendentale: la ragione della assenza nella macchina di autotrascendimento coscienziale è appunto l'assenza del trascendentale. La macchina può trascendere, non se stessa, ma il linguaggio cui dà luogo, procedendo meta-discorsivamente; se non che, il "meta" è solo il surrogato monodimensionale del trascendentale.[197]

[195] Cfr. P. Pagani, *Sulla attualità del concetto di "anima"*, in «Aquinas», LVI, 2013, no. 2, pp. 425-439.

[196] Cfr. *ivi*, p. 433.

[197] Cfr. *ivi*, p. 434.

Il concetto di monodimensionalità rende molto bene l'idea per la quale una macchina, relativamente alle sue capacità di dimostrazione e di estensione definitoria di una teoria, ha a disposizione solamente una procedura meccanica e lineare, la quale, mediante gödelizzazione, consente di costruire l'enunciato indecidibile del sistema e di aggiungerlo agli assiomi di quest'ultimo, costruendone così un'estensione più potente. La creatività del potere dimostrativo della mente umana, invece, sembra dipendere dal suo procedere attraverso vie alternative e parallele. E abbiamo anche già notato che il parallelismo dei processi computazionali di una macchina connessionista non ha niente a che vedere con la pluridimensionalità del pensiero umano; esso, cioè, è soltanto apparente, in quanto è effettivamente riducibile alla linearità seriale dei calcoli eseguibili da una macchina di Turing. Il calcolo computazionale segue delle regole finitarie, come quella della bivalenza, mentre la mente umana è in grado di adottare dei principi infinitari, come quello del "terzo escluso" o di "non contraddizione" e di concepire nozioni astratte. Un esempio che abbiamo già incontrato è quello della serie dei numeri naturali. Da un lato, un calcolatore è sì in grado di produrre meccanicamente tutta la serie infinita dei termini per i numeri naturali, ma lo può fare soltanto in potenza, ovvero soltanto a condizione di disporre di un tempo illimitato, e, per così dire, restando appiattito come tra due binari, all'interno di un'unica dimensione lineare. Dall'altro lato, al contrario, la mente umana pare in grado di abbracciare la struttura "planare" dell'insieme infinito dei numeri naturali. Questa divergenza è confermata anche da Galvan quando afferma l'impossibilità di formalizzare all'interno di una teoria il modello astratto dei numeri naturali. In particolare, Pagani sostiene che *élenchos* rappresenti una modalità di argomentazione fondativa non formalizzabile nei termini della logica formale, e che costituisca, perciò, il luogo teorico in cui si manifesta in modo chiaro l'irriducibilità della verità alla dimostrabilità formale. In altri termini, attraverso *élenchos* si rende evidente che il discorso logico non può fare a meno di presuppore un orizzonte ontologico come proprio fondamento ultimo, ovvero che la dimensione dell'infinitarietà, che caratterizza la struttura intuitiva di una teoria come quella sui numeri naturali, è imprescindibile come condizione di possibilità della formalizzazione teorica in termini finitari.

Nel saggio intitolato *La forma dell'*élenchos,[198] Pagani offre una lucida e dettagliata analisi della figura di *élenchos*, e mostra come essa rappresenti la forma essenziale del concetto di infinito "trascendentale". "Trascendentale" significa "non trascendibile", "inevadibile" e, come sottolinea l'autore, si riferisce ad un "luogo", od "orizzonte", rispetto a cui non è possibile collocarsi altrove. Il trascendentale, inoltre, coincide con la dimensione dell'essere in quanto tale. A questo proposito, Pagani afferma:

[198] Cfr. P. Pagani, *La forma dell'*élenchos, in «Rivista di Filosofia Neo-scolastica», CXXII, 2020, pp. 969-994.

> l'argomentare che dall'interno dell'essere progetta l'evasione da esso, e ne constata l'impraticabilità – cioè constata che la *heterotes*, rispetto all'essere come tale, può darsi solo come *tautotes* –, sembra in un certo senso capace (problematicamente) della ulteriorità autoescludentesi. Già formulare l'ipotesi della evasione presuppone infatti la previa ipotesi di un *altro*, che stia oltre l'essere e che *altrimenti* si strutturi.[199]

Élenchos consiste, perciò, in un progetto, in un'ipotesi, che si esclude da sé in quanto non può che riconoscere la propria impossibilità. La possibilità dell'esistenza di un luogo oltre l'essere, e quindi altro da esso, infatti, è soltanto esprimibile, ma non anche pensabile. In altre parole, il contenuto del pensiero che pensa il totalmente altro dall'essere (il non-essere assoluto), è un contenuto vuoto. «L'essere in quanto propone una *heterotes* che gli sia relativa, e ne supera l'ipotetica autonomia, è il pensiero».[200] Il pensiero va inteso, dunque, come una dimensione imprescindibile dell'essere che ne rispetta l'analogicità. Ogni determinazione finita, o, il che è lo stesso, ogni differenza, non è mai assoluta, bensì sempre e soltanto relativa, cioè commisurata al pensiero che la comprende all'interno di una dimensione infinita. In questo movimento di auto-differenziazione e auto-comprensione, in questo "differire" delle determinazioni che trapassa in un loro "conferire", consiste la natura dialettica del pensiero. Non è dunque possibile concepire qualcosa che stia al di fuori del pensiero, come non è possibile esprimere qualcosa se non attraverso il linguaggio (inteso in ogni sua forma).

Inoltre, le strutture semantiche, sintattiche e pragmatiche del linguaggio vanno intese anch'esse come dimensioni trascendentali dell'essere in quanto linguaggio. «Queste emergono quali costanti apofantiche ma, appunto per questo, devono essere considerate quali costanti ontologiche. L'essere in quanto tale, infatti, si rende noto attraverso l'apofansi».[201] Sullo sfondo di questa considerazione, naturalmente, è presente la lezione del filosofo britannico John L. Austin intorno teoria degli S*peech Acts*, ripresa poi da John R. Searle, che mette a tema la distinzione fondamentale, interna ad ogni atto linguistico, tra la dimensione locutoria, ovvero la parte proposizionale sintattico-semantica, e quella illocutoria, che rimanda alla sfera pragmatica e intenzionale. Quest'ultima è

---

[199] Cfr. *ivi*, p. 970. Il corsivo è dell'autore.

[200] Cfr. *ibidem*.

