Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Alessio Plebe

Inquietudini e misteri della mente artificiale

1. Introduzione

L'attuale sviluppo dell'Intelligenza Artificiale (IA) presenta due aspetti di cui ci si intende occupare. Uno è piuttosto conosciuto e dibattuto, riguarda una certa inquietudine che non pochi provano a riguardo. L'altro ha finora suscitato decisamente meno attenzioni, ma si ritiene sia non meno rilevante, è la mancanza di spiegazioni su come l'IA riesca ad essere così intelligente.

Tornando alle preoccupazioni per l'IA, una di quelle che maggiormente tengono banco è la stessa che ricorre ogni volta che si affaccia all'orizzonte una discontinuità tecnologica significativa: la perdita di posti di lavoro. Non è questo il genere di inquietudine di cui si vuol parlare, per affrontarla seriamente occorrerebbero competenze di sociologia ed economia di cui chi scrive ne è carente. Ci si limita a suggerire, con leggerezza, che l'eventuale effetto di riduzione delle necessità lavorative da parte dell'IA potrebbe venire incontro alla prospettiva delineata da Bertrand Russell nel suo delizioso elogio dell'ozio, sottoscrivendo le parole con cui lo introduce: "Io penso che in questo mondo si lavori troppo, e che mali incalcolabili siano derivati dalla convinzione che il lavoro sia cosa santa e virtuosa" (Russell 1932). Utopia filosofica? Forse, ma non sono pochi a coltivare l'idea che l'IA possa essere una strada per avvicinarla (Srnicek and Williams 2015). L'inquietudine di cui si intende parlare è decisamente meno contingente, ma forse ancor più profonda, è un potenziale duro colpo al senso di primi della classe, profondamente radicato in noi umani, soprattutto nella cultura occidentale (Lovejoy 1936). Quello che in filosofia viene spesso denominato "eccezionalismo umano" aveva avuto già una serie di batoste non da poco. La prima si deve a Copernico, il quale, con la sua teoria eliocentrica, ha mostrato come l'uomo non sia affatto al centro dell'universo, ma collocato in suo infimo anonimo angolino. Ancor peggio, Darwin ha rivelato che condividiamo antenati comuni con tutti gli altri esseri viventi, ci ha insistito la recente cognizione comparata nello sperimentare in altri animali tutto il repertorio di comportamenti che si ritenevano nostra unica prerogativa. Aggiungiamoci le neuroscienze, che mostrano impietosamente come non vi sia nemmeno un recondito neurone fisiologicamente eccezionale rispetto ad altri animali (Finlay and Workman 2013). Ora l'unicità ed eccezionalità del nostro mondo mentale, quello di cui siamo più fieri e gelosi, viene insidiata addirittura da entità che non sono nemmeno viventi. Un'avvisaglia si è avuta nel giugno 2022, quando un ingegnere di Google, Blake Lemoine, ha esternato la sua convinzione che un nuovo programma chiamato LaMDA, di cui lui era incaricato di eseguire verifiche, era cosciente, provava emozioni, ed dotato di una propria personalità. La notizia fece scalpore, e sul momento la quasi totalità dei commenti furono negativi, tacciando Lemoine di ingenuità, e venne persino licenziato da Google. Ma il rapido progresso di modelli del genere di LaMDA ha posto pesanti dubbi sull'archiviazione troppo sbrigativa delle convinzioni di Lemoine. Lui poteva essere tacciato da ingenuo buontempone, ma nessun poteva osar insinuare qualcosa del genere nei confronti del più conclamato filosofo della coscienza contemporaneo, David Chalmers, che pochi mesi dopo ha posto in termini rigorosi l'interrogativo su forme di coscienza posseduta da questi nuovi modelli dell'IA (Chalmers 2023).

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

LaMDA è l'acronimo di *Language Model for Dialogue Applications*, dove con la dicitura *Language Model* vengono oggi indicati dei particolari modelli neurali artificiali dedicati alla comprensione e produzione di linguaggio naturale. In qualche modo incarnano il punto di congiunzione con l'idea primordiale di intelligenza in un computer, che secondo Alan Turing avrebbe dovuto esplicarsi proprio nella capacità di conversare amabilmente con umani. La sezione successiva racconterà questi primordi, e le iniziali diatribe filosofiche che ha suscitato. Diatribe accese, ma del tutto salottiere, considerando che l'IA per oltre mezzo secolo è rimasta di fatto abissalmente lontana dalle capacità mentali dell'uomo. Nella terza sezione si traccerà il recente percorso che ha condotto, in modo repentino e del tutto inaspettato, alle prestazioni attuali, quelle che avvicinano in modo inquietante la mente umana.

La quarta sezione affronta l'altra faccenda che si considera centrale per l'IA: la mancanza di spiegazioni su come sia in grado di esibire prestazioni mentali talmente simili a quelle umane. E' una situazione per certi versi paradossale, si può accettare non saper spiegare in modo esaustivo nemmeno i più semplici comportamenti della mente umana, ma la prerogativa della mente artificiale è proprio di essere progettata e realizzata da esseri umani. Come mai il suo funzionamento risulta invece talmente elusivo? Un motivo risiede nella tipologia dei modelli della recente IA, si tratta di modelli neurali artificiali. Pur se il termine "neurale" allude a similarità molto deboli con i loro omonimi biologici, pare che una caratteristica li accomuni al cervello. Così come–nonostante gli enormi progressi delle neuroscienze–il cervello rimane sostanzialmente un mistero, altrettanto refrattari alle indagini risultano i modelli dell'IA.

2. Mente e conversazione

Gli sviluppi più recenti ed eclatanti dell'IA sembrano esaudire alla lettera la futuristica concezione di un computer intelligente, proposta da Alan (Turing 1950) nel celebre articolo *Computing Machinery and Intelligence*. Nel porsi l'interrogativo se un computer possa essere ritenuto intelligente, Turing va al cuore dell'obiezione che potrebbe essere sollevata qualunque sia il compito sofisticato, chiamiamolo X, che un computer dimostra di svolgere: "d'accordo, questo computer sa svolgere il compito X, ma non sbilanciamoci, per *intelligenza* si intende ben altro". Cosa sia il "ben altro" si può prestare ad infinite ridefinizioni tali da conservare sempre l'intelligenza nello scrigno delle proprietà unicamente umane. Per sbarazzarsi del comodo scudo offerto dalla vaghezza del termine "intelligenza", Turing adotta una strategia ben nota in filosofia della scienza con il nome di *operazionalizzazione*. Introdotta dal premio Nobel per la fisica Percy (Bridgman 1927), consente di attribuire un significato rigoroso a termini di per se vaghi ed ambigui. Per Bridgman il significato di un termine può essere interamente specificato da un insieme di *operazioni* che lo coinvolgono. Questo procedimento è pienamente adottato in fisica, dove termini come forza, massa, carica elettrica, trovano specificazione in una serie di operazioni in cui sono coinvolti in modo non ambiguo. Turing fa lo stesso con il termine "intelligenza", lo equivale ad una operazione, che con leggerezza delinea nella fattispecie di un gioco.

