SAGGI

L'intelligenza artificiale generativa e il gioco dell'imitazione

di Vincenzo Fano

This paper discusses the recent experiments by Jones and Bergen that showed ChatGPT-4 passed the Turing test. To this end, the game proposed by Turing is placed in the proper philosophical dimension by showing how it tries to answer a well-formulated and circumscribed question. The details of the experiment are then investigated. Finally, some important consequences are drawn regarding the problem of explainability of neural networks.

Keywords: Turing Test; ChatGPT-4; Explainability of Neural Networks; Imitation Game; Philosophy of Psychology

Introduzione

Il tentativo di sviluppare l'intelligenza artificiale in modo non logico ha origini coeve rispetto al progetto mainstream basato su modelli logici[1]. "Non logico" nel senso duplice di basarsi sull'addestramento e sull'apprendimento più che su un software o un hardware già strutturati e prendere le mosse da una configurazione semplice e simmetrica, che poi si sviluppa mediante l'addestramento. Questo approccio, che proviene più dalla psicologia che dall'informatica, venne aspramente criticato in un celebre libro del 1968[2]. La critica, pur corretta, non teneva in considerazione il fatto che il percettrone poteva essere sviluppato con gli strati intermedi, come venne poi proposto sempre dagli psicologi, negli anni Ottanta[3]. I limiti di quest'approccio, ispirato in parte dal funzionamento del nostro cervello, erano invece altri due: in primo luogo, la scarsità di dati etichettati e in secondo luogo, la limitata potenza dei calcolatori. Solo nel 2012, AlexNet ottenne alla celebre competizione ImageNet un risultato strabiliante – 95% di immagini classificate correttamente. AlexNet era una rete neurale convolutiva, basata su un'idea di Ian LeCun, realizzata da Geoffrey Hinton, Alex Krizhevsky e Ilya Sutskever: tutti nomi che hanno fatto la storia della rivoluzione tecnologica a cui stiamo assistendo.

Da quel momento in poi le reti neurali addestrate sono entrate nella nostra vita quotidiana in modo sempre più capillare. L'idea non logica degli psicologi ha di fatto vinto rispetto a quella logica proposta dalla maggior parte degli studiosi di intelligenza artificiale. Vedremo però che in un certo senso la logica si prenderà la rivincita.

Di solito si fa risalire la fondazione del progetto dell'intelligenza artificiale al famoso lungo congresso del 1956 a Dartmouth, durante il quale venne coniato il termine a cui parteciparono studiosi del calibro di Claude Shannon, Marvin Minsky e Herbert Simon. D'altra parte, nessuno può negare che questo nuovo programma di ricerca abbia avuto fra le pagine del celebre articolo di Alan Turing comparso sulla rivista filosofica *Mind* nel 1950 una spinta decisiva[4]. Questo straordinario saggio, fra le altre cose, proponeva quello che venne poi chiamato "test di Turing", cioè un gioco ideato come cartina di tornasole per stabilire se le macchine possano o meno pensare.

---

1 - W.S. McCulloch - W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, «Bulletin of Mathematical Biophysics», 5 (1943), pp. 115-133 e F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain*, «Psychological review», 65 (1958/6), p. 386.

2 - M. Minsky - S.A. Papert, *Perceptrons*, MIT Press, Cambridge MA 1968.

3 - D.E. Rumelhart - J.L. McLelland (eds.), *Parallel, Distributed Processing,* MIT Press, Cambridge MA 1986, Vol. I e II.

4 - A. Turing, *Computing machinery and intelligence*, in «Mind», 59 (1950), pp. 433-460.

Come spiega Melanie Mitchell nel suo importante libro del 2019[5], il test di Turing non è preso troppo sul serio dagli informatici, in quanto si presta a trovate pubblicitarie di dubbio valore tecnologico. Inoltre, l'intelligenza è tante altre cose, non solo superare il test di Turing. Infine, anche chatbot estremamente poco sofisticati riescono a simulare conversazioni umane.