[201] Cfr. *ivi*, p. 971. Per apofansi si intende il discorso enunciativo, o assertivo. La dimensione apofantica del linguaggio, perciò, riguarda la distinzione semantica tra verità e falsità. «Tra le costanti apofantiche che possono essere difese elencticamente vi sono strutture semantiche, sintattiche, pragmatiche. Semantiche sono quelle costanti la cui presenza consente all'apofantico di avere un contenuto effettivo, ovvero determinato, e di essere così una manifestazione ontologica. Sintattiche sono quelle costanti che assicurano la coerenza e la consistenza del locutorio. Pragmatiche sono quelle costanti che garantiscono la collocazione del locutorio entro un determinato registro illocutorio e, più in generale, comunicativo» (Cfr. *ivi*, p. 978).

particolarmente importante in quanto determina la concretezza dell'atto linguistico, ovvero il suo dipendere dal contenuto intenzionale del parlante e dallo specifico contesto linguistico in cui esso si trova. Esistono tipi diversi di illocuzione che variano a seconda degli stati intenzionali del parlante, come ad esempio la credenza (o asserzione), la promessa, l'ordine, ecc. Ciascuno di essi implica un determinato impegno pragmatico-intenzionale da parte del parlante nei confronti delle condizioni di verità del contenuto proposizionale; in particolare, l'asserzione implica un impegno alla verità di quanto viene asserito. Inoltre, la differenziazione tra locutorio e illocutorio è ciò che consente di distinguere anche tra la contraddizione "in termini", pertinente cioè alla sola dimensione proposizionale, e la contraddizione "performativa", che avviene, invece, quando viene esplicitata la contraddizione fra le due componenti (l'*actus signatus* e l'*actus exercitus*), mediante la proiezione della seconda sulla prima, ossia l'esplicitazione della dimensione intenzionale su quella proposizionale.[202] Questo tipo di contraddizione, poi, si dice "strutturale", quando costituisce una «violazione di condizioni strutturali dell'atto di discorso in quanto tale».[203] Tuttavia, il momento della contraddizione performativa non deve essere confuso con quello della fondazione elenctica, poiché, infatti, mentre la prima costituisce soltanto una fase di transizione dialettica e finisce con l'identificarsi o «come intenzionale autosottrazione all'orizzonte noetico-linguistico […], o come atto linguistico mancato», la seconda, invece, coincide con il «riconoscimento della intrascendibilità degli elementi negati».[204] Sulla base di questo sfondo teorico, Pagani riformula la struttura elenctica scandendola in due momenti: quello della «negazione esplicitamente formulata (NF)» e quello della «negazione effettuale (NE)»:

> (a) la negazione esplicitamente locutoria delle costanti apofantiche si realizza all'interno di quelle stesse costanti: essa cioè fa ciò che nega, ovvero performativamente non riesce a evadere da ciò che locutoriamente rifiuta; (b) il progetto di una negazione effettuale delle costanti apofantiche coincide con l'ipotesi di una illocuzione priva di contenuto locutorio: una pura intenzione di dire, cui non può corrispondere niente di effettivamente detto.[205]

Se la negazione esplicita delle costanti apofantiche trova una formulazione, ma è costretta ad ammettere quelle stesse costanti, a meno di non degenerare in un atto linguistico banale e senza alcun significato, la negazione effettuale, al contrario, non può nemmeno realizzarsi sul piano linguistico,

[202] Cfr. la nota no. 6 a piè pagina in *ivi*, p. 975. La proiezione consiste nell'esplicitazione delle intenzioni illocutorie sulla componente locutoria.

[203] Tali condizioni, afferma Pagani, «riguardano: *vel* (a) le particolari forme dell'illocutorio *vel* (b) l'atto di discorso in quanto tale» (Cfr. *ivi*, p. 974).

[204] Cfr. *ivi*, 975-977.

[205] Cfr. *ivi*, 977-978.

né tantomeno su quello «noetico», ovvero del pensiero. Quest'ultima, perciò, si auto-esclude dall'orizzonte del trascendentale, ovvero dal luogo in cui tutte e sole le determinazioni dell'essere trovano la propria sede. In altre parole, la negazione effettuale è destinata ad un esito estremo, ovvero alla totale *aphasia*, ad un silenzio assoluto.[206] Ciò deriva dalla completa astrazione della componente illocutoria di NE rispetto alla componente locutoria, ossia dell'intenzione rispetto all'esecuzione. Come il suo contenuto inteso risulta non solo indicibile, ma anche impensabile, così anche la sua esecuzione non si riduce che ad un nulla.

Abbiamo visto che *élenchos* costituisce una forma di fondazione indiretta, ossia di "eterofondazione", mediante la quale viene alla luce l'impossibilità di un trascendimento della dimensione dell'essere, ovvero di una sua progettuale negazione sul piano intenzionale. Ciò è di estrema importanza per comprendere ciò che affronteremo nel prossimo paragrafo in relazione agli esiti del programma di fondazione della logica.

8.2 L'impossibilità di una autofondazione della logica per via elenctica

Nel libro intitolato *Non contraddizione e terzo escluso*,[207] in particolare nel capitolo settimo – dal titolo «Per una filosofia delle logiche» –, Galvan si concentra sul tema dell'eterofondazione della logica, ovvero della sua fondazione a partire dal contenuto d'evidenza, e analizza quella figura argomentativa centrale nella storia del pensiero filosofico che è *élenchos*.

Come spiega l'autore, prima della nascita, nel Novecento, delle logiche non standard, come ad esempio quella intuizionistica, ha prevalso la credenza nella validità della logica classica e nella sua imprescindibilità per l'esercizio della ragione in forza dell'appello ad una forma di intuizione immediata, ovvero mediante una «giustificazione *via evidentiae*».[208] La struttura dell'argomentazione elenctica che informa implicitamente tale credenza costituisce la pretesa per la quale «la tesi segue dalla sua stessa negazione senza il ricorso ad altro che alle regole essenziali alla istituzione del gioco dialettico tra proponente ed opponente della tesi stessa».[209] È per questo motivo che, tradizionalmente, *élenchos* è stato considerato l'unica vera modalità di autofondazione logica. Tuttavia, Galvan sostiene che questa conclusione sia errata, poiché in questo caso non è legittimo parlare di un'autofondazione, bensì di un'eterofondazione facente perno su contenuti extra-logici. A

[206] Cfr. *ivi*, p. 979.
[207] Cfr. S. Galvan, *Non contraddizione e terzo escluso. Le regole della negazione nella logica classica intuizionistica e minimale*, FrancoAngeli, Milano 1997[6].
[208] Cfr. *ivi*, p. 124.
[209] Cfr. *ivi*, p. 125.

fini dimostrativi, egli avanza perciò una proposta di formalizzazione in termini logici della fondazione elenctica del principio di non contraddizione (per comodità lo abbreviamo con NC) che egli chiama «calcolo *e*», e mostra, in seconda battuta, che una simile procedura non costituisce un metodo di autofondazione conclusivo.