2.1 *Giocare conversando*

The Imitation Game è il titolo del film di Morten Tyldum del 2014, che ha fatto conoscere la figura di Turing al grande pubblico, concentrandosi però più sulle vicende legate alla seconda guerra mondiale che su questo geniale gioco apparentemente semplice ed innocuo. Turing lo introduce nella forma di

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

gioco di società del genere fiorente in Inghilterra, noti come *Victorian parlor games* (Beaver 1974). Viene giocato da tre protagonisti: un uomo (A), una donna (B) e uno che interroga (C), che sta in una stanza separata dagli altri due, a lui noti solamente come X e Y. L'interrogante C deve cercare di scoprire chi sia l'uomo e chi la donna, pronunciandosi quindi con un "X è A e Y è B" o viceversa. Per orientarsi nella sua scelta C ha facoltà di porre qualsiasi domanda ad A e B (esclusa l'ovvia domanda se sia uomo o donna), i quali debbono rispondere sinceramente. La mossa di Turing è sostituire gli attori A e B, ora A è un essere umano, e B è un computer. Tale computer computer risulta degno della proprietà dell'intelligenza se l'interlocutore C conversando con A e B riguardo qualunque argomento, non sa dire quale dei due sia una macchina oppure un conversante in carne ed ossa. Fin qui l'elegante strategia dell'operazionalizzazione, poi arriva la provocazione più scandalosa: Turing ci crede davvero che verrà un futuro in cui i computer diventeranno intelligenti e lo dimostreranno sapendo conversare come noi, e lo sostiene con una serie di argomenti teorici di notevole spessore.

Il gioco perde tutta la sua innocenza trasformandosi in una scandalosa provocazione. Nel prendere la capacità di conversare come marchio dell'eventuale intelligenza di una macchina, Turing aveva mirato dritto alla facoltà che per molti è la più sublime e unica della specie umana: il linguaggio. Immaginare che un computer possegga la facoltà del linguaggio, come aveva spudoratamente fatto Turing, voleva quindi dire profanare il santuario della mente umana.

Non pochi filosofi hanno sottolineato come il linguaggio non sia solamente un modo di comunicare, ma l'impianto con cui è costruita la nostra visione del mondo. Ludwig (Wittgenstein 1922) affermava che "i limiti del mio linguaggio significano i limiti del mio mondo". Soprattutto, il linguaggio ha costituito il baluardo ultimo dell'eccezionalismo umano. Cartesio, parafrasato da Noam (Chomsky 1966) nell'aderire pienamente al suo pensiero, sostiene che "la parola è l'unico segno e il marchio più sicuro della presenza di pensiero". Chomsky ha lungamente capeggiato la dura battaglia contro l'insorgere di tante evidenze dall'etologia sulla continuità tra le capacità comunicative umane e di altri animali (Savage-Rumbaugh, Shanker, and Taylor 1998; Pepperberg 1999), costituendo una agguerrita scuola di pensiero (Pinker 1994; Hauser, Chomsky, and Fitch 2002; Penn, Holyoak, and Povinelli 2008). Non bastavano pappagalli, bonobo e altri primati non umani, ora ad avanzare pretese verso la preziosa e supposta unicamente umana facoltà del linguaggio ci si mettono anche i computer.

2.2 *La mente artificiale anima discussioni da salotto*

La sfida lanciata da Turing ha provocato uno dei più ampi ed accesi dibattiti filosofici del secolo scorso, non sopiti nemmeno oggi. Ne è ovviamente responsabile l'intera impresa dell'IA, che coglieva il guanto lanciato da Turing giusto un anno dopo la sua scomparsa (McCarthy et al. 1955). Non furono necessari troppi anni perché si levassero alte voci di reazione. Il filosofo Hubert Dreyfus ha speso l'intera sua vita a trovare ragioni per negare al computer la possibilità di raggiungere mai capacità umane come il linguaggio. In uno dei suoi più famosi scritti, *What Computers Can't Do: A Critique of Artificial Reason* (Dreyfus 1972), la parte centrale mira proprio a ridicolizzare i tentativi pionieristici dell'IA nel campo del linguaggio, in particolare la traduzione automatica – a quell'epoca davvero fallimentari – concludendo che mai il linguaggio umano sarà alla portata di un computer.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Agli albori dell'IA sia critiche che simpatie erano soprattutto in linea di principio, ma nel 1966 lo scandalo di Turing pare prendere una prima forma concreta. *Eliza* è il software che impersona uno psicoterapeuta computerizzato (Weizenbaum 1966) in grado di fornire brevi risposte abbastanza credibili ad interlocutori umani. La genialità era di impersonare uno psicoterapeuta di scuola rogeriana, la cui strategia è grosso modo di portare sempre il paziente ad autointerrogarsi senza sbilanciarsi mai troppo. In questo modo usando poche parole delle frasi scritte dagli interlocutori, Eliza può formulare blande e generiche controdomande, senza doversi preoccupare di alcuno dei dettagli sulla situazione effettiva del suo paziente, tantomeno dei riferimenti nel mondo delle parole lette e prodotte. Nonostante la sua palese limitazione, Eliza fu un successo di pubblico come pochi altri nella storia del software, e proprio questo destò preoccupazioni. Per Ned (Block 1981) c'era qualcosa di sbagliato, di immorale, nel lasciar credere alla gente che potesse esserci qualche forma di intelligenza dietro ad un astuto gioco di rimaneggiamento di pochi simboli. Block andò anche oltre il caso specifico di Eliza, sostenendo che anche programmi più potenti non avrebbero potuto arrogarsi la proprietà dell'intelligenza a pieno titolo. Su questo terreno però fu un altro argomento, proposto dal suo collega John (Searle 1980), a diventare la lama più affilata puntata contro il test di Turing. Divenuta celebre sotto il nome di "stanza cinese", la situazione immaginaria proposta da Searle vede se stesso chiuso in una stanzetta, a cui vengono passati fogli con scritte in cinese. Vi è a disposizione un meraviglioso manuale, da usare cercando nella pagina a sinistra un testo identico a quello appena ricevuto, basta poi ricopiare quel che si legge nella pagina a destra, e consegnarlo a chi sta fuori dalla stanza. Bene, i fogli in entrata sono frasi di un interlocutore, e quelle in uscita le risposte, la stanza chiusa funziona quindi come una macchina in grado di conversare perfettamente, idonea quindi a superare il test di Turing, con il piccolo particolare, osserva Searle, che lui non sa nemmeno una parola di cinese. Ovvero un software potrebbe rispondere meccanicamente senza aver capito nulla.

Diversi trovarono l'espediente di Searle tanto convincente da considerare la possibilità per un computer di comprendere il linguaggio un traguardo irraggiungibile. Non tutti. Per esempio, secondo Daniel (Dennett 1980) se Eliza poteva essere tacciata di impostura, lo era in modo ben più subdolo la *stanza cinese*. L'impostura è di condurre il lettore in una immaginazione, falsa, costellata di passaggi impossibili, per poi convincerlo della tesi che vuol sostenere, e Dennett coniò un nome specifico per queste raffinate imposture: *intuition pump* (Dennett 2012). Il passaggio platealmente impossibile è il manuale, che in teoria, per qualunque contenuto di una conversazione, dovrebbe avere la risposta pronta e appropriata da fornire. Ovviamente non è realisticamente realizzabile un tale manuale, data l'infinita possibilità di costrutti linguistici in una conversazione.