Con l'avvento delle nuove reti neurali, come la serie di ChatGPT, il test di Turing è tornato a far parlare di sé. Questa tecnologia è molto complessa e simula in modo sorprendente la capacità di conversare umana. Nello Cristianini, nel suo bel libro, *Machina Sapiens*[6], racconta che gli stessi ingegneri e programmatori che la hanno realizzata sono rimasti sorpresi dalle sue abilità di dialogo. Resta però il fatto che spesso si è ancora scettici sulla capacità di ChatGPT di superare il test di Turing. Ad esempio, Tommaso Poggio, nel suo ottimo volume, *Cervelli, menti, algoritmi*, scritto con Marco Magrini, sostiene che «i *large language models* […] almeno per ora non sono in grado di superare una severa versione del test di Turing»[7]. "Per ora" nel 2023, ma già nel 2024 sono usciti due articoli che descrivono esperimenti molto seri, dai quali risulta, invece, che ChatGPT-4 ha superato ampiamente il test di Turing[8]. Di questo parleremo nel presente breve testo. Nel prossimo paragrafo delineeremo una metodologia generale in filosofia della psicologia che consente di collocare correttamente il significato filosofico del test di Turing. In quello successivo descriveremo e discuteremo il gioco dell'imitazione proposto da Turing nel 1950. Nella sezione 4 analizzeremo gli esperimenti da cui risulta che il test è stato superato. Alcune brevi considerazioni concludono l'articolo.

Il quadro generale

Se decidiamo di provare a comprendere qualcosa degli elettroni, dobbiamo studiare la meccanica quantistica, cioè la migliore teoria che possediamo riguardo a queste particelle. "Migliore" nel doppio senso che è molto ben confermata e altamente esplicativa. Se accettiamo almeno in parte la tesi del realismo scientifico, allora, pur consapevoli che la teoria quantistica è una creazione umana, siamo convinti che essa dica anche qualcosa su come sono effettivamente fatti gli elettroni. In che senso possiamo essere realisti scientifici moderati non lo discutiamo in que-

5 - M. Mitchell, *Intelligenza artificiale,* Einaudi, 2022 pp. 38-41. L'ed. originale è del 2019.

6 - N. Cristianini, *Machina Sapiens,* Il Mulino, Bologna 2024

7 - T. Poggio - M. Magrini, *Cervelli, menti, algoritmi*, Sperling & Kupfer, Milano 2023, p. 111.

8 - C.R. Jones - B. K. Bergen, *Does GPT-4 pass the Turing test?*, in «*arXiv*» 2023: *arXiv:2310.20216*; C.R. Jones - B.K. Bergen, *People cannot distinguish GPT-4 from a human in a Turing test, in* «arXiv» 2024: *arXiv:2405.08007.*

sto contesto[9]. Basta dire che la meccanica quantistica è almeno in parte vera. Dove il termine "vera" va inteso in senso corrispondentista.

Stiamo indugiando sull'epistemologia di una scienza naturale altamente confermata, al fine di chiarire alcuni punti che torneranno utili rispetto al nostro problema, che in fondo è di filosofia della psicologia.

All'interno di un paradigma genericamente aristotelico di scienza, per comprendere gli elettroni dovremmo stabilire quali sono le loro caratteristiche essenziali – quella che potremmo chiamare "sostanza seconda" dell'elettrone – e allo stesso tempo fornire una definizione che ne colga la forma, cioè la "sostanza prima". La prima istanza fa ancora parte della scienza moderna. Un elettrone ha una certa massa e una certa carica, che lo contraddistinguono, ma queste sue peculiarità di fatto non dicono quasi nulla del reale suo comportamento; la seconda è stata completamente abbandonata. Infatti, conoscere gli elettroni significa oggi stabilire le equazioni che governano il loro comportamento nelle diverse situazioni, come ad esempio, quando l'elettrone è libero o fa parte di un atomo di idrogeno. Avere a disposizione queste equazioni, tuttavia, non ha nulla a che fare con sapere l'essenza dell'elettrone; significa invece, conoscerne il comportamento e saperlo manipolare.

Questo esempio mostra un punto molto importante che caratterizza l'epistemologia della scienza moderna in generale: non proviamo più ad afferrare l'essenza della realtà, ma solo a comprenderne alcuni comportamenti e a fornirne delle spiegazioni funzionali.

Dobbiamo tenere presente questa scelta metodologica in parte rinunciataria, ma vincente, quando prendiamo in considerazione lo studio della mente umana.

La psicologia nasce come scienza sperimentale nella seconda metà dell'Ottocento. Di solito come data discriminante si fa riferimento all'istituzione a Lipsia del primo laboratorio di psicologia sperimentale fondato da Wilhelm Wundt nel 1879[10]. A grandi linee, Wundt sostiene che ci sia un piano dell'esperienza vissuta precedente al costituirsi dei concetti scientifici. La psicologia sarebbe quindi la scienza empirica e sperimentale di questo livello di realtà primario, basata sull'auto-osservazione. Questo progetto epistemologico ha dato qualche frutto, ma è rapidamente andato incontro a problemi insolubili. Già Comte aveva osservato che la psicologia intesa in questo senso è impossibile, dato che l'osservatore e l'osservato coincidono e quindi interferiscono[11]. Inoltre Wittgenstein mostra che di fatto una scienza

9 - Vedi però V. Fano, *A Regional Scientific Image of Time*, in «Argumenta», 8 (2/2023), pp. 407-425 [https://www.argumenta.org/article/a-regional-scientific-image-of-time-special-issue/].