In primo luogo, per la riuscita dell'argomentazione elenctica, è necessario che venga definito un «comune terreno d'intesa» tra colui che intende negare il NC (ovvero lo «scettico»), e colui che vuole invece confutare il negatore («l'opponente dello scettico»). Il calcolo *e* contiene le regole di base condivisibili da entrambe le parti e comuni ai diversi tipi di logica: quella classica, quella intuizionistica e quella minimale. Come punto di partenza si consideri la «regola di *autocontraddizione*» (AC):

$$\frac{X\ \alpha \vdash \neg\alpha}{X \vdash \neg\alpha}$$

Dal momento che NC è un principio logico incondizionato, è possibile riformulare AC prescindendo dall'insieme di assunzioni X di una teoria. Inoltre, sostituendo α con τ (la tesi dello scettico), si ottiene la «regola di *autocontraddizione elenctica*» (AE):

$$\frac{\tau \vdash \neg\tau}{\vdash \neg\tau}$$

Oltre ad AE bisogna assumere anche tutte quelle regole che assicurano «quella *condizione minimale di significanza* del discutere a cui neppure lo scettico può sottrarsi».[210] A tal scopo, è necessario assumere le regole appartenenti alla «base comune b» proprie delle tre diverse tipologie di logica sopra menzionate, e inoltre la regola E∀ e la regola S (di sostituzione proposizionale). Infine, lo scettico deve soddisfare anche «una *condizione minimale di coerenza*; non certamente una condizione basata sul rispetto del principio di non contraddizione – che lo scettico si propone al contrario di negare –, ma quella richiesta dall'esigenza per lo scettico (1) di impegnarsi su tutto ciò che è *implicato* dalla sua tesi – il darsi della contraddizione – e (2) di non accettare niente che sia *incompatibile* con essa. […] Nelle parole dello scettico: affermo tutto ciò che è condizione necessaria della mia posizione e nego tutto ciò che la elimina».[211] (1) è formalizzabile nei termini di S1:

[210] Cfr. *ivi*, p. 127. Il corsivo è dell'autore.
[211] Cfr. *ivi*, p. 128. Il corsivo è dell'autore.

$$\frac{\tau \vdash \alpha}{\vdash \alpha}$$

Mentre (2) lo è nei termini di S2:

$$\frac{\alpha \vdash \neg\tau}{\vdash \neg\alpha}$$

Galvan sottolinea che S2, la quale definisce la negazione all'interno della teoria scettica, non va confusa con la regola generale della negazione minimale (¬j), la quale nega ciò che implica contraddizione. Infatti, lo scettico ammette proprio la possibilità della contraddizione in generale, sebbene si impegni però a negare ciò che contraddice τ. Inoltre, si noti che S2 contiene AE, nel senso che quest'ultimo rappresenta una sua specificazione.

Una volta definito il terreno comune di argomentazione, è necessario mostrare che dal calcolo *e* deriva ¬τ. Questo, tuttavia, dipende dalla formulazione di τ, e ciò è possibile soltanto per la formulazione "generale" della tesi scettica, non altrettanto per la sua formulazione «locale, l'unica che ha un effettivo interesse […] anche nell'ottica dello scettico».[212] Nel caso della formulazione generale, τ coincide con l'affermazione della contraddittorietà di "tutte" le proposizioni esprimibili nel linguaggio del calcolo *e*. La dimostrazione di ¬τ, come la illustra l'autore, procede come segue:[213]

$\forall p(p\wedge\neg p) \vdash \forall p(p\wedge\neg p)$	per assunzione
$\forall p(p\wedge\neg p) \vdash p\wedge\neg p$	per E∀
$\forall p(p\wedge\neg p) \vdash \neg p$	per E∧
$\forall p(p\wedge\neg p) \vdash \neg\tau$	per S
$\tau \vdash \neg\tau$	per def. di τ
$\vdash \neg\tau$	AE

Questo risultato, come è stato anticipato, non consente di confutare in maniera effettiva la tesi scettica, poiché, infatti, la formulazione locale di τ afferma che ad essere contraddittorie non sono "tutte" le proposizioni (il che comprenderebbe la stessa τ), ma solo alcune o, al limite, tutte tranne quella stessa. Per ottenere la contraddizione locale occorrerebbe «ottenere $\vdash_e \neg(p \wedge \neg p)$, ma ciò è impossibile; per ottenere ¬τ è necessario qualcosa di aggiuntivo rispetto al calcolo *e*». «Ciò significa»

[212] Cfr. *ivi*, p. 130.
[213] Cfr. *ivi*, p. 131.

– conclude Galvan – «che neppure a proposito del principio di non contraddizione si può a rigore parlare di autofondazione, vale a dire di fondazione incondizionata […]. Dalla autofondazione si deve, dunque, passare a qualche forma *condizionata* di *eterofondazione*».[214]

A questo proposito, Pagani giustamente osserva:

> La capacità autofondativa non può appartenere ad alcun principio, se questo è inteso in chiave logico-formale – cioè come semplice regola di calcolo […]. La figura dell'autofondazione, piuttosto, non può che realizzarsi in ambiente tematicamente ontologico, dove il PDNC [il principio di non contraddizione] può essere considerato per quel che esso è originariamente e concretamente: e cioè come l'essere stesso in quanto trascendentalmente coerente con sé.[215]

La prospettiva dell'autentico fondamento, in altri termini, è da ricercarsi al di fuori della sfera logica, ovvero all'interno della dimensione ontologica dell'essere, prestando attenzione però al fatto, messo in luce da Pagani, per il quale quest'ultima non va intesa come esterna, o meglio estranea, alla prima, ma piuttosto ne è come il suo alveo naturale e analogico. La struttura trascendentale, inevadibile, dell'essere è, infatti, condizione di possibilità e di esistenza di tutte le differenti determinazioni che si collocano al suo interno. Su questo punto concorda, del resto, lo stesso Galvan quando rileva che quella ontologica «tradizionalmente era considerata la dimensione corrispondente al punto di vista dell'intero, la prospettiva […] dalla quale si colgono gli aspetti più generali e strutturali della realtà».[216]

Ora, i contenuti extra-logici su cui si basa l'eterofondazione della logica rimandano ad un determinato impianto semantico di una teoria, il quale, a sua volta, dipende dall'assunzione di una specifica prospettiva ontologica. Del resto, questo è un punto su cui abbiamo cercato di portare l'attenzione lungo tutto il corso del presente lavoro. Il dibattito intorno all'equiparabilità tra la mente e la macchina, non può dirimersi esclusivamente sulla base del punto di vista della logica della dimostrazione, ma presuppone, in aggiunta, un preliminare impegno riguardo l'accettazione o il rifiuto di determinate considerazioni di natura filosofica relativamente alla realtà di certe tipologie di enti o di fenomeni, come quella dei numeri naturali o degli stati mentali coscienti.

Come rileva Galvan, esistono due tipologie ontologiche fondamentali: il realismo e il costruttivismo. Egli si dichiara a favore del realismo e propone una serie di argomenti volti a mostrare alcuni aspetti di questa posizione dai quali, né la semantica sottostante alla logica classica né quella

---

[214] Cfr. *ivi*, pp. 132-134. Il corsivo è dell'autore.
[215] Cfr. P. Pagani, *La forma dell'*élenchos, cit., p. 989.
[216] Cfr. S. Galvan, *Non contraddizione e terzo escluso*, cit., p. 135.

che sta alla base del costruttivismo, possono, in ultima istanza, prescindere. Innanzitutto, il realismo poggia sulle seguenti due tesi:

- Una tesi fondamentale, comune a tutte le ontologie realiste, per la quale esiste un piano della realtà che è totalmente indipendente da quello del pensiero, ovvero che non è prodotto da quest'ultimo. Nell'ottica della trascendentalità della struttura del pensiero, o eidetica, è possibile quindi concepire l'oggetto del pensiero: da un lato, come essere in sé; dall'altro, come essere intenzionale in quanto pensato da un soggetto.
- Una tesi secondaria diversa per ogni forma di realismo, che stabilisce «quanto sia estesa, rispetto al pensiero, la sfera dell'essere in sé».[217]