Negli ultimi due decenni dello scorso millennio si assiste ad una curiosa dissociazione riguardo l'IA. Dopo il fuoco fatuo dell'eroica quanto effimera Eliza, più si accumulavano ricerche sul dotare computer di linguaggio, più l'impresa si rivelava improba e velleitaria. Dopo Eliza, la ricerca sulla comprensione artificiale del linguaggio aveva intrapreso una strada molto diversa, che non voleva più lasciarsi tentare da trucchi e stratagemmi. Si è tentato di riversare nei computer le teorie linguistiche del funzionamento del linguaggio umano, soprattutto quelle di ispirazione chomskiana che nascono già parzialmente formalizzate (Winograd 1972). Pur essendo teorie di grande portata nel descrivere la miriade di intricati fenomeni delle lingue umane, si rivelarono sostanzialmente fallimentari nel dotare i computer di capacità linguistiche. Nonostante la mancanza di evidenze concrete su cui discutere, il tema in astratto sulla possibilità per l'IA di padroneggiare un linguaggio era diventato un

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

irresistibile divertimento salottiero per diversi filosofi. Searle e il suo esperimento mentale hanno continuato a costituire il fulcro per innumerevoli confutazioni e contro-confutazioni, con interi libri dedicati al suo dibattito (Preston and Bishop 2002). Da notare che, per quanto acceso, il dibattito era tutto sommato esente da inquietudini, si trattava di esercitare le proprie abilità filosofiche su uno scenario – il computer intelligente in grado di conversare – del tutto ipotetico, privo nel panorama tecnologico di alcun indizio di una sua prossima concretizzazione.

3. Il fenomeno *Deep Learning*

Per meglio apprezzare gli aspetti di inquietudine e mistero dell'attuale IA è opportuno raccontare brevemente la svolta avvenuta nell'ultimo decennio. Si è detto come ad inizi del 2000 circolasse una diffusa disillusione sulle possibilità di dotare un computer di capacità linguistiche, ma ancor più in generale l'intera impresa di una mente artificiale era passata in secondo piano. Era addirittura caduta in un tale discredito, per cui settori della ricerca informatica una volta accolti sotto il suo ombrello, preferivano evitare la sua etichetta, e trovare altre sigle specifiche con cui caratterizzarsi. Le ricerche sull'elaborazione del linguaggio tramite computer facevano parte di NLP (*Natural Language Processing*), quelle sulla visione artificiale si etichettavano come IP (*Image Processing*) e l'analisi di grandi moli di dati era l'ambito *Big Data Mining*. Il lessema IA era accuratamente evitato. In poco più di una decina di anni la situazione si è rapidamente e sorprendentemente ribaltata grazie ad una invenzione che va sotto il nome di *Deep Learning* (DL).

3.1 *Il rinascimento dell'Intelligenza Artificiale*

Così è stato letteralmente definito in (Tan and Lim 2018) l'inaspettato risorgere dell'IA, sulla spinta del DL. Il termine *Learning* in DL esplicita la sua matrice radicalmente empirista, in piena continuità con le reti neurali artificiali degli anni '80 (Rumelhart and McClelland 1986b), di cui ne è una diretta discendenza. La storica contrapposizione tra corrente empirista e razionalista in IA ha ricalcato in piccolo quella di ben più lunga tradizione in ambito filosofico. Si è assistito ad un'alternanza tra le due, con la compagine razionalista dominante nel ventennio tra 1960 e 1980. Verso la fine degli anni '80 la situazione si ribalta grazie all'invenzione delle reti neurali artificiali, basate su semplici unità di calcolo disposte su più livelli, interconnessi tra di loro. La loro strategia era l'apprendimento, di qualunque genere di funzione di cui fossero disponibili esempi da imparare, non richiedendo di includere l'esplicitazione di alcun genere di regola.

Passando al termine *Deep*, non dispiace certamente agli esponenti di questa tecnica che *profondo* possa essere interpretato come *perspicace*, capace di andare a fondo delle questioni, ma il suo uso deriva da faccende prettamente tecniche. Le reti neurali artificiali degli anni '80 erano come detto costituite da strati di neuroni, ed era in uso catalogarle come *shallow* se il numero complessivo di strati era tre, o *deep* se vi erano quattro o più strati. Una rete neurale artificiale puo' svolgere compiti via via più sofisticati aumentando il numero complessivo delle sue unità, ma facendo i conti con le maggiori richieste di potenza di calcolo. A parità del numero complessivo il progettista puo' scegliere se avere solo tre strati ciascuno con molte unità, oppure ripartire le unità di cui puo' disporre su molti strati. Fino a pochi anni fa la seconda opzione era poco praticabile, perché il metodo matematico inventato per realizzare l'apprendimento, noto come *backpropagation* (Rumelhart, Hinton, and Williams 1986), funzionava bene per tre strati, e sempre meno bene passando ad un numero di strati maggiori.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Proprio questa limitazione aveva condotto all'esaurirsi del progresso delle reti neurali artificiali, divenute sempre più marginali all'interno dell'IA all'inizio di questo millennio.

Nel 2006 avviene la svolta, Geoffrey Hinton, già tra gli inventori della *backpropagation*, escogita una serie di artifizi matematici che consentono di addestrare reti neurali artificiali con quattro e cinque strati (Hinton and Salakhutdinov 2006), innescando un nuovo impulso di ricerca nel filone delle reti artificiali, raffinando rapidamente i metodi di apprendimento per un numero sempre più elevato di strati (Hinton et al. 2012; Kingma and Ba 2014). In quegli anni il DL esplode come fenomeno, grazie ad un numero di successi che lo rendono rapidamente popolare. Un modello basato su strati convolutivi profondi sviluppato da Hinton e suoi allievi domina la competizione più famosa nel campo del riconoscimento di immagini, facendo crollare l'errore dal 26.0% del precedente vincitore a ben 16.4% (Krizhevsky, Sutskever, and Hinton 2012). Da allora ogni anno modelli via via più perfezionati hanno continuato ad abbassare l'errore, arrivando ad eguagliare le prestazioni umane (VanRullen 2017). L'impatto del DL nella visione artificiale è formidabile, alcune applicazioni che fino a pochi anni prima erano poco più che progetti futuristici, sono diventati a portata di mano, come il caso dei veicoli a guida autonoma (Kuutti et al. 2019).

3.2 *Il difficile incontro tra modelli neurali e linguaggio*

Il linguaggio naturale è stato l'ambito in cui il DL ha avuto inizialmente difficoltà ad imporsi come era successo per la visione artificiale. L'unico settore dove subito il DL era risultato vincente è il riconoscimento del parlato (Veselý et al. 2013), dove le reti neurali artificiali hanno potuto beneficiare anzitutto della loro superiore capacità nel processamento di segnali, senza la necessità di cimentarsi con le complessità del linguaggio nel suo insieme.