10 - Vedi, ad esempio, E.G. Boring, *A history of experimental psychology*, Appleton-Century-Crofts, New York 1957[2].

11 - A. Comte, *Cours de philosophie positive*, Bachelier, Paris 1864[2], vol I, pp. 30ss.

dei propri vissuti è impossibile, poiché il linguaggio pubblico e condiviso li modifica e li costituisce[12]. L'introspezione dunque in generale si dimostra un metodo tutt'altro che affidabile per fare psicologia scientifica. L'idea di una scienza prima dei vissuti passa progressivamente alla filosofia. In varie forme la sosterranno Bergson, Dilthey, Mach, Husserl e Carnap[13], fra gli altri. D'altra parte, la psicologia, al fine di indirizzarsi su una base più scientifica, soprattutto negli Stati Uniti, si avvicina a posizioni comportamentiste, come quella prima di Watson e poi di Skinner. Questi ultimi abbandonano completamente il progetto di mettere a punto una scienza della soggettività, concentrandosi invece sugli stimoli e sui comportamenti, cioè sulla parte pubblica e condivisa del fenomeno psichico.

Negli anni Sessanta del Novecento il modello comportamentista si dimostrerà insufficiente. Tuttavia questo non darà origine a un ritorno della psicologia come scienza dell'esperienza soggettiva, bensì all'introduzione di modelli mentali non osservabili che intervengono a spiegare l'elaborazione umana degli stimoli e la produzione di comportamenti. Questi modelli possono essere suggeriti dai report dell'esperienza vissuta da parte dei soggetti sperimentali, ma possono essere anche, come si dice oggi, subpersonali, cioè di fatto inconsci. Anzi, per la maggior parte sono modelli di come lavorerebbe il nostro cervello, che avrebbero poco a che fare con quello che percepiamo soggettivamente, che, come abbiamo accennato, è sfuggente e non facilmente oggettivabile. Inoltre negli ultimi decenni abbiamo acquisito molti nuovi dati sul funzionamento neurofisiologico del nostro cervello, per cui è oggi anche possibile confrontare, almeno in parte, i modelli cognitivi con la nostra neurofisiologia. Resta il fatto che la psicologia scientifica non è più una scienza della nostra esperienza soggettiva[14].

Questa situazione generale ha molte eccezioni, su cui vale la pena soffermarsi brevemente. Ad esempio, Gerald Edelman, nel suo classico libro *La materia della mente*[15], prova a sostenere l'esigenza di una scienza "non galileiana", cioè che provi a tenere conto anche dell'esperienza soggettiva. Il progetto neuro-fenomenologico di Varela[16] si muove nella stessa direzione. Recentemente Dan Zahavi ha provato a riproporre una psicologia fenomenologica[17]. I gestaltisti, che ormai sono una spa-

12 - L. Wittgenstein, *Ricerche filosofiche*, Einaudi, Torino 2009. L'ed. originale è del 1953.

13 - Vedi ad esempio H. Spiegelberg, *The phenomenological movement*, Springer, Berlin 1971 e W. Stegmüller, *Die Hauptströmungen der Gegenwartsphilosophie*, Kröner, Stuttgart 1965.

14 - M.R.W. Dawson, *Mind, Body, World. Foundations of cognitive science*, AU Press, Edmonton AB 2013.

15 - G. Edelman, *La materia della mente*, Adelphi, Milano 1992.

16 - F.J. Varela, *Neurophenomenology: a methodological remedy for the hard problem,* in «Journal of Consciousness Studies», 4 (3/1996), pp. 330-349.

17 - S. Gallagher - D. Zahavi, *The Phenomenological Mind*, Routledge, London 2008.

ruta minoranza, non hanno mai abbandonato il progetto di tenere conto dell'esperienza soggettiva, soprattutto quando si tratta di psicologia della percezione; ecc.

Tutti questi tentativi di oggettivare l'esperienza soggettiva hanno ottenuto risultati parziali e controversi. Proviamo a indagare meglio le ragioni di questa difficoltà.