Il costruttivismo mantiene la distinzione tra pensiero e pensato, ma concepisce il secondo come un prodotto del primo, e quindi dotato di solo essere intenzionale. Tuttavia, la stessa distinzione che il costruttivismo ammette «discende […] dalla tesi realistica fondamentale sopra esposta, secondo la quale il pensiero non può costituire se stesso e, perciò, *almeno* il pensiero presenta una forma d'essere che non coincide con l'essere pensato».[218] Inoltre, e qui torna un argomento già incontrato nei capitoli precedenti seppure in altra veste, Galvan evidenzia che ricondurre – alla maniera dei costruttivisti – la nozione di verità a quella dimostrabilità, espone alla difficoltà, messa particolarmente in luce dai teoremi limitativi gödeliani, per la quale, lungo la progressiva costruzione di metateorie sempre più potenti, si giunge prima o poi a riscontrare l'esistenza di un insieme di proposizioni che le nostre capacità dimostrative non consentono di decidere, ripresentando in questo modo una forma di dualismo propria della logica classica e dell'ontologia realistica. È evidente, perciò, che la tesi fondamentale del realismo sta al fondo anche dell'ontologia costruttivista.

Al fine di mostrare la legittimità di tale tesi, Galvan propone un argomento per assurdo. Se il costruttivista rifiutasse anche la basilare distinzione tra pensiero e pensato, allora egli sarebbe costretto ad ammettere un'unica forma di realtà intenzionale all'interno della quale si darebbe l'insieme dei contenuti di pensiero, ovvero degli universali astratti e degli oggetti determinati che ne costituiscono l'estensione. Tuttavia, tenendo sullo sfondo la celebre lezione hegeliana sul trapasso di ogni determinazione nel suo opposto, ciò equivarrebbe ad accettare un'essenziale indeterminatezza del reale come una proprietà strutturale di esso, ovvero come una sorta di «*continuum*»; il che condurrebbe, in ultima istanza, ad ammettere l'esistenza ontologica della contraddizione. Ciò, non solo «dovrebbe comportare il rifiuto della logica intuizionistica […], ma urterebbe contro il solido

[217] Cfr. *ivi*, p. 152.
[218] Cfr. *ivi*, p. 153. Il corsivo è dell'autore.

dato d'evidenza che la contraddizione non possa appartenere all'ordine dell'essere reale».[219] A questo proposito, Pagani giustamente osserva che l'impossibilità di un'autofondazione in termini logico-formali del principio di non contraddizione, «è qualcosa che non intacca il senso proprio di *élenchos*».[220] L'argomento elenctico, infatti, trae la sua piena forza dal rimando allo sfondo trascendentale dell'ontologia, di cui rappresenta la manifestazione formale sul piano noetico.

Da ultimo, per superare la nota difficoltà epistemologica di eredità kantiana riguardante la conoscibilità della realtà in sé, Galvan sostiene che occorra abbracciare una concezione per la quale il reale non è qualcosa di «*attualmente*» inaccessibile, ma, al contrario, di strutturalmente «*disposizionale*». Ciò significa che non si dà una singola e privilegiata modalità d'accesso alla realtà, bensì differenti "aperture" sul mondo legate ai particolari punti di vista che i soggetti conoscenti incarnano; aperture che, in forza dell'unicità referenziale[221] del reale, ne rivelano determinate modalità strutturali.

8.3 L'infinito trascendentale e l'anima

A sostegno della necessità di oltrepassare la logica formale in direzione della riflessione ontologica, e per chiarire meglio che cosa si intenda per visione "disposizionale" della realtà, è fondamentale tornare a prendere in considerazione il saggio intitolato *Sulla attualità del concetto di "anima"*. Il nucleo centrale della riflessione consiste nel mettere in luce come la marca che contraddistingue l'essere umano in quanto tale è una forma di capacità infinita e trascendentale che costituisce la dimensione interiore dell'io.

Riproponendo una categorizzazione tradizionale, è possibile distinguere tra quattro diverse figure dell'infinito: l'infinito potenziale; il transfinito; l'infinito di perfezione; l'infinito trascendentale. L'infinito potenziale costituisce qualcosa che l'intelletto è in grado di cogliere, per l'appunto, soltanto in potenza e mai in atto, e, in un certo senso, delimita il dominio di azione della mente, intesa come ciò che è capace di misura (*mensura*) e di calcolo (*ratio*). Come nota Pagani, ogni possibile determinazione di questa categoria si dà come finita e incompleta, mai «*tota simul*», e quindi costituisce sempre qualcosa di incrementabile e indeterminato. Un esempio è rappresentato dalla

[219] Cfr. *ivi*, p. 155.
[220] Cfr. P. Pagani, *La forma dell'*élenchos, cit., p. 994.
[221] «La realtà in sé non è nient'altro che quel qualcosa di referenzialmente unico che a seconda del punto di vista assunto manifesta queste o quest'altre proprietà, ma che in sé non è caratterizzato da un insieme di predicati alternativi – rispetto a quelli che si manifestano a noi – ed inaccessibili alla conoscenza» (Cfr. S. Galvan, *Non contraddizione e terzo escluso*, cit., p. 156).

successione dei numeri naturali. Questo tipo di infinità è riproducibile in forma di processo meccanico all'interno di un sistema formale per l'aritmetica, o da una macchina di Turing mediante dei processi lineari, ma non è mai totalmente "catturabile" (tutt'al più è "inseguito") dai mezzi espressivi di un linguaggio formalizzato. La nozione di transfinito è stata sviluppata da Cantor in risposta all'idea kantiana per la quale ciò che delimita il dominio della finitezza è l'assoluto, ovvero l'infinito attuale. Come rileva Gianfranco Basti, «questo sarebbe vero se esistesse un solo modo di considerare l'infinito attuale, ma per Cantor ciò non è vero: esiste […] anche l'infinito *transfinito*. Ovvero, l'infinito *determinato*, specificato entro un limite, eppure *incrementabile*, ovvero dove non tutti i termini della serie sono *attualmente definiti (enumerati)*».[222] Un insieme infinito i cui elementi stiano in corrispondenza biunivoca con gli elementi dell'insieme $\mathbb{N}$ dei numeri naturali, si dice avere «la *potenza del numerabile*», ovvero «lo stesso *numero cardinale infinito* di elementi».[223] Il transfinito, dunque, è un infinito attuale, determinato rispetto alle proprietà essenziali che lo definiscono, ma indeterminato rispetto all'enumerazione dei suoi elementi, la quale è incrementabile *ad infinitum*.[224] L'infinito attuale tradizionale è invece riferibile al concetto di Assoluto, e denota la perfezione ontologica: ciò di cui non si può pensare alcunché che gli sia superiore. Esso è perciò determinato e non passibile di incremento. Come evidenzia Pagani, l'infinito trascendentale, invece, a differenza di quelli precedenti, denota un concetto del tutto peculiare, che definisce un aspetto essenziale dell'anima, e costituisce la «condizione di costruibilità» sul piano noetico delle altre nozioni di infinito. In altri termini, esso rappresenta la capacità infinita dell'anima di abbracciare in un luogo interiore inesteso la potenziale infinità delle determinazioni finite. Per questo motivo, l'infinito trascendentale è indeterminato, in quanto differente rispetto ad ogni determinazione finita, e non incrementabile, in quanto dimensione "limite" ed onnicomprensiva di ogni possibile determinazione.