Vi sono due ragioni di fondo che hanno reso difficile per le reti neurali entrare in confidenza con il linguaggio. Una è l'intrinseco conflitto tra il formato delle rappresentazioni delle reti neurali artificiali e il linguaggio. Quest'ultimo è realizzato da simboli, le parole, a loro volta composte tramite simboli morfologici, su scala inferiore, e ad una granularità ancora inferiore da simboli fonetici. La valuta corrente delle reti neurali sono invece vettori di numeri reali, tipicamente normalizzati nell'intervallo tra 0 e 1. Come conciliare vettori di numeri con la natura simbolica del linguaggio è un problema che ha afflitto le reti neurali artificiali sin dal loro primo affacciarsi nel mondo del linguaggio. Un primo tentativo, ma non certo timido. Il diciottesimo capitolo della raccolta *Parallel Distributed Processing* non solo è di gran lunga il più citato tra i vari capitoli, ma ha costituito una pietra dello scandalo attorno cui si è sviluppata una memorabile guerra intellettuale (Rumelhart and McClelland 1986a). Il modello intendeva mostrare come le regole morfologiche della forma passata dei verbi inglesi potessero essere imparate semplicemente dall'esperienza, non solo, che la traiettoria di apprendimento del modello ricalcava quella tipica osservata nei bambini. Un risultato che suonava come una inaccettabile provocazione in un periodo in cui la linguistica era dominata dalla visione di Chomsky, che sanciva l'impossibilità di apprendere il linguaggio senza ipotizzare un insieme di regole innate. In un articolo di oltre cento pagine due campioni del fronte razionalista, (Pinker and Prince 1988), smontarono il modello di Rumelhart e McClelland con una serrata critica. La parte principale dell'articolo è dedicata proprio al modo con cui le parole venivano convertite in vettori numerici, terreno dove Pinker aveva legittimamente gioco facile. Trovandosi per primi alle prese con questo bel problema, Rumelhart e McClelland avevano fatto ricorso ad una rappresentazione piuttosto

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

macchinosa e ben poco funzionale, denominata *Wickelfeature* in onore al suo inventore Wayne (Wickelgren 1969). Soprattutto, è una codifica palesamene priva di ogni plausibilità biologica.

La codifica vettoriale delle parole è il pesante inconveniente che ha gravato le possibilità di applicare il DL al linguaggio, fino all'introduzione di una soluzione particolarmente efficace, nota come *word embedding* (Mikolov et al. 2013). L'idea principale è di estendere alla codifica stessa il principio empirista, lasciando quindi che il modo di transitare da parole a vettori non fosse imposto a priori, ma appreso. Lo schema di questo apprendimento è noto come *skip-gram*, in cui data una parola il compito della rete neurale è di predire quelle che immediatamente la precedono e la seguono. Quel che viene modificato per attenuare l'errore di predizione è proprio il vettore che codifica la parola. E' evidentemente un compito impossibile con esattezza, ma nel tentare di svolgerlo la rete progressivamente affina vettori che possano svolgerlo al meglio su tutti gli esempi in cui compare una parola. La codifica emerge quindi con l'esposizione del modello ad un adeguato corpus testuale. I vettori risultanti da questo metodo presentano sorprendenti proprietà di semantica lessicale: vettori numericamente simili risultano semanticamente collegati, semplici operazioni aritmetiche sui vettori danno luogo a fenomeni di composizionalità semantica.

Vi è un'altra ragione, ancor più seria, che aveva reso le reti neurali artificiali poco adatte al linguaggio. Esse sono intrinsecamente statiche, mentre il linguaggio è tipicamente dinamico. Il significato si va formando parola su parola nel corso di una frase, di un discorso, di una conversazione. Il cervello di un ascoltatore deve mettere in campo una serie di meccanismi dinamici su molteplici scale temporali per catturare il senso di quel che sta udendo. Una elegante soluzione fu escogitata dallo psicolinguista Jeffrey (Elman 1990), affiancando al consueto strato nascosto di una rete neurale artificiale uno di egual lunghezza, le cui attivazioni conservano quelle dello strato nascosto all'intervallo temporale precedente. I due strati sono connessi da una matrice di pesi sinaptici che viene affinata durante l'apprendimento, facendo uso della convenzionale *backpropagation*. Questo schema permise ad Elman di costruire dei primi modelli di reti neurali in grado di cogliere fenomeni sintattici, dove l'ordine seriale delle parole è essenziale. Nonostante successivi affinamenti (Hochreiter and Schmidhuber 1997), questi tipi di rete, denominate genericamente *ricorsive*, hanno successo solamente se le frasi sono relativamente semplici e brevi. Non appena intercorrono relazioni importati tra parole troppo distanti tra di loro, diventa problematico catturarle. La difficoltà è nell'avere una memoria che sia al contempo di tempo sufficientemente lungo per mantenere relazioni distanti, ma non troppo nel conservare tracce di parole divenute irrilevanti nel corso del testo.

3.3 *La svolta vincente: il Transformer*

La svolta radicale è di pochi anni fa, con l'invenzione da parte del team di Google dell'architettura denominata *Transformer* (Vaswani et al. 2017). Lo schema complessivo del Transformer segue l'idea di *autoencoder*, introdotta originariamente da (Hinton and Zemel 1994), che consiste semplicemente nel compito di riprodurre in uscita di una rete lo stesso suo input, transitando per una prima parte detta *encoder* e una seconda, *decoder*. In questo caso l'input è una frase codificata vettorialmente, da riprodurre tal quale in uscita. Anche se il compito appare inutile, il vantaggio dell'autoencoder è di ottenere, nel vettore al confine tra encoder e decoder, delle rappresentazioni interne compatte e

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

ricche di significato, senza dover ricorrere alla supervisione. Tutto questo nel Transformer è associato ad un complesso meccanismo per tessere relazioni tra le varie parole. Viene definito come *attenzione*, e si basa su una serie di vettori ausiliari, denominati *query*, *key* e *value*, derivanti dal word embedding mediante matrici separate. Senza entrare nei dettagli, sostanzialmente questi vettori combinati tra di loro forniscono un punteggio numerico, per ogni parola della frase, riguardo alla rilevanza instaurata con qualunque altra parola. Nel pieno spirito delle reti neurali artificiali, non vi è nulla di predefinito nelle matrici che collegano i vettori query, key e value, vengono semplicemente apprese dall'esperienza. Anche i loro nomi, così come quello complessivo di *attenzione*, sono semplicemente suggestivi ma privi di un legame concreto con il loro significato comune. Il modo più corretto di caratterizzare il Transformer è come euristica, ma un'euristica che si è dimostrata straordinariamente efficace, di gran lunga oltre le aspettative dei suoi creatori. Inclusa nell'euristica è l'idea di replicare il sistema dell'attenzione tante volte, applicandolo solo ad una porzione della rappresentazione delle parole. Per via del *word embedding* (vedi §3.2) le parole sono rappresentate mediante vettori, tipicamente con migliaia di elementi. Questi vettori vengono ripartiti in porzioni, e a ciascuna porzione viene dedicato un *attention head*, in pratica una replica per intera del meccanismo. Alla fine, per calcolare quanto le parole stanno in relazione tra di loro, si sommano semplicemente i contributi di tutte le *attention head*. L'idea è che vi siano diverse dimensioni linguistiche che legano le parole tra di loro: sintattiche, semantiche, allineamento morfologico, e avere una dotazione di diverse attenzioni separate potrebbe essere vantaggioso. Lo sviluppo del Transformer mirava ad una precisa applicazione: la traduzione automatica, certamente non una applicazione da poco, ma comunque limitata rispetto alla portata che hanno presto assunto i modelli neurali del linguaggio. Il Transformer è stato rapidamente adottato in modelli via via più complessi, passando per BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al. 2019) e GPT (*Generative Pre-trained Transformer*) (Brown et al. 2020). Si tratta a questo punto di modelli, qui indicati con l'acronimo NLM (*Neural Language Model*), la cui caratteristica fondamentale è di incorporare una comprensione complessiva del linguaggio naturale, che può essere valorizzata in una varietà di modi diversi, non più solamente per la traduzione. Una delle modalità di impiego, il conversare, consegna soddisfazione postuma a Turing, ne sono protagonisti LaMDA e ChatGPT. Mentre il primo è rimasto nel mondo poco conosciuto della ricerca, il secondo grazie all'interfaccia pubblica ha aperto a tutti la possibilità, finora solo sognata, di conversare con il computer.