Al fine di comprendere questo problema, torniamo brevemente all'epistemologia della fisica. Fin dai tempi di Newton questa scienza ha introdotto per ragioni esplicative entità non direttamente misurabili, come le forze, i campi e le particelle. Questo significa che in psicologia non dobbiamo avere paura a fare altrettanto. E in effetti, il cognitivismo, come abbiamo visto brevemente prima, si è mosso proprio in questa direzione. Le entità fisiche non misurabili hanno due caratteristiche salienti: in primo luogo, sono descrivibili in maniera esatta, la maggior parte delle volte mediante un linguaggio matematico; in secondo luogo, sono connesse a quelle misurabili da precise leggi funzionali o indirettamente o direttamente. Qualcosa di simile capita per i modelli cognitivi, che sono spesso formulati in termini logici o computazionali e sono ben connessi agli stimoli e ai comportamenti che vorrebbero spiegare.

Quando invece passiamo all'esperienza soggettiva, tutto cambia. Abbiamo buone ragioni epistemologiche per ritenere che le altre persone abbiano un'esperienza soggettiva simile alla nostra. In effetti le altre persone sono molto simili a noi dal punto di vista biologico; noi abbiamo un'esperienza soggettiva, quindi viene da dire che anche loro la abbiano. È vero che questa è una strana inferenza a partire da un solo caso, tuttavia la nostra esperienza soggettiva ci accompagna per buona parte della nostra vita, è fortemente correlata a ciò che capita al nostro corpo e l'analogia fisico-biologica fra noi e gli altri è molto dettagliata. Possiamo dunque concludere che gli altri vivano un'esperienza soggettiva simile alla nostra. A questo punto siamo, dal punto di vista epistemologico, in una situazione che ha vantaggi e svantaggi rispetto a quella della fisica. In primo luogo, non possediamo un linguaggio esatto per descrivere la nostra esperienza soggettiva, come la matematica o la logica. Tuttavia, in un certo senso, anche se deformato dal linguaggio e dalle nostre idiosincrasie, abbiamo, almeno in prima approssimazione un accesso privilegiato alla realtà della nostra soggettività, che, invece, nel caso, ad esempio, di un elettrone è del tutto assente. Questo accesso parzialmente privilegiato è un'arma a doppio taglio: da un lato ci convince che non sia così difficile cogliere l'essenza dei fenomeni mentali, dall'altro, però, non abbiamo alcun modo per oggettivarli. Infatti, non solo ci manca un linguaggio adeguato, come la matematica o la logica, ma non siamo neanche in grado di dire in modo chiaro che cosa sarebbero dei contenuti soggettivi oggettivati. Ad esempio, che significato potrebbe avere "percepire la mente di un'altra persona"? Se stiamo percependo, stiamo percependo un nostro stato mentale, non quello di un altro. Inoltre, che senso avrebbe l'espressione "so quel che pensi", se di fatto non possediamo un linguaggio ogget-

tivo per descrivere i nostri pensieri? Insomma, al di là dell'ingannevole intuizione che avremmo un accesso privilegiato agli stati mentali, fare una scienza della soggettività sembra un compito oltremodo difficile.

Questo non significa che sia del tutto impossibile. Da un lato notiamo che per conoscere non esiste solo la scienza, ma anche altre attività tipicamente umane. Come hanno mostrato, ad esempio Nelson Goodman e Hans-Georg Gadamer[18], anche l'arte insegna. Certo sono conoscenze con un basso grado di esattezza e giustificazione, ma non sono del tutto assenti. In altre parole, raccontare storie sulla nostra soggettività è tutt'altro che inutile, anche se non può avere i crismi della scientificità. Dall'altro, in un certo senso la nostra esperienza soggettiva, anche se non in modo diretto e costitutivo, gioca un ruolo indiretto e regolativo nella psicologia scientifica. Faccio un esempio. Chiaramente non possiamo utilizzare una nostra esperienza soggettiva per giustificare un asserto psicologico; tuttavia il nostro vissuto ci guida nell'indagine dei comportamenti e nella costruzione dei modelli.

Questa prima parte sull'epistemologia della psicologia scientifica ci è servita per introdurre il test di Turing. Come vedremo nella prossima sezione, questo test è rinunciatario dal punto di vista epistemologico, ma è proprio ciò di cui abbiamo bisogno per fare buona scienza. In psicologia, esattamente come in fisica, dobbiamo seguire il principio galileiano di "non tentar l'essenza".