In proposito sarebbe opportuno riconsiderare alcuni elementi chiave dell'antropologia tommasiana, la quale a sua volta si rifà a quella aristotelica. In questa sede ci limiteremo ad illustrarne soltanto alcuni tra quelli di maggiore interesse. In primo luogo, Tommaso concepisce l'umano come un composto di corpo e anima, in cui è l'anima a dare forma al corpo secondo le possibilità che le sono concesse da quest'ultimo. L'anima, perciò, non è qualcosa di riconducibile al corpo né va intesa, in senso materialistico, come qualcosa di spiegabile in termini fisici. Del resto, ciò è attestato in modo evidente dalla capacità intellettiva propria dell'anima di trascendere il piano della *physis*, ovvero di non essere vincolata (all'interno di una relazione necessaria di causa ed effetto) al mutamento condizionato da fattori materiali. In secondo luogo, tale capacità va concepita nei termini di un

[222] Cfr. G. Basti, *Le radici forti del pensiero debole*, cit., p. 186. Il corsivo è dell'autore.
[223] Cfr. *ivi*, p. 187. Il corsivo è dell'autore.
[224] Gli insiemi che hanno la medesima cardinalità di $\mathbb{N}$ si dicono transfiniti del primo ordine.

“vedere” riferito all’orizzonte dell’essere, di un “attestare” che qualcosa si dà come presente nella nostra interiorità. «Conoscere è diverso da essere modificati o da reagire a uno stimolo fisico. Conoscere è “aver presente”: è accogliere, dar luogo, trattenere […]. Non c’è una conoscenza senza soggettività, cioè senza una capacità di aver presente, di immedesimarsi con ciò che si dà».[225]

La dimensione fisica è quella a cui appartiene una macchina, o un calcolatore, il quale opera soltanto all’interno dei limiti imposti da essa. Come sottolinea Searle, un *computer*, in sé, non possiede nemmeno quella basilare natura logico-sintattica secondo cui, normalmente, riteniamo che esso operi. La sintassi delle sue computazioni, cioè, è un aspetto che emerge soltanto in relazione a un nostro atto di osservazione. Al contrario, noi intendiamo l’“essere soggetto”, nel senso autentico del termine, come l’avere presente sia il piano della sintassi che quello della semantica, ovvero come l’essere in grado di cogliere un orizzonte di significato che si manifesta attraverso l’evidenza. La soggettività, cioè, si configura come intenzionalità, ovvero come una forma di «disposizione» conoscitiva per la quale il soggetto intrattiene una relazione con ciò che si presenta allo sguardo dell’intelletto. L’intenzionalità presuppone un’articolazione del soggetto in oggetto intenzionato e agente intenzionante: «l’io non è semplice coincidenza con sé, ma è anche coincidenza con l’orizzonte della presenza». «Ora, il fatto di essere entrambe le cose implica nell’io due dimensioni: in primo luogo, esso è nell’orizzonte, e in questo senso è un ente […]; in secondo luogo, proprio per poter concepire tutto ciò […] l’io deve aderire all’orizzonte stesso».[226] L’idea per la quale il soggetto aderisce all’orizzonte infinito della presenza è particolarmente significativa, e riteniamo che sia utile concepirla attraverso l’immagine di una sfera come rappresentazione della totalità dell’essere. Da questa prospettiva possiamo interpretare l’estensione interna alla sfera come il dominio delle specifiche determinazioni, o degli enti finiti, di cui l’io, come soggetto individuale, è parte. La superficie perimetrale della sfera, invece, sia in quanto zona liminale (al di là della quale, stando alla rappresentazione suggerita, nulla è dato) che in quanto dimensione limite (ovvero inestesa, in quanto delimitante l’intero dell’estensione), rappresenterebbe l’orizzonte dell’infinito trascendentale che accoglie in sé la totalità degli enti. L’aderenza da parte dell’io a tale orizzonte, allora, va intesa come un co-incidervi “dal di sotto”, in modo tale che l’io sia, al contempo, parte “limitata” del dominio dell’estensione, sia “limite” inesteso del dominio stesso.

L’intenzionalità «è una dimensione di me, e, in particolare, è la relazione del mio io con la manifestazione dell’essere in quanto tale», per cui «tutto ciò che accade a me […] è anche una manifestazione intenzionale».[227] Il riferimento, qui, è alla tradizione scolastica, e in particolare al

[225] Cfr. la nota no. 4 in P. Pagani, *Sulla attualità del concetto di “anima”*, cit., p. 426.
[226] Cfr. *ivi*, p. 428.
[227] Cfr. *ivi*, p. 429.

termine latino *omnitudo*, che indica la capacità infinita propria dell'anima di abbracciare con l'intelletto tutte le cose. L'infinito trascendentale, perciò, costituisce quella apertura intrascendibile e analogica dell'essere sull'essere stesso che contraddistingue la soggettività umana.

Coerentemente con quanto abbiamo finora esposto circa il carattere di infinitarietà che appartiene alla natura dell'anima e dell'intelletto umano, in merito al dibattito intorno alla riproducibilità delle mente in termini computazionali, Pagani conclude evidenziando il risultato fondamentale che abbiamo cercato di mettere in luce in questa sede, ovvero che «la mira teorica delle riflessioni post-gödeliane di Lucas e Penrose sembra essere la seguente: conoscere non equivale a derivare».[228] Come abbiamo visto, anche in ambito matematico, la mente sembra procedere facendo appello a principi e metodi infinitari, che implicano il riferimento ad un orizzonte, o modello, semantico di significato. La capacità di aderire a questo orizzonte, ovvero di avere presente attraverso l'evidenza il modello semantico dei numeri naturali che la logica della dimostrazione non è in grado di formalizzare, coincide con quell'apertura trascendentale dell'anima che è l'alveo all'interno della quale trova spazio l'esercizio della ragione.