3.4 *Discussioni che si rianimano, non più nei salotti*

Improvvisamente e inaspettatamente la prospettiva del computer in grado di carpire la facoltà umana da molti ritenuta più preziosa, il linguaggio, stava diventando reale, con una progressione continua di evidenze sempre più impressionanti. Una memorabile è datata 8 settembre 2020, giorno in cui il quotidiano *The Guardian* pubblica un articolo intitolato *A robot wrote this entire article. Are you scared yet, human?*, effettivamente scritto da un NLM (GPT-3).

L'aspetto ironico è che l'editoriale avrebbe come tema proprio il dissipare inquietudini, era volto a convincere sull'inoffensività dell'IA, nonostante il circolare di visioni apocalittiche. Ma proprio il fatto che l'articolo fosse stato realizzato da una mente artificiale, con pretese di argomentare e convincere mediante l'uso del linguaggio, costituisce l'elemento inquietante. Si è assistito ad un miscuglio di indubbie curiosità ed ammirazioni per questo progresso, unitamente a preoccupazioni, se non proprio irritazioni, per lo scenario che si andava prospettando. Le reazioni non si fecero

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

attendere, e tra il 2018 e il 2021 si è accumulato un impressionate coro di critiche. Generalmente quel che viene imputata è l'assenza di capacità mentali, nascoste dietro una pura apparenza di conoscere il linguaggio. Come emerso da un'analisi puntuale di questo corpo di critiche (Perconti and Plebe 2023), esse in buona parte ricalcano quelle dei critici dell'idea originale di Turing, ma i toni non sono più da salotto. Il fenomeno è davanti agli occhi di tutti, l'invasione da parte delle macchine del mondo mentale è ora una preoccupazione concreta, e le critiche sono diventate decisamente più accorate. Si rintraccia una varietà di categorie di critiche. Si ritrovano, per esempio, posizioni che pregiudizialmente negano ogni possibilità di intelligenza ai NLM, e all'IA in generale, semplicemente in quanto computazioni (Bishop 2021), realizzate da macchine (Larson 2021). Vi sono anche dei critici che nutrono simpatie per l'IA, ma ne prediligono l'ispirazione razionalista, tendendo pertanto a negare l'attribuzione di intelligenza ai NLM non tanto per il pregiudizio di essere computazioni, ma di essere reti neurali artificiali, pertanto non sistemi basati su regole razionali (Pearl and Mackenzie 2018; Landgrebe and Smith 2019; Marcus and Davis 2019). Da un punto di vista strettamente filosofico non si registrano sostanziali progressi rispetto al dibattito che si era sviluppato negli anni '90, al contrario pare ora più circoscritto e superficiale. L'omaggio alla *stanza cinese* di Searle è immancabile: nell'analisi sopra citata (Perconti and Plebe 2023) viene rinvenuta in ben dodici dei lavori esaminati, ignorandone però l'ampiezza e sofisticazione del dibattito che ne era scaturito. Vi è stato persino un tentativo di emulazione, da parte di (Bender and Koller 2020), che inventano una loro versione dell'esperimento mentale di Searle. Stavolta protagonista non è Bender in persona, ma un polipo. L'animale se ne sta in fondo al mare, vicino ad un cavo di comunicazione attraverso cui conversano due parlanti inglesi, sfortunatamente naufragati su due lontane isole, per fortuna ben dotate di interconnessione. Il polipo ha curiosità per le telecomunicazioni, e quindi impara in fretta come gli impulsi elettrici che transitano nel cavo in una direzione siano seguiti da altri treni di impulsi nell'altra direzione. Ad un certo punto ha acquisito una tale confidenza con le sequenze di impulsi da decidere di tagliare il cavo, e provare lui stesso a mandare segnali in risposta ad uno dei due naufraghi. Il quale, dice Bender, potrebbe anche continuare a credere che dall'altra parte del cavo ci sia il suo sventurato amico, per quanto se la cava bene il polipo, che però non capisce nulla della conversazione in corso. E' evidente quanto il polipo della storia sia ancor più improbabile del manuale nella "stanza cinese", ed è ben difficile sia destinato ad altrettanta fortuna. Tuttavia Dennett sembra aver ragione nel segnalare l'attrazione delle *intuition pump*, infatti il lavoro di Bender e Koller ha avuto un notevole successo, ottenendo il premio come miglior lavoro al convegno annuale della *Association for Computational Linguistics* e ad oggi vanta quasi 500 citazioni.

Altre critiche hanno seguito invece una via più empirica, puntando a verificare sperimentalmente carenze *intellettive* del NLM. Anche i lavori di Bender comprendono ampie sezioni con minuziose collezioni di casi di conversazioni in cui il computer fornisce risposte sbagliate, soprattutto quando si tratta di errori che difficilmente farebbe un parlante umano. Di per se si tratta dell'ordinaria e fondamentale attività di esplorazione e verifica delle capacità dei vari modelli del linguaggio, in cui è importante individuare, più delle risposte corrette, quelle sbagliate. Si trasforma invece in retorica ideologica quando i casi di errore vengono cristallizzati come testimonianze inoppugnabili del fallimento complessivo nel dotare un computer del linguaggio umano. E' la pratica che il linguista computazionale Samuel (Bowman 2022) ha chiamato *the dangers of underclaiming*, l'esibizione poco scientifica di casi di errore allo scopo di denigrare i modelli del linguaggio. Il fenomeno peggiore è che molte di queste critiche usano come supporto critiche precedenti, citate sempre come fatti

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

empirici consolidati. In questo modo continuano ad essere presentati come errori casi ampiamente superati dal continuo e rapido progresso di questi modelli. Con il passare degli anni questa pratica si sta esaurendo in modo naturale, oggi è diventato arduo offrire evidenze empiriche dell'impaccio dei sistemi artificiali con il linguaggio umano, oramai la competenza di un parlante comune è quotidianamente replicata da agenti conversazionali basati sui NLM. E' sintomatico quanto le critiche siano spesso espressione dell'inquietudine posta dall'attuale IA, e infatti in varie di queste si coniuga la negazione di sue genuine capacità linguistiche, con la denuncia del suo potenziale pericolo (Bender et al. 2021). A ben vedere i due aspetti dovrebbero essere divergenti: se le pretese linguistiche dell'IA sono pura illusione, e non vi è dietro nulla che possa lontanamente competere con la mente umana, allora non ci si dovrebbe nemmeno preoccupare troppo.