Il gioco dell'imitazione

Alan Turing pubblica sulla rivista filosofica *Mind* nel 1950 un articolo straordinario, che ancora oggi influenza il nostro modo di concepire i computer e il loro rapporto con la mente umana. Il titolo del saggio è *Macchine calcolatrici e intelligenza*. Di fatto però Turing si pone una domanda diversa: "possono le macchine pensare?" Non sappiamo se Turing fosse consapevole della distinzione kantiana fra "intelletto" e "pensiero". Il primo sovraintende l'attività argomentativa della mente umana, mentre il secondo ha un senso molto più ampio, che comprende anche l'immaginazione e le idee della ragione. Lo sviluppo del saggio e degli esempi che l'Autore propone portano a ritenere che Turing abbia in mente proprio la nozione ampia di pensiero. Tuttavia questa domanda così formulata è troppo ambigua. Turing propone quindi una questione molto diversa, che è indagabile e precisa. Mettiamo in due stanze chiuse una macchina e una persona. Dopo di che un'altra persona, chiamata l'interrogante, può comunicare con le due stanze solo mediante messaggi scritti a macchina. Dopo un po' di dialogo l'interrogante deve stabilire in

18 - N. Goodman, *I linguaggi dell'arte,* Il Saggiatore, Milano 1976 e H.-G. Gadamer, *Verità e metodo*, Bompiani, Milano 1983.

quale stanza ci sia la macchina e in quale la persona. Di fatto la persona prova ad aiutare l'interrogante nel suo compito, mentre la macchina non deve farsi scoprire.

Dunque, la domanda, "possono le macchine pensare?" viene sostituita da "può una macchina ingannare l'interrogante nel gioco dell'imitazione?". Prima di proseguire, sottolineiamo l'aspetto metodologico centrale della mossa di Turing: egli ha evitato di rispondere alla domanda troppo ambiziosa "che cosa è il pensiero?". Come abbiamo già detto, oggi non ci chiediamo "che cosa è un elettrone?", ma proviamo a spiegare e prevedere come esso si comporta. Analogamente, noi non siamo in grado di rispondere in modo univoco alla domanda "che cosa è il pensiero?"; però sappiamo come si comportano le persone che pensano. Nella visione di Turing, pensare è un'attività prettamente intellettuale che si esprime linguisticamente.

Oggi, la scienza cognitiva, dopo aver accettato questa idea per decenni, la sta mettendo in discussione, attribuendo un'importanza sempre maggiore al corpo e all'azione[19]. Probabilmente Turing era consapevole che la sua nozione prettamente linguistica di pensiero fosse limitata, ma taglia corto: questa è la domanda che vuole discutere, senza entrare nel merito se sia o meno quella giusta. Per Turing, dunque, il pensare coincide con i comportamenti linguistici di una persona. Notiamo anche che Turing si disinteressa completamente di come la macchina riesca a simulare i comportamenti linguistici umani. Essa può utilizzare hardware e software sostanzialmente diversi da quelli di cui si avvalgono le persone. Attualizzando il suo modello, il fatto che oggi le immense reti neurali dietro ai chatbot funzionino in modo sostanzialmente diverso dal cervello umano è irrilevante; l'importante è che i comportamenti linguistici vengano effettivamente simulati in modo adeguato.

È importante anche sottolineare che Turing non attribuisce alcun ruolo alla soggettività e alla coscienza. Abbiamo visto nella sezione precedente che non dobbiamo farci ingannare dal nostro accesso ai nostri stati mentali. Infatti, il mondo in prima persona, pur essendo molto importante, non è oggetto di scienza. Qualche pagina dopo, Turing sarà categorico: «Il solo modo di conoscere che un uomo pensa è essere quell'uomo»[20]. In altre parole, se volessimo coinvolgere la coscienza nella nostra discussione scientifica della nozione di pensiero saremmo costretti a una visione solipsistica. Questo non significa, prosegue Turing, che non ci siano misteri riguardo alla coscienza, ma solo che ci si può porre la domanda sul pensiero delle macchine anche senza aver risolto quei misteri. In altre parole, come spesso accade nella metodologia scientifica, la strategia *divide et impera* è vincente.

---

19 - Vedi ad esempio, M.R.W. Dawson, *Mind, Body, World: Foundations of Cognitive Science*, Athabasca University Press, Edmonton AB 2013, cap. 5.

20 - A.M. Turing, *Computing machinery and intelligence*, in «Mind» 59 (1950), pp. 433-460, qui p. 446. Trad. nostra.

Da questa presa di posizione di Turing segue che tutti quegli argomenti, come la "stanza cinese" di Searle o la "testa di legno" di Block[21], sono sostanzialmente irrilevanti. Per essere più precisi, la nozione di pensiero di cui Turing si vuole occupare non comprende la soggettività, che resta un concetto sostanzialmente estraneo al suo approccio parzialmente operazionale.