A questo punto, riteniamo opportuno riportare il seguente passo, in cui Pagani espone il quadro generale del suo discorso intorno all'anima:

> Oggi il senso comune è dominato dalla immagine dello Zarathustra nicciano, secondo cui l'uomo sarebbe un ponte tra la scimmia (si pensi al riduzionismo di certa vecchia "teoria sintetica") e l'oltre-uomo, che può essere prefigurato oramai nei termini di una raffinata macchina computazionale […]. In tale prospettiva, l'umano è evacuato di ciò che lo specifica, ed è schiacciato – *a parte ante* – sull'*animal brutum*, che è un certo *suppositum* cognitivo e affettivo, ma privo di mente; e – *a parte post* – sulla macchina computazionale, che è, se non attualmente, almeno nelle prospettazioni estremistiche che se ne fanno, una mente (almeno in quanto capacità di *mensurare*), ma priva di *suppositum* cognitivo e affettivo […]. Ora, il vantaggio di parlare di "anima" (ovvero di riscoprire – con Tommaso – la pertinenza di questo antico concetto) è quello di restituire – quando si parla dell'umano – un corpo alla *mens* e una *mens* al corpo. In altre parole, la figura dell'anima sfugge al *dilémmaton* tra biologismo e mentalismo, indicando come la fenomenologia dell'umano non possa essere captata né entro un quadro biologistico riduzionista, né entro un quadro cibernetico sovradeterminato, risultando, rispetto a entrambi, disomogenea.[229]

Qui l'autore intende difendere l'importanza, per il discorso filosofico contemporaneo, di tornare a prendere in seria considerazione concetti come quello di "anima" (che equivale a quello moderno di "soggetto"), che oggi paiono, per così dire, obsoleti. Ciò al fine di ripristinare i termini

[228] Cfr. P. Pagani, *Sulla attualità del concetto di "anima"*, cit., p. 436.
[229] Cfr. *ivi*, p. 426.

dell'essenza dell'uomo che le teorie dominanti che si sono imposte sul panorama intellettuale e culturale hanno, in un certo senso, sfumato fino quasi a dissolvere. La critica, in particolare, può essere rivolta a quella corrente di pensiero post-moderna, che va sotto il nome di "transumanesimo". Il transumanesimo – che sposa la visione fisicalista per la quale la coscienza e l'autocoscienza dell'uomo non sarebbero che configurazioni o schemi ricorrenti di informazione – sostiene che il compito dell'uomo sia quello di trascendere la propria forma fisica mediante un utilizzo della tecnologia,[230] teso al miglioramento progressivo (*enhancement*) ad al perfezionamento della specie. L'obiettivo è quello giungere ad una modalità di riproduzione, non più per via biologica, bensì per mezzo di reti di informazione all'interno delle quali l'autocoscienza individuale possa sopravvivere indefinitamente nel tempo, mediante la trasmissione della propria configurazione essenziale di informazione da un sostrato fisico divenuto obsoleto ad un altro.[231] In questo senso, il "transumano" il progetto di chi intende rendersi artefice esclusivo del proprio destino, divenendo l'unico fatto determinante della propria evoluzione. Da questo punto di vista, il successore prossimo all'uomo lungo la scala evolutiva può essere rappresentato da una forma di autentica intelligenza artificiale come discendente della specie umana in una duplice direzione: da un lato, come prodotto creato dall'ingegno umano in grado di ergersi al fianco del proprio creatore quanto a prestazioni razionali; dall'altro come generazione di umanoidi (*cyborg*), collocantisi a metà tra uomo e macchina. Ma una simile visione, se ad un primo sguardo pare voler mettere in risalto l'autocoscienza come l'aspetto principale di ciò che definisce e distingue l'umano, e quindi come qualcosa da custodire mediante una sua progressiva "eternalizzazione" grazie alla tecnica, da un altro punto di vista, appiattisce l'io sul piano fisico, interpretandolo come uno schema sinaptico-informazionale riproducibile e trasmissibile indefinitamente, e finisce con il dissolvere la dimensione più propria della soggettività – la spiritualità –, intendendola alla stregua di una forma di ricorsività totalmente decifrabile e tecnicamente riproducibile.

Ciò che il concetto di postumano, o transumano, porta con sé è l'idea che ogni categorizzazione troppo rigida (che distingua tra uomo e animale, tra uomo e macchina, tra maschio e femmina, ecc.) debba essere fatta saltare, e quindi che ogni limite che determina l'uomo debba essere tendenzialmente eliminato. Tuttavia, un tale concetto non può prescindere dall'assumere pur sempre un ideale regolativo di uomo, e quindi dal reintrodurre ciò che si proponeva di eliminare: un limite che definisca un'essenza. È inevitabile implicare una certa concezione della natura umana anche da parte di chi non ne riconosca esplicitamente alcuna: è condizione di senso, per poter dare

[230] Ci riferiamo alla nanotecnologia, alla biotecnologia, alle scienze cognitive e all'informatica, come ai settori principali della ricerca scientifico-tecnologica.

[231] In proposito, si veda R. Kurzweil, *The Singularity Is Near. When Humans Transcend Biology* (2005), Duckworth, 2018.

valutazioni (giudizi di valore) che riguardino la condizione umana. La visione del transumanesimo, dunque, mentre sul piano pratico riesce soltanto ad inseguire un incessante programma di eliminazione di quei vincoli che limitano l'esistenza dell'uomo, sul piano teorico, invece, è costretta ad una continua ridefinizione, e quindi riaffermazione, di ciò che ne delimita l'essenza.

La versione mondana dell'immortalità prospettata dal transumanesimo apre ad una serie di interrogativi filosofici ed etici urgenti, dal momento che la sua vulgata, insieme a quella di altre concezioni postumanistiche e fisicalistiche, sembra essersi imposta sul panorama culturale odierno in modo assai esteso. Inoltre, la sempre più diffusa fede scientista ha fatto sì che il notevole e rapido progresso nel campo dell'intelligenza artificiale rafforzasse una convinzione per la quale, con le parole di Agazzi, «il sostituire l'uomo mediante automi anche sul terreno dell'attività pensante vera e propria non appare altro che il logico e prevedibile coronamento di questo cammino».[232] Questa convinzione è stata e continua ad essere anche al centro di un certo tipo di retorica mediatica che propone, con accenti o catastrofisti o utopistici, l'ineluttabilità di un tale coronamento. È per questo motivo che consideriamo di primaria importanza il recupero di una riflessione critica sull'essenza dell'uomo e sulla sua irriducibilità.

[232] Cfr. E. Agazzi, *Alcune osservazioni sul problema dell'intelligenza artificiale*, in «Rivista di Filosofia Neo-Scolastica», LIX, 1967, no. 1, p. 2.

Conclusioni

Un primo intento di questo lavoro è stato quello di riunire in un'unica sede alcuni tra gli elementi teorici fondamentali utili per comprendere le linee principali del dibattito tra meccanicisti ed anti-meccanicisti intorno all'equiparabilità della mente umana ad un modello computazionale: dibattito che si è sviluppato in seguito ai risultati di Gödel circa l'incompletezza della matematica e al lavoro di Turing nel campo della teoria della computazione, e che ha continuato a suscitare l'interesse dei commentatori fino ad oggi. Il secondo intento è consistito nel mettere in luce l'importanza che assumono differenti (se non opposte) credenze di tipo ontologico che stanno alla base delle argomentazioni proposte in difesa di entrambe le parti, e che di per sé costituiscono qualcosa che non può essere formalizzato dal punto di vista logico. Perciò, abbiamo cercato di mostrare come i teoremi di incompletezza, su cui poggiano i cosiddetti "argomenti gödeliani", non consentano, presi da soli, di dirimere la questione in maniera ultimativa.