Curiosamente pare che le inquietudini poste dall'IA siano prese molto sul serio in ambito intellettuale, giornalistico, e in parte anche politico, ma al momento emergano poco tra la gente comune. Mentre LaMDA e diversi altri sistemi di dialogo sono rimasti noti solo al mondo della ricerca, ChatGPT ha esposto al mondo intero come oggi il computer sappia conversare. Si è trattato di una sorta di implicito Turing test corale. Si è andati oltre, Turing è oramai dimenticato, e ChatGPT deve il suo apprezzamento al saper conversare esattamente come un umano, anzi, come un umano dalla cultura prodigiosa. Ha stabilito il record assoluto di prodotto digitale con la più rapida crescita di adesioni, avendo raggiunto un milione di utenti dopo solo una settimana di attività dal suo lancio a fine novembre 2022, e attualmente riceve 300 milioni di visite al mese. Secondo un primo studio sulle sensazioni degli utenti ChatGPT basato su interazioni Twitter (Haque et al. 2022), un 80% degli utenti attribuiscono una genuina intelligenza al sistema, e oltre il 90% trova il conversarci un ottimo intrattenimento e uno stimolo per la creatività personale. Non si vuol prendere questi dati preliminari come segno di una trasformazione culturale, e una disposizione della gente ad estendere l'eccezionalismo umano a simpatici compagni di conversazione artificiali. Pur se i numeri degli utenti di ChatGPT sono impressionanti, si tratta sempre di una piccola frazione della popolazione umana, ed è probabile che chi nutra sospetti e preoccupazioni nei suoi confronti, semplicemente non lo utilizzi.

4. Una mente inspiegabile

Ashish Vaswani, il giovane ricercatore indiano formatosi in informatica all'università di Mesra, e poi spostatosi in California dove ha conseguito il dottorato, avrà sicuramente, come qualunque suo collega, nutrito l'ambizione di sviluppare modelli destinati al successo. Ma ne lui ne i suoi sei collaboratori potevano mai immaginare che la loro invenzione, il Transformer (Vaswani et al. 2017), avrebbe trascinato l'IA a padroneggiare il linguaggio umano. Come descritto in §3.3, quel progetto, nato nell'ambito del Google Brain team, aveva come obiettivo unicamente il progredire nella traduzione automatica di testi da una lingua ad un'altra. Le architetture Transformer sono di una semplicità disarmante se comparate con le sofisticazioni, escogitate nell'ambito del NLP, per risolvere le innumerevoli insidie del linguaggio naturale, facendo appello alla loro analisi linguistica. Come sia possibile che il Transformer abbia avuto in pochi anni un tale successo, nemmeno lontanamente avvicinato da più di mezzo secolo di intense ricerche in NLP, è del tutto misterioso.

Occorre dire che non si tratta di un mistero isolato, i NLM fanno parte della grande famiglia del DL, a sua volta privo di una conclamata spiegazione del suo successo, in particolare del salto di

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

prestazioni nel transitare dalle reti neurali artificiali degli anni '80 a quelle "profonde" (Plebe and Grasso 2019). In qualche modo l'indecifrabilità dei modelli neurali artificiali li accomuna al cervello, da cui hanno preso ispirazione. Nonostante l'impressionante progresso delle neuroscienze nel secolo scorso, per la maggior parte delle funzioni cognitive non esistono spiegazioni su come siano realizzate dai neuroni biologici. Sono disponibili numerose ed importanti informazioni sulla localizzazione di funzioni, nonché sulle loro connessioni con altre parti del cervello, ma quali calcoli neurali conducano a tali funzioni rimane in genere sconosciuto. Tra le pochissime eccezioni possiamo citare la visione a basso livello (Hubel and Wiesel 2004) e la rappresentazione di luoghi nell'ippocampo (O'Keefe and Burgess 2005). Vi sono comunque delle differenze che rendono particolarmente imbarazzante non saper spiegare come funzionano i modelli neurali artificiali. Anzitutto hanno un'architettura estremamente più semplice rispetto al cervello, e poi sono progettati e realizzati da esseri umani. Tentare di spiegare come funzionino diventa ancor più impellente nel caso del NLM, che minano all'ultima risorsa per cui ci sentiamo eccezionali nel mondo, il linguaggio. Se dobbiamo condividerla con artefatti informatici, almeno vorremo sapere come sia possibile.

In diversi hanno avvertito questa esigenza, per esempio Stephen (Wolfram 1988), e i tentativi di porre rimedio al vuoto di spiegazione si moltiplicano, anche se al momento con risultati molto preliminari e limitati. Le direzioni di ricerca sono disparate, si sono raggruppate qui a seguito in tre tipi diversi di spiegazione.

4.1 *Psicologia per le macchine*

Un genere di spiegazione consolidato in scienze cognitive è quello denominato *funzionale* (Cummins 1983), in cui una certa capacità mentale viene spiegata cercando di individuare funzioni più semplici in grado complessivamente di realizzarla. L'analisi della disposizione di un individuo a realizzare, secondo certe condizioni e determinati input, una funzione, ha costituito nel tempo, un formidabile repertorio di protocolli sperimentali in psicologia. Non tutti, ma una rilevante parte di questi protocolli consistono esclusivamente in interazioni di tipo linguistico tra lo sperimentatore e il soggetto umano. Ecco quindi l'idea di scrutare la misteriosa mente dei NLM "fingendo" che siano dei soggetti umani, e sottoponendoli a esperimenti, impiegando i consolidati strumenti metodologici della psicologia. In modo preteorico diversi studi si sono mossi effettivamente lungo questa strada, ma una sua esplicita proposizione e giustificazione teorica è stata per la prima volta formulata da Thilo (Hagendorff 2023), che ha battezzato questo nuovo filone scientifico *Machine Psychology*.

Una funzione psicolinguistica particolarmente rilevante è quella etichettata come *tracciamento di entità del discorso* (Groenendijk and Stokhof 1991), essenziale per la comunicazione linguistica sociale, che comprende un numero di distinte abilità, come l'individuazione dell'inserimento di entità nuove in un discorso, la risoluzione delle co-referenze, e il tracciamento di cambi di stato delle varie entità di discorso. (Kim and Schuster 2023) hanno progettato una serie di esperimenti derivati da quelli classici in psicolinguistica, per verificare se NLM posseggano, e in che misura, la capacità di tracciamento di entità del discorso. Per esempio nei racconti si parla di scatole contenenti oggetti, che via via vengono tolti, aggiunti, o scambiati. Tra i vari NLM analizzati solamente GPT-3.5 dimostra un buon livello di questa funzione.

Una funzione prettamente mentale è quella nota come *induzione di proprietà*, costitutiva del

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

ragionamento induttivo, e la sua comparsa nei bambini demarca un momento importante dello sviluppo delle loro capacità cognitive (Carey 1985). Questa funzione consiste nel saper estendere o meno certe proprietà, sicuramente possedute da certe categorie di oggetti, ad altre categorie. (Han et al. 2023) hanno sfruttato il protocollo sperimentale messo a punto da (Osherson et al. 1990) per valutare l'induzione di proprietà, con esempi in categorie di animali e piante, applicandolo anziché ad umani a diversi NLM. Di questi GPT-4 è quello che ha fornito risultati molto vicini a quelli ottenuti da soggetti umani.

Particolarmente interessante è il caso della teoria della mente, considerato elemento centrale della cognizione sociale e dell'autocoscienza (Carlson, Koenig, and Harms 2013), di cui si dispone di una valutazione scientifica effettuata da (Kosinski 2023), impiegando i classici test di falsa credenza impiegati nei bambini. I risultati hanno mostrato che GPT-1 non possedeva nessuna teoria della mente, GPT-3 nella versione 2020 dava qualche risultato ma decisamente scarso, mentre la versione del gennaio 2022 superava il 70% dei test, e il modello davinci-003 del novembre 2022 ne risolveva ben il 93%. Naturalmente le diverse versioni di GPT non erano affatto dedicate a raffinare questo specifico comportamento, ma erano un generico aumento della sua parametrizzazione, pertanto Kosinski può correttamente speculare sulla teoria della mente come una sorta di proprietà emergente dal raffinamento delle capacità linguistiche. È stato davvero sorprendente constatare segni di teoria della mente nei NLM, conseguentemente vi sono stati immediati tentativi di repliche, con risultati divergenti, e ne è nato un vero e proprio piccolo filone di ricerca a se stante (Brunet-Gouet, Vidal, and Roux 2023; Holterman and van Deemter 2023; Ullman 2023; Trott et al. 2023).