Turing prosegue spiegando che cosa intende con il termine "macchine" nella sua domanda. Egli si concentra sulle cosiddette "macchine digitali" universali. Turing non è esplicito al riguardo, ma sembra di capire che egli abbia in mente un computer che sia una macchina universale di Turing, cioè che sia in grado di simulare tutte le possibili computazioni nel senso di Turing.

Resta aperta la questione delle macchine non digitali, come ad esempio il regolo calcolatore. Può infatti essere che il nostro cervello non sia una macchina a stati discreti, ma a stati continui. Turing sottolinea però che l'interfaccia comunicativa di una macchina a stati continui non può che essere a stati discreti. In altre parole, anche se le macchine a stati continui fossero essenzialmente diverse rispetto a quelle a stati discreti, difficilmente questa differenza risulterebbe evidente all'interrogante[22].

Dal punto di vista della metodologia della psicologia, di cui abbiamo discusso nella sezione precedente, dove si colloca il gioco dell'imitazione? Di certo non tiene conto né della psicologia introspettiva, né della neuropsicologia. Turing non è interessato né alla coscienza, come abbiamo visto, né a come i neuroni influenzino il pensiero. O meglio, in questo articolo tiene fuori gioco tali questioni. D'acchito potrebbe sembrare che Turing sia un comportamentista, poiché in fondo stabilisce la presenza del pensiero mediante i comportamenti linguistici palesi. Come hanno mostrato Oppy e Dowe[23] non è così. Infatti il compito di programmare una macchina digitale capace di produrre quei comportamenti linguistici è estremamente complesso. In altre parole, Turing non sta immaginando gli stupidi arnesi proposti da Searle e Block, che non sono altro che degli enormi registri di stimoli e risposte; egli sta proponendo di mettere a punto un software altamente strutturato che si basi su un vero e proprio modello cognitivo. Possiamo quindi affermare che questo

21 - G. Oppy - D. Dowe, *The Turing Test*, in «The Stanford Encyclopedia of Philosophy» (Winter 2021 Edition), edited by Edward N. Zalta, in https://plato.stanford.edu/archives/win2021/entries/turing-test/, § 6.

22 - Va inoltre notato inoltre, che per quel poco che si sa sulle macchine a stati continui, come ad esempio lo Shannon's General Purpose Analog Computer (GPAC), sembra che siano in grado di calcolare esattamente come una macchina di Turing: O. Bournez - M.L. Campagnolo - D.S. Graça - E. Hainry, *Polynomial differential equations compute all real computable functions on computable compact intervals*. in «Journal of Complexity», 23 (3/2007), pp. 317-335.

23 - G. Oppy - D. Dowe, *The Turing Test*, cit.

saggio di Turing è una delle prime espressioni di quel movimento che ha dominato la psicologia della seconda metà del Novecento, che chiamiamo "cognitivismo".

Concludiamo questa sezione con le previsioni formulate da Turing. La prima è che nel 2000 sarà possibile programmare macchine che dopo 5 minuti del gioco dell'imitazione non vengano identificate nel 70% dei casi. Nella prossima sezione vedremo che di fatto questo non è successo. La seconda è che riuscire a superare il gioco dell'imitazione è soprattutto una questione di software e non di hardware. E anche su questo sappiamo che Turing si sbagliava, poiché le odierne reti neurali sono anche un importante progresso dal punto di vista ingegneristico. Tuttavia, anche se Turing non è del tutto esplicito al riguardo, il punto che l'intero articolo vuole sottolineare sembra essere un altro: le macchine digitali universali saranno sempre più brave a simulare il pensiero umano. E su questo, soprattutto gli ultimi dieci anni di ricerche tecnologiche hanno dato completamente ragione a Turing.

ChatGPT-4 ha superato il test di Turing?

Prima di entrare nel vivo della discussione del nostro tema, ancora alcune considerazioni generali.

Non è chiaro che cosa realmente rileverebbe il superamento del test di Turing. Come spesso capita in questo tipo di esperimenti, non possiamo certo dire che superare il test di Turing sia condizione necessaria e sufficiente per attribuire intelligenza o pensiero alle macchine. Neanche, però possiamo affermare che sia necessaria, poiché esistono senz'altro altre forme di intelligenza[24]. Neppure forse possiamo ritenere che sia sufficiente, in quanto si possono immaginare macchine del tutto stupide che potrebbero superarlo, come la "testa di legno" di Block o la "stanza cinese" di Searle. Questo argomento standard[25] non è però del tutto convincente, poiché questo tipo di macchine immaginate in esperimenti mentali filosofici è molto poco realistico. Come abbiamo già detto, un programma di computer in grado di parlare non può essere una lista enorme di input e di output, ma deve avere una sua struttura interna. Possiamo quindi dire che in linea di massima superare il test di Turing è una condizione sufficiente per attribuire a una macchina l'intelligenza. Resta il fatto, come abbiamo già sottolineato, che l'intelligenza di cui stiamo parlando è una nozione scientifica e cognitiva, che non ha molto a che fare con la nostra soggettività. Questa è stata la giusta scelta metodologica di Turing, che va rispettata.