Nella prima Parte del presente lavoro, a scopo meramente introduttivo, abbiamo illustrato, a grandi linee, gli aspetti-chiave del problema filosofico classico del rapporto tra mente e corpo e dell'identità tra uomo e macchina. Innanzitutto, attraverso l'analisi di un'opera di Popper, abbiamo preso in considerazione gli elementi che caratterizzano la visione del materialismo e delle versioni in cui esso è venuto a differenziarsi nel corso del tempo, oltre che alcune serie obiezioni che l'autore muove nei confronti di ciascuna di queste. Inoltre, abbiamo esposto altre due differenti posizioni sulla questione dell'emergenza della coscienza e della sua interazione con il mondo fisico: l'"interazionismo" di Popper ed il "naturalismo biologico" di Searle. Queste visioni prendono le distanze dalla visione materialista nelle sue diverse declinazioni, in quanto non solo ammettono l'esistenza degli stati coscienti, ma tentano anche di fornire una spiegazione coerente del modo in cui questi ultimi interagiscono con i processi fisici. Infine, abbiamo illustrato in che cosa consista il programma di ricerca dell'intelligenza artificiale nelle sue due versioni, che Searle definisce tesi dell'"IA forte" e tesi dell'"IA debole".

La seconda Parte, invece, è giunta al nucleo principale della nostra trattazione: il dibattito intorno agli argomenti gödeliani, avanzati da Lucas e Penrose nel tentativo di rigettare il meccanicismo, ovvero la tesi per cui il funzionamento della mente sarebbe riproducibile, se non addirittura replicabile, per mezzo di un algoritmo computabile da una macchina di Turing. Dapprima abbiamo posto le basi storico-concettuali in relazione alla cosiddetta "crisi dei fondamenti" della matematica di fine Ottocento e al programma formalista hilbertiano di inizio Novecento. Ciò è servito per comprendere il retroterra culturale in cui è venuto a delinearsi il lavoro di Gödel. Una volta esposti i teoremi di limitazione e le loro implicazioni dal punto di vista strettamente formale, siamo passati

ad analizzare nel dettaglio le argomentazioni di Lucas e Penrose, insieme alle obiezioni mosse loro da parte di Putnam, Benacerraf, Lindström e Shapiro. Naturalmente, abbiamo cercato di fornire un'esposizione il più chiara possibile, senza con ciò sacrificare il rigore che si addice alla trattazione di temi di carattere logico e formale. Da questo esame è emerso che, sebbene tali argomenti non riescano realmente a confutare l'ipotesi del meccanicismo, ciononostante essi consentono di mettere in evidenza due aspetti cruciali della questione. In primo luogo, al netto di considerazioni aggiuntive di stampo più propriamente filosofico ed extra-logico – riguardanti la realtà degli enti matematici, differenti teorie epistemiche, certe idealizzazioni circa l'operare della ragione e il rapporto intenzionale della mente con la realtà –, la tesi meccanicistica non risulta determinata a sufficienza affinché su di essa possa instaurarsi un dibattito entro il quale sostenere posizioni che abbiano la pretesa di essere conclusive e che si basino su un comune terreno di riflessione. In secondo luogo, la capacità del pensiero di comprendere e di attestare certi contenuti matematici di carattere infinitario sulla base dell'evidenza o dell'intuizione sembra essere qualcosa di irriducibile alle abilità dimostrative di un sistema formale o di una macchina di Turing. Sebbene questi ultimi siano in grado di fornire una riproduzione a posteriori in termini computazionali dei risultati ottenuti *via evidentiae* dal procedere razionale, ciononostante ad essi rimane precluso l'accesso a processi di dimostrazione creativi e aperti, che facciano appello alla visualizzazione e alla comprensione di strutture matematiche infinite. A sostegno di tali considerazioni, inoltre, ci siamo serviti delle preziose riflessioni di Galvan e Pagani per mostrare come non sia possibile pervenire ad una propria e reale autofondazione della logica, nemmeno attraverso la procedura elenctica. *Élenchos*, al contrario, costituisce una procedura fondativa eteronoma rispetto al dominio della logica, in quanto poggia su dei postulati propri dell'ontologia e, in particolare, del realismo.

Da ultimo, ci siamo soffermati sul tema classico della natura infinitaria dell'anima, e in particolare sulla nozione di "infinità trascendentale" che Pagani intende come autentica dimensione che funge da discrimine tra l'essere umano da una parte e il regno animale e gli automi dall'altra. Secondo la tradizione neoscolastica, l'anima costituisce una «apertura» sull'essere. Tale apertura rappresenta un orizzonte, in primo luogo, infinito, nel senso indicato dalla capacità dell'anima di abbracciare in sé ogni cosa che le si manifesti, e, in secondo luogo, trascendentale, in quanto condizione di possibilità del darsi di ogni determinazione finita.

Come è noto, nel ventesimo secolo il progresso scientifico e tecnologico ha subito una marcata accelerazione contribuendo notevolmente al miglioramento del benessere economico dei Paesi industrializzati o in via di sviluppo. Di converso, perciò, sempre più risorse sono state incanalate e destinate al finanziamento della ricerca in diversi campi e settori. Ciò ha fatto sì che il complesso e le linee guida del sapere scientifico, passando attraverso un processo di divulgazione e di

banalizzazione, venissero a influenzare, in maniera, potremmo dire, quasi totalizzante, diversi settori delle sfere della vita pubblica e privata, determinandone così un processo graduale ma radicale di "tecnicizzazione". Successivamente, la ricerca nel campo d'avanguardia dell'intelligenza artificiale ha raggiunto un livello di crescita esponenziale che continua sino ad oggi, in forza dei numerosi successi che l'applicazione di questa tecnologia ha ottenuto in ambito sperimentale e, in misura sempre maggiore, anche commerciale. Oggi, in un'epoca in cui la comunicazione e la condivisione dei contenuti è istantanea, e grazie ad una diffusa ma spesso imprecisa opera di divulgazione, la conoscenza dei risultati raggiunti in questo campo è divenuta oramai una sorta di bagaglio culturale comune, sebbene poco approfondito e accolto in modo generalmente acritico. Mentre analoga diffusione su larga scala non è stata data alle riflessioni critiche intorno alle possibilità che un tale campo di ricerca apre effettivamente. Pertanto, è di fondamentale importanza contribuire a fornire un quadro preciso di che cosa si intenda quando si parla di intelligenza artificiale, in modo tale da demistificare un certo tipo di retorica che in buona parte alimenta le pubblicazioni rivolte al grande pubblico, il più delle volte spinte da interessi commerciali più che da attenzione alla verità.

Bibliografia

E. Agazzi, *Alcune osservazioni sul problema dell'intelligenza artificiale*, in «Rivista di Filosofia Neo-Scolastica», LIX, 1967, no. 1, pp. 1-34, https://www.jstor.org/stable/43068656.

E. Agazzi, *The Rise of the Foundational Research in Mathematics*, in «Synthese», XXVII, 1974, no. 1/2, pp. 7-26, https://www.jstor.org/stable/20114903.

G. Basti – A.L. Perrone, *Le radici forti del pensiero debole: dalla metafisica, alla matematica, al calcolo*, Il Poligrafo, Padova 1996.