4.2 *Confrontando il cervello*

I semplici stratagemmi di cui è composto il Transformer, descritti in §3.3, sono pura invenzione dei ricercatori guidati da Vaswani, tutti di estrazione informatica, senza nessuna conoscenza di neuroscienze, e senza la benché minima ispirazione da qualche aspetto del cervello umano. Pur consapevoli di questo, alcuni studiosi hanno provato a verificare a posteriori, se il funzionamento dei NLM presentasse qualché affinità con il modo con cui il cervello umano processa il linguaggio. Ne sono emerse evidenze davvero sorprendenti.

Il materiale preparatorio per i confronti tra NLM e cervello lo hanno realizzato (Nastase et al. 2021), raccogliendo immagini fMRI (*functional Magnetic Resonance Imaging* in oltre 300 soggetti, mentre ascoltavano 27 diverse storie. Le immagini erano temporalmente sincronizzate sia alle varie parole che componevano i racconti, sia al dettaglio dei fonemi che si succedevano nel tempo. (Caucheteux, Gramfort, and King 2022) utilizzarono questi dati per verificare se il modello GPT-2 andava in qualche modo d'accordo con quel che succedeva nel cervello. Più precisamente hanno impiegato una mappatura lineare tra le attivazioni dei neuroni (artificiali) in GPT-2, nel momento in cui forniva la previsione per la prossima parola, e i valori dei voxel (i minimi volumi cerebrali a cui è associata un'attivazione nella misura fMRI) nell'istante in cui veniva ascoltata la stessa parola. I risultati mostrarono una scarsa correlazione per le aree cerebrali coinvolte nei processi a basso livello, uditivo e fonologico, ma una correlazione significativa per aree di elaborazione semantica, come il giro superiore frontale, e il giro posteriore supero-temporale.

Lo stesso gruppo di ricerca è andato oltre, provando a verificare se l'ampiezza della finestra di

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

attenzione usata da GPT-2 nel predire la prossima parola, conducesse a correlazioni differenti nel cervello (Caucheteux, Gramfort, and King 2023). È risultato che in maniera sistematica, i legami tra parole calcolati dal meccanismo attentivo del NLM, correlano maggiormente con le aree corticali frontoparietali quando erano legami distanti, e con le aree temporali quando riguardavano parole ravvicinate.

È entrato in maggior dettaglio il gruppo di ricerca a Princeton, che aveva realizzato la raccolta di immagini fMRI, cercando di discriminare diverse correlazioni delle componenti dell'attenzione dei NLM, le *attention head* descritte in §3.3 (Kumar et al. 2023). Il modello analizzato è BERT (Devlin et al. 2019), che avendo 12 strati e 12 attention head in ciascun strato, comprende 144 componenti separate di attenzione. Impossibile averne un colpo d'occhio, viene in aiuto la statistica con l'analisi dei componenti principali, che proietta i 144 valori nelle sole due dimensioni più importanti. In due dimensioni si ragiona bene, ed è stato possibile da un lato verificare i punti in questo spazio che genere di funzione linguistica principale esplicano in BERT, e dall'altro con quali aree cerebrali correlano. Ne sono emerse diverse informazioni, per esempio le attention head che predicono l'esistenza di una relazione con un soggetto nominale correlano con il giro medio-frontale e non con la porzione dorso-mediale della corteccia prefrontale; la relazione con l'oggetto di una proposizione ben correla con la corteccia supero-temporale posteriore; la relazione con l'oggetto diretto di un verbo è fortemente correlata con il giro angolare e in modo debole con la parte ventro-mediale della corteccia prefrontale.

4.3 *Spiegazioni meccanicistiche*

È sicuramente illuminante scoprire che i NLM realizzano in qualche misura alcune delle funzioni cruciali studiate in psicologia e scienze cognitive, accennate in §4.1. Ma non ci dice nulla su come questo succeda, tra i semplici calcoli neurali realizzati nel Transformer e le funzioni prima descritte vi è un abisso esplicativo. Forse ancor più intrigante è scoprire che qualcosa che succede nei NLM va di pari passo con qualcos'altro che si verifica nel cervello quando si processa la stessa narrazione. Ma qui non solo non si inizia a capire qualcosa, forse il mistero addirittura si infittisce. Sapere che in modo sistematico le relazioni con un soggetto nominale attivate nelle attention head di BERT vanno a braccetto con giro medio-frontale, ma non con la porzione dorso-mediale della corteccia prefrontale, non solo non ci spiega come faccia BERT a identificare le relazione con soggetti nominali, ci pone l'ulteriore interrogativo su come mai le sue attivazioni assomiglino ai segnali circolanti nel giro medio-frontale del cervello.

Nell'ambito della biologia in generale, ma ancor più specificatamente in neuroscienza, il genere di spiegazione ritenuto in filosofia della scienza più pertinente è quello denominato *meccanicistico*, che consiste nell'identificare parti, in un sistema complesso, responsabili del fenomeno sotto indagine, e di studiare quali relazioni tra tali parti siano in grado di dar luogo al fenomeno (Machamer, Darden, and Craver 2000; Craver and Darden 2001; Craver 2007; Kaplan and Craver 2011). È una terza strada che alcuni hanno iniziato a percorrere nel tentativo di svelare i misteri del Transformer. Si tratta di un percorso estremamente arduo, in cui viene messa in campo una poderosa matematica per tentare di analizzare quel che succede nelle varie componenti del Transformer, di cui si dispensa qui il lettore. Si illustrerà solo uno dei primissimi risultati finora acquisiti, ancora una piccola goccia nel mare di mistero che ricopre il Transformer.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Il primo tentativo di analisi meccanicistica dei NLM è dovuto a (Elhage et al. 2021), ed ancora oggi rimane il più profondo e rivelatore. La strategia è stata di partire dal minimo numero possibile di componenti di un Transformer, iniziando da un solo strato e una sola attention head, cercando via via di individuare "parti", in questo caso componenti matematiche, potenzialmente responsabili di qualche fenomeno interessante. È stato nel passare da uno a due strati che un fenomeno è emerso, denominato *induction head*. Consiste nel predire una parola *B* quando la parola corrente è *A*, se nella sequenza di parole precedenti si è verificata la stessa successione *AB*. Probabilmente chiunque sia incuriosito dai misteri della mente artificiale che comprende e produce linguaggio, troverà poco serio chiamare "interessante" un fenomeno così elementare, e non si può certo dargli torto. La scoperta va comunque valorizzata per due motivi. Anzitutto per questo fenomeno è stato individuato in modo preciso il meccanismo che lo realizza, con certe "parti" matematiche derivate dalle matrici che calcolano le attivazioni dell'attenzione nei due livelli di questo mini-Transformer. Inoltre, lo stesso gruppo ha studiato la stretta correlazione tra *induction head* e l'emergere di un fenomeno di ben più alto livello, conosciuto come *in-context learning* (Olsson et al. 2022). E' il fenomeno per cui, più si aggiunge un ricco contesto nell'interrogare un NLM, e più perspicue risultano le sue risposte. Questo costituisce un forte sintomo di una sorta di "pensiero" che guida la produzione linguistica del NLM. Si è ben lontani da produrre una catena causale che dal fenomeno *induction head* conduca al *in-context learning*, però Olsson e coautori hanno studiato come, in modelli Transformer completi, l'emergere del *in-context learning* durante l'addestramento, coincida precisamente con l'emergere del fenomeno *induction head*.

