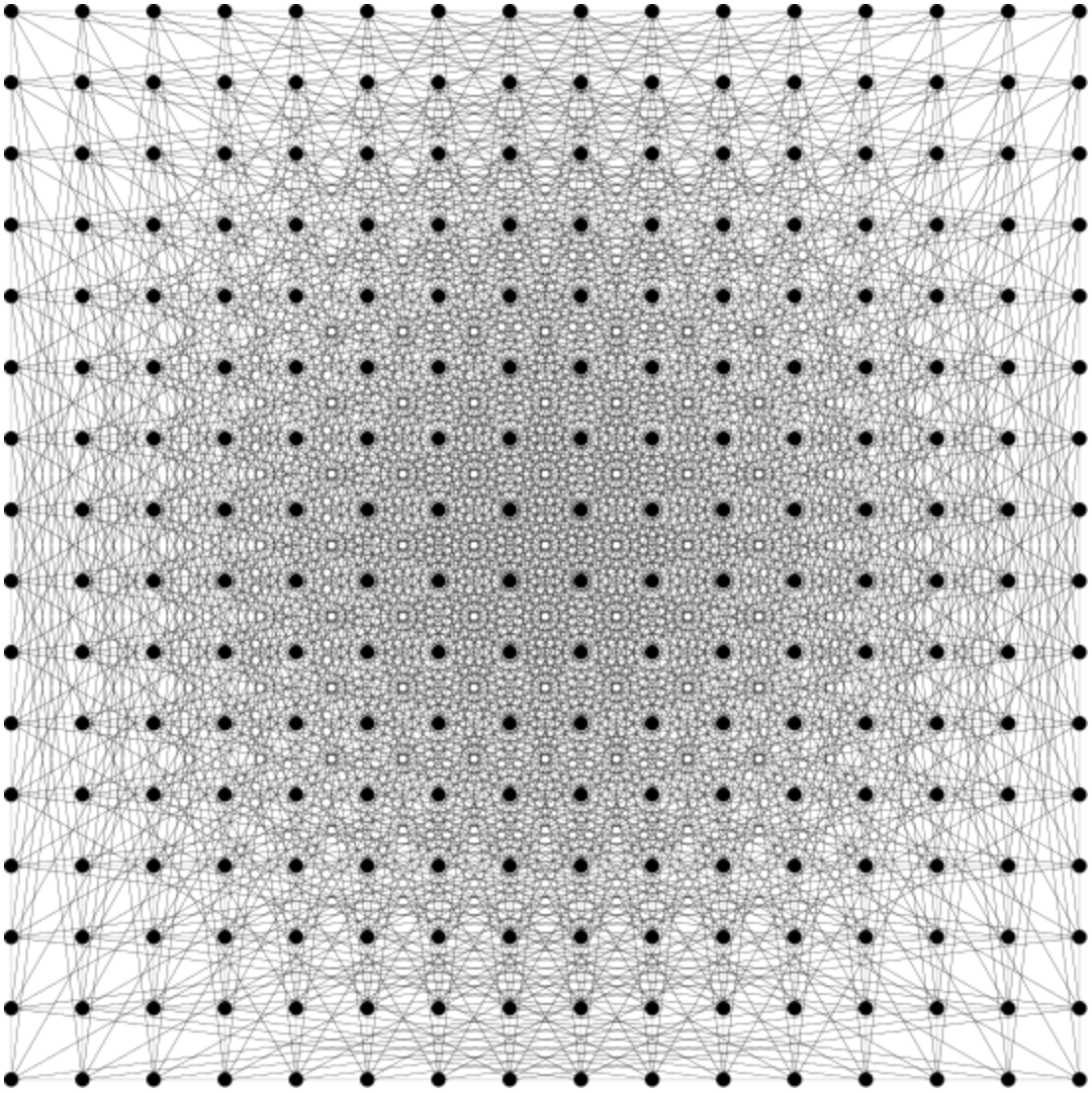
Sir Roger Penrose ha fatto tante cose, oltre che vincere un Nobel. Dal punto di vista della filosofia della matematica è famoso per la sua affermazione secondo cui la coscienza umana non è computazionale; come corollario si ha che il cervello non è una macchina di Turing, e quindi è in grado di fare qualcosa che nessun algoritmo può fare per le limitazioni del primo teorema di incompletezza di Gödel. Non che la maggior parte dei filosofi della matematica concordi con lui, non foss'altro che perché la mente umana non è un sistema coerente; però bisogna dargli atto di aver provato a vedere le cose da un altro punto di vista. Un suo recente commento mi pare però sia molto più condivisibile. Come si può leggere in questo tweet, in un suo video Penrose sostiene che dovremmo smetterla di parlare di intelligenza artificiale e usare piuttosto il termine "artificial cleverness".

Prima di andare più a fondo nel pensiero di Penrose permettetemi una considerazione personale. Per tradurre "artificial cleverness" ho scelto l'espressione "ingegno artificiale" per due motivi, a parte il restare con l'acronimo IA. Il primo è banale: "clever" viene spesso tradotto come "intelligente" ma qui non potevo certo farlo. Il secondo è più sottile: "ingegno artificiale" mi ricorda gli "ingegni minuti" narrati da Giovanbattista Vico e ripresi da Benedetto Croce nella sua crociata :-) contro gli scienziati, che sono sì tanto bravi ma non portano davvero conoscenza. Anche in questo caso infatti chi parla dà un giudizio esplicito su cosa fa una categoria diversa dalla propria: i non-filosofi per Croce, i non-umani per Penrose. Sicuramente Croce lo faceva per sminuire gli scienziati; Penrose probabilmente per sminuire l'IA; per me la cosa è un po' diversa.