P. Benacerraf, *God, the Devil, and Gödel*, in «The Monist», LIII, 1967, vol. 51, no. 1, pp. 9-32, https://www.jstor.org/stable/27902014.

F. Berto, *Tutti pazzi per Gödel! La guida completa al Teorema di Incompletezza*, Laterza, Bari 2008[8].

D.J. Chalmers, *The Conscious Mind. In Search of a Fundamental Theory*, Oxford University Press, 1996.

M. Davis, *Il calcolatore universale. Da Leibniz a Turing*, trad. it. di G. Rigamonti, Adelphi, Milano 2003[8]; tit. orig. *The Universal Computer. The Road from Leibniz to Turing*, 2000.

S. Feferman, *Are There Absolutely Unsolvable Problems? Gödel's Dichotomy*, in «Philosophia Mathematica», XXXV, 2006, vol. 14, no. 2, pp. 134-152, https://doi.org/10.1093/philmat/nkj003.

S. Feferman, *Transfinite Recursive Progressions of Axiomatic Theories*, in «The Journal of Symbolic Logic», XXVII, 1962, vol. 27, no. 3, pp. 259-316, https://www.jstor.org/stable/2964649.

S. Galvan, *Gödel e il modello computazionale della mente*, in «Rivista di Filosofia Neo-Scolastica», XCVI, 2004, no. 1, pp. 145-174, https://www.jstor.org/stable/43063049.

S. Galvan, *Introduzione ai teoremi di incompletezza*, FrancoAngeli, Milano 1992.

S. Galvan, *Non contraddizione e terzo escluso. Le regole della negazione nella logica classica intuizionistica e minimale*, FrancoAngeli, Milano 1997[6].

S. Galvan, *Pensiero e Logica*, in «Rivista di Filosofia Neo-Scolastica», LXXXIX, 1987, pp. 118-128, https://www.jstor.org/stable/43061647.

A. Giordani, *Logica dell'evidenza*, in «Rivista di Filosofia Neo-Scolastica», XCII, 2000, no. 3/4, pp. 582-626, https://www.jstor.org/stable/43063247.

K. Gödel, *Proposizioni formalmente indecidibili dei* Principia mathematica *e di sistemi affini I*, in *Opere. Volume 1: 1929-1936*, trad. it. di E. Ballo, S. Bozzi, G. Lolli e C. Mangione, Bollati Boringhieri, Torino 1999; tit. orig. *On formally undecidable propositions of* Principia mathematica *and related systems I*, in *Collected Works. Volume I: Publications 1929-1936*, Oxford University Press, New York, 1986.

K. Gödel, *Alcuni teoremi basilari sui fondamenti della matematica e loro implicazioni filosofiche (1951)*, in *Opere. Volume 3: Saggi inediti e conferenze*, trad. it. di E. Ballo, G. Lolli, C. Mangione e P. Pagli, Bollati Boringhieri, Torino 2006; tit. orig. *Some basic theorems on the foundations of mathematics and their implications*, in *Collected Works. Volume III: Unpublished Essays and Lectures*, Oxford University Press, New York, 1995.

D. Hilbert, *Sull'infinito*, trad. it. di A. Frigo, Castelvecchi, 2013; tit. orig. *Über das Unendliche*, in «Mathematische Annalen», LVII, 1925, vol. 95, pp. 161-190.

R. Kurzweil, *The Singularity Is Near. When Humans Transcend Biology* (2005), Duckworth, 2018.

P. Lindstrom, *Penrose's New Argument*, in «Journal of Philosophical Logic», XXX, 2001, vol. 30, no. 3, pp. 241-250, https://www.jstor.org/stable/30226726.

J.R. Lucas, *Minds, Machines and Gödel*, in «Philosophy», XXX, 1961, vol. 36, no. 137, pp. 112-127, https://www.jstor.org/stable/3749270.

J.R. Lucas, *Satan Stultified: A Rejoinder to Paul Benacerraf*, in «The Monist», LIV, 1968, vol. 52, no. 1, pp. 145-158, https://www.jstor.org/stable/27902069.

P. Mancosu, *Il programma di Hilbert e i teoremi di incompletezza di Gödel*, in «Rivista di Filosofia Neo-Scolastica», XCVIII, 2006, no. 3, pp. 489-531, https://www.jstor.org/stable/43063730.

E. Nagel, J.R. Newman, *La prova di Gödel*, trad. it. di L. Bianchi, Bollati Boringhieri, Torino 1961[4]; tit. orig. *Gödel's Proof*, 1958.

P. Pagani, *La forma dell'*élenchos, in «Rivista di Filosofia Neo-scolastica», CXXII, 2020, no. 4, pp. 969-994.

P. Pagani, *Sulla attualità del concetto di "anima"*, in «Aquinas», LVI, 2013, no. 2, pp. 425-439.

R. Penrose, *Shadows of the Mind. A Search for the Missing Science of Consciousness* (1994), Vintage, Londra 2005.

R. Penrose, *Beyond the Doubting of a Shadow. A Reply to Commentaries on Shadows of the Mind*, in «Psyche», II, 1996, http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html.

K.R. Popper – J.C. Eccles, *L'io e il suo cervello*, trad. it. di G. Mininni e B. Continenza, Armando Armando, Roma 1981, voll. I-III; tit. orig. *The Self and Its Brain. An Argument for Interactionism*, Springer, 1977.

H. Putnam, *Minds and Machines*, in *Dimensions of Mind*, ed. S. Hook, New York 1960, e in *Philosophical Papers. Volume 2: Mind, Language and Reality*, Cambridge University Press, 1975.

S. Russell, *Human Compatible. Artificial Intelligence and the Problem of Control*, Penguin Books, 2020[2].

J.R. Searle, *Il mistero della coscienza*, trad. it. di E. Carli, Raffaello Cortina Editore, Milano 1998; tit. orig. *The Mistery of Consciousness*, New York Review of Books, New York 1997.

J.R. Searle, *Il mistero della realtà*, trad. it. di P. Di Lucia e L. P. Glazel, Raffaello Cortina Editore, Milano 2019; tit. orig. *The Basic Reality and the Human Reality*, 2019.

S. Shapiro, *Mechanism, Truth, and Penrose's New Argument*, in «Journal of Philosophical Logic», XXXIII, 2003, vol. 32, no. 1, pp. 19-42, https://www.jstor.org/stable/30226779.

S. Shapiro, *Incompleteness, Mechanism, and Optimism*, in «The Bulletin of Symbolic Logic», IV, 1998, vol. 4, no. 3, pp. 273-302, https://www.jstor.org/stable/421032.

P. Slezak, *Minds, Machines and Self-Reference*, in «Dialectica», XXXVIII, 1984, vol. 38, no. 1, pp. 17-34, https://www.jstor.org/stable/42971428.

A.M. Turing, *Computing Machinery and Intelligence*, in «Mind», LXXV, 1950, vol. 59, no. 236, pp. 433-460, https://doi.org/10.1093/mind/LIX.236.433.

H. Wang, *Dalla matematica alla filosofia*, trad. it. di A. Giacomelli, Bollati Boringhieri, Torino 2016; tit. orig. *From Mathematics to Philosophy*, 1974.