Va inoltre sottolineato che, quando si realizza concretamente il test di Turing sono presenti molte variabili, che vanno opportunamente controllate. Di fatto stori-

24 - Vedi, ad esempio, G. Vallortigara, *Altre menti*, Il Mulino, Bologna 2022.

25 - G. Oppy - D. Dowe, *The Turing Test*, cit.

camente ci sono stati tentativi di somministrare il test, a partire da quello di ELIZA[26] a quelli legati al famoso premio istituito da Loebner[27]. In letteratura si trovano altri esperimenti, ma nessuno è conclusivo, fino ai due articoli che discuteremo di Jones e Bergen[28], che arrivano alla conclusione che ChatGPT-4 ha superato il test. Vediamo ora come è stato realizzato l'esperimento.

Nel primo articolo i due autori avevano provato diversi prompt. Nel secondo si concentrano su quello che meglio aveva performato, che è un prompt di ChatGPT-4. Non riporteremo tutti i dettagli dell'esperimento, solo quelli per noi rilevanti.

Una prima differenza essenziale rispetto alla situazione immaginata da Turing è che l'interrogante si trova di fronte a un solo "testimone", che può essere o ChatGPT-4 oppure una persona. L'interrogante, come proposto originariamente, ha a disposizione 5 minuti per indovinare se sta chattando con una persona o con una macchina. Alla persona-testimone vien chiesto di esprimersi come persona. L'interrogante ha graficamente di fronte una chat in cui scrive e ottiene risposte con un certo ritardo, anche quando sta messaggiando con una macchina, in modo da non notare la differenza. Il tasso di successo dell'interrogante viene misurato come il numero di volte che l'interrogante afferma di avere di fronte una persona. Il risultato più importante del primo esperimento è che nel caso in cui il testimone è un certo tipo di prompt di ChatGPT-4 su 855 casi l'interrogante ha sostenuto di avere davanti una persona il 50% delle volte. Per contro, quando aveva davanti una persona lo ha sostenuto il 66% delle volte. Non è immediato confrontare questo risultato con la previsione di Turing del 30% di identificazioni sbagliate. Teniamo presente che nell'esperimento proposto da Turing, il massimo che una macchina può ottenere è che l'interrogante sbagli il 50% delle volte, cioè che la sua ipotesi sia come il caso. Invece, nella situazione sperimentale di Jones e Bergen, che è diversa, il meglio per la macchina sarebbe performare come le persone, cioè essere scambiata per una persona il 66% delle volte. Facile allora fare la proporzione 50 : 66 = x : 50, cioè il tasso di successo dell'esperimento di Jones e Bergen è analogo al 38% in quello originale di Turing. Il test nella forma originale di Turing non è ancora stato realizzato, per cui questo confronto è solo virtuale. Tuttavia sembra che il test di Turing sia ampiamente superato. Si noti che ovviamente la dimestichezza con i *large language models* degli interroganti fa la differenza, nel senso che gli interroganti più esperti vengono ingannati di meno.

---

26 - J. Weizenbaum, *ELIZA – a computer program for the study of natural language communication between man and machine,* in «Communications of the ACM», 9 (1/1966), pp. 36-45.

27 - S.M. Shieber, *Lessons from a restricted Turing test*, in «*arXiv*» (1994) [cmp-lg/9404002].

28 - C.R. Jones - B.K. Bergen, *Does GPT-4 pass the Turing test?*, cit.; C.R. Jones - B.K. Bergen, *People cannot distinguish GPT-4 from a human in a Turing test*, cit.

Il secondo esperimento condotto da Jones e Bergen sostanzialmente conferma i dati del primo. Per questa ragione i due autori passano dal titolo interrogativo del primo articolo a quello affermativo del secondo. Su 500 testimoni del tipo migliore di prompt di ChatGPT-4 l'interrogante afferma di aver di fronte una persona – dopo una chat di 5 minuti – in più del 50% dei casi. Resta il fatto che le persone riescono a "sembrare" persone in una percentuale significativamente superiore di casi.