Torniamo a Penrose. Il suo punto di vista è che l'intelligenza implica la coscienza, ma una macchina non ha coscienza. *Modus tollens*: una macchina non è intelligente. (Questo perché per lui la coscienza non è computazionale, come scrivevo all'inizio). Chiaramente il problema si è così semplicemente spostato: ma è proprio vero che una macchina non ha coscienza? Non entro nel merito, non avrei le competenze, ma continuando nella lettura di Penrose qualche pezzo del puzzle si è finalmente messo a posto. Ecco cosa dice: «Prendiamo gli studenti di matematica. Alcuni capiscono quello che fanno. Altri sono semplicemente ingegnosi. Sanno ripetere ciò che hanno imparato, e lo sanno fare anche molto bene; ma non è detto che capiscano quello che fanno.» Chi ha scritto il tweet chiosa dicendo che quel divario tra il saper calcolare bene e il capire davvero è esattamente il divario che Penrose vede tra quanto i computer sanno fare oggi e la vera intelligenza.

Devo avere già scritto in passato che io mi sono laureato in matematica nella categoria degli "ingegnosi", ma avrò cominciato a capire davvero un po' di matematica solo dopo altri quindici anni, quindi mi ritrovo nelle parole di Penrose. Credo però di avere una

differenza fondamentale sul valore dell'ingegnosità dell'IA, almeno nel mondo matematico. Prendiamo l'annuncio della scorsa settimana di OpenAI: un loro modello interno ha dimostrato che una congettura affermata da Paul Erdős nel 1946 era falsa. Come a volte capita nella teoria dei numeri, la congettura è anche facile da presentare: scelti n punti nel piano, quanti al più possono trovarsi a distanza 1? A un matematico non interessa tanto il numero esatto come funzione di n ma il suo andamento: Erdős congetturò che la funzione avesse andamento \(n^{1+o(1)} \), che tradotto dal matematichese vuol dire "cresce più di Mn per un qualsiasi M, ma meno di n^k per un qualsiasi k maggiore di 1. La struttura ottimale si pensava fosse quella a griglia quadrata che sembra un op-art mostrata qui sotto; essa si costruisce partendo da un ipercubo di lato 1 e proiettandolo sul piano. Con il quadrato abbiamo quattro punti e quattro distanze unitarie; col cubo arriviamo a otto punti e dodici distanze unitarie; e così via. Il numero di segmenti unitari che si riesce a mantenere con la proiezione cresce più lentamente del numero di punti: da qui si ricava il limite indicato da Erdős. Negli anni questo limite minimo non era stato superato: Erdős aveva offerto 500 dollari a chi l'avesse dimostrato, ma solo 50 dollari a chi avrebbe trovato un controesempio.



Il modello di OpenAI ha trovato che invece per infiniti n c'era una configurazione che permetteva di avere $n^{1+\delta}$ punti a distanza 1, con δ strettamente maggiore di zero anche se l'articolo non specificava il valore; fatti i conti, era circa \(6.24\times 10^{-38} \). La dimostrazione, o meglio la validità del controesempio, è stata verificata automaticamente con Lean e si è rivelata corretta. Ma la parte davvero interessante è che la costruzione ha usato concetti della teoria dei numeri algebrica, cosa che non era mai venuta in mente di usare da parte di nessun matematico. Se ho ben capito cosa fa, invece che partire dagli ipercubi prende dei campi di numeri algebrici con certe proprietà che una volta proiettati sul piano ottengono la struttura voluta.

Il risultato è davvero interessante, e non scherzo. Il problema era noto e studiato, e nessuno si aspettava che la congettura di Erdős fosse falsa. Inoltre non pare ci siano stati passaggi human-assisted: il modello ha fatto tutto da solo. D'accordo, è un controesempio

e non una dimostrazione, ma a differenza di tanti io non sono così tranchant e accetto la loro validità come veri progressi in matematica. Insomma, il modello è stato davvero ingegnoso. Ma ora viene il resto. Per prima cosa, il problema non se l'è inventato il modello, ma qualcuno gliel'ha dato. In secondo luogo, ventiquattr'ore dopo la presentazione del risultato il teorico dei numeri Will Sawin ha postato un preprint dove ha migliorato la soluzione, dando un valore esplicito per $\delta=1{,}014$. Questo non è certo venuto in mente al modello OpenAI. Ecco: possiamo discutere se "decidere autonomamente di risolvere un problema" e "una volta trovata una risposta, vedere se e come si può usare quel metodo per fare di meglio" siano indice di coscienza; ma sicuramente questo manca ancora agli LLM, e quindi io do ragione alla conclusione di Penrose anche se non accetto necessariamente la sua premessa implicita.

Questo vuol dire sminuire questi modelli? No, niente affatto. Come ho scritto, sono in grado di dare risposte a problemi che prima non l'avevano, e il loro agnosticismo - nel senso che non hanno delle linee di attacco predefinite - è davvero utile. Semplicemente ognuno (umani e IA) ha il suo ruolo: se volete, adesso c'è una sorta di simbiosi ma i ruoli restano ben precisi. La figura qui sotto, presa dal substack di Ben Lorica, dà qualche idea in più su come la pensa la comunità matematica.