Tirando le somme, possiamo affermare che le previsioni di Turing si sono avverate, anche se con un certo ritardo. Le macchine digitali sono in grado di conversare con noi in un modo che offusca la nostra capacità di capire se stiamo dialogando con una macchina o con una persona. Al di là dell'entusiasmante risultato tecnologico, che con il prosieguo della ricerca non può che migliorare, il superamento del test di Turing pone problemi di carattere etico estremamente importanti. Infatti queste chatbot ormai quasi indistinguibili dalle persone possono essere usate in modi ingannevoli e dannosi per gli utenti.

Considerazioni conclusive

In *Machina sapiens*, Nello Cristianini usa un efficace apologo per descrivere la situazione che stiamo vivendo. È come se un'intelligenza aliena fosse atterrata fra di noi. Non aliena nel senso della sua origine, visto che la abbiamo messa a punto noi, ma nel senso che di fatto non sappiamo bene come funziona. ChatGPT era basata su 175 miliardi di parametri, cioè una rete neurale i cui pesi delle connessioni sono numeri fra 0 e 1 di 175 miliardi di nodi. Si suppone che ChatGPT-4 ne abbia almeno un ordine di grandezza in più. In pratica, anche se avessimo davanti a noi il tabulato di questi quasi 2000 miliardi di parametri e la loro assegnazione ai nodi della rete non sapremmo che cosa ChatGPT-4 stia facendo. Nel 1997, il calcolatore dedicato Deep Blue dell'IBM batté a scacchi il campione del mondo in carica Gary Kasparov. Fu un evento epocale. Gli ingegneri e i programmatori che avevano costruito Deep Blue, anche se non potevano seguire tutti i dettagli dei suoi processi, avevano ben chiaro come funzionasse la loro creatura. Essa aveva memorizzato migliaia di partite; considerata la posizione dei pezzi sulla scacchiera, analizzava la maggior parte degli sviluppi possibili per parecchie mosse e poi valutava in base a certi parametri decisi dai programmatori la qualità della posizione finale; sceglieva quindi la variante con il punteggio più alto. Per contro, il programma AlphaGo della DeepMind, che ha sconfitto 4 a 1 nel 2016 uno dei più forti giocatori del mondo di Go, cioè Lee Se-Dol, anche se ha una struttura di massima ben nota, nello scegliere le mosse del gioco ragiona in un modo sostanzialmente ignoto ai suoi costruttori.

Questa è una novità profonda nel mondo della tecnologia. Non sapendo bene come lavorano le reti neurali, non solo fatichiamo a modificarle, ma siamo di fronte a un notevole scacco morale per la nostra consapevolezza. Uno dei compiti che

ci aspetta è proprio quello di incrementare la nostra conoscenza di questi strumenti che ormai usiamo quotidianamente. Tanto più che essi, come abbiamo visto, hanno superato la barriera epocale del test di Turing. Questa maggiore consapevolezza può essere un problema informatico, nel senso di riuscire a trovare dei software che prendendo in input i parametri e la struttura di una rete neurale, diano in output una rappresentazione computazionale più comprensibile di come la rete stia operando. In questo senso nell'introduzione parlavamo della rivincita della logica. La logica, infatti, la capiamo, mentre le reti neurali no, quindi rappresentare logicamente le reti neurali è oggi un compito fondamentale.

Non dobbiamo però trascurare un altro aspetto interessante della comprensione delle reti neurali. Sean Trott e Timoty B. Lee, in questo bel post on line[29], spiegano in modo chiaro ed efficace come funzionano le reti neurali tipo ChatGPT. Non solo, nel breve testo ci sono anche i link alla maggior parte dei paper che nello spazio di pochi anni hanno portato nel 2017 alla prima messa a punto di ChatGPT. Dare un'occhiata a questi articoli è molto interessante, perché si entra nei meccanismi di lavoro delle teste pensanti che hanno messo a punto questa nuova tecnologia. Di fatto, sembra che questi ricercatori abbiano sviluppato strumenti linguistici naturali per descrivere e comprendere i *large language models* che stavano manipolando. In altre parole, essi stessi hanno elaborato dei parziali strumenti di comprensione che sono serviti a guidare le ricerche. Questi strumenti linguistici meritano la nostra attenzione e ulteriori ricerche, perché una loro disamina aiuterà ad attenuare il problema della comprensione delle reti neurali in cui oggi siamo immersi.

Ringrazio Marco Giunti e Massimo Mugnai, dai quali ho appreso molto riguardo a questi temi.

VINCENZO FANO
Professore ordinario di Logica e Filosofia della scienza presso il Dipartimento di Scienze Pure e Applicate, Università di Urbino
vincenzo.fano@uniurb.it

29 - [https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/].