5. Conclusioni

Da diversi anni l'esplosione dell'IA l'ha resa argomento abituale sia nella stampa scientifica che in quella giornalistica, con un'ulteriore impennata alla sua popolarità quest'anno, a seguito della disponibilità di ChatGPT. I punti di vista da cui parlare di IA sono variegati, spaziano da prospettive tecniche, economiche, sociali, politiche, con il tema etico sicuramente il più gettonato. Qui ci si è concentrati su due specifici aspetti, ritenuti filosoficamente di rilievo. Uno riguarda la constatazione che, forse, abbiamo a che fare con una nuova entità dotata di una mente, e soprattutto della facoltà a cui siamo particolarmente affezionati, il linguaggio, e questo inevitabilmente innesca inquietudini. L'altra è l'imbarazzante mistero su come, certi semplici calcoli neurali che costituiscono i NML, diano luogo a capacità linguistiche. I due aspetti non sono del tutto indipendenti, l'ignoto risulta inevitabilmente sempre più inquietante, poter disporre almeno della conoscenza su come funziona ciò di cui ci preoccupiamo, sarebbe già un certo elemento di rassicurazione.

Riferimenti bibliografici

Beaver, Patrick. 1974. *Victorian Parlor Games*. Edinburgh: Thomas Nelson.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 Acm Conference on Fairness, Accountability, and Transparency*, 610–623. ACM.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Bender, Emily M., and Alexander Koller. 2020. "Climbing Towards Nlu: On Meaning, Form, and Understanding in the Age of Data." In *58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Somerset (NJ): Association for Computational Linguistics.

Bishop, J. Mark. 2021. "Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It." *Frontiers in Psychology* 11: 513474.

Block, Ned. 1981. "Psychologism and Behaviorism." *Philosophical Review* 90: 5–43.

Bowman, Samuel R. 2022. "The Dangers of Underclaiming: Reasons for Caution When Reporting How Nlp Systems Fail." In *Proceedings of the 60th Meeting of the Association for Computational Linguistics*, 1:7484–99. Association for Computational Linguistics.

Bridgman, Percy W. 1927. *The Logic of Modern Physics*. New York: Macmillan.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and Prafulla Dhariwal et.al. 2020. "Language Models Are Few-Shot Learners." *arXiv* abs/2005.14165.

Brunet-Gouet, Eric, Nathan Vidal, and Paul Roux. 2023. "Can a Conversational Agent Pass Theory-of-Mind Tasks? A Case Study of ChatGPT with the Hinting, False Beliefs, and Strange Stories Paradigms." *Zenodo* DOI 10.5281/zenodo.8009748.

Carey, Susan. 1985. *Conceptual Change in Childhood*. Cambridge (MA): MIT Press.

Carlson, Stephanie M., Melissa A. Koenig, and Madeline B. Harms. 2013. "Theory of Mind." *Wiley Interdisciplinary Reviews: Cognitive Science* 4: 391–402.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Remi King. 2022. "Deep Language Algorithms Predict Semantic Comprehension from Brain Activity." *Scientific Reports* 12: 16327.

———. 2023. "Evidence of a Predictive Coding Hierarchy in the Human Brain Listening to Speech." *Nature Human Behaviour* 7: 430–41.

Chalmers, David. 2023. "Could a Large Language Model Be Conscious?" *arXiv* abs/2303.07103.

Chomsky, Noam. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper; Row Pub. Inc.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford (UK): Oxford University Press.

Craver, Carl F., and Lindley Darden. 2001. "Discovering Mechanisms in Neurobiology: The Case of Spatial Memory." In *Theory and Method in Neuroscience*, edited by Peter Machamer, Rick Grush, and P. McLaughlin. Pittsburgh (PA): Pittsburgh University Press.

Cummins, Robert. 1983. *The Nature of Psychological Explanation*. Cambridge (MA): MIT Press.

Dennett, Daniel C. 1980. "The Milk of Human Intentionality." *Behavioral and Brain Science* 3: 429–30.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

———. 2012. *Intuition Pumps and Other Tools for Thinking*. New York: W. W. Norton & Company.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–86. Association for Computational Linguistics.

Dreyfus, Hubert. 1972. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper; Row Pub. Inc.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. 2021. "A Mathematical Framework for Transformer Circuits." *Transformer Circuits Thread*.

Elman, Jeffrey L. 1990. "Finding Structure in Time." *Cognitive Science* 14: 179–221.

Finlay, Barbara L., and Alan D. Workman. 2013. "Human Exceptionalism." *Trends in Cognitive Sciences* 17: 199–201.

Groenendijk, J., and M. Stokhof. 1991. "Dynamic Predicate Logic." *Linguistics and Philosophy* 14: 39–100.

Hagendorff, Thilo. 2023. "Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods." *arXiv* abs/2303.13988.

Han, Simon J., Keith Ransom, Andrew Perfors, and Charles Kemp. 2023. "Inductive Reasoning in Humans and Large Language Models." *arXiv* abs/2306.06548.

Haque, Mubin Ul, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. "I Think This Is the Most Disruptive Technology: Exploring Sentiments of Chatgpt Early Adopters Using Twitter Data." *arXiv* abs/2212.05856.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. 2002. "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" *Science* 298: 1569–79.

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 28: 504–7.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors." *arXiv* abs/1207.0580.

Hinton, Geoffrey, and Richard S. Zemel. 1994. "Autoencoders, Minimum Description Length and Helmholtz Free Energy." In *Advances in Neural Information Processing Systems*, 3–10.

Hochreiter, Sepp, and Jurgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9: 1735–80.

Holterman, Bart, and Kees van Deemter. 2023. "Does ChatGPT Have Theory of Mind?" *arXiv*

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

abs/2305.14020.

Hubel, David H., and Torsten N. Wiesel. 2004. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford (UK): Oxford University Press.

Kaplan, David M., and Carl F. Craver. 2011. "Towards a Mechanistic Philosophy of Neuroscience." In *Continuum Companion to the Philosophy of Science*, edited by Steven French and Juha Saatsi, 268–92. London: Continuum Press.

Kim, Najoung, and Sebastian Schuster. 2023. "Entity Tracking in Language Models." *arXiv* abs/2305.02363.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." In *Proceedings of International Conference on Learning Representations*.

Kosinski, Michal. 2023. "Theory of Mind May Have Spontaneously Emerged in Large Language Models." *arXiv* abs/2302.02083.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, 1090–8.

Kumar, Sreejan, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. 2023. "Reconstructing the Cascade of Language Processing in the Brain Using the Internal Computations of a Transformer-Based Language Model." *bioRxiv* DOI: 10.1101/2022.06.08.495348.

Kuutti, Sampo, Saber Fallah, Richard Bowden, and Phil Barber. 2019. "Deep Learning for Autonomous Vehicle Control – Algorithms, State-of-the-Art, and Future Prospects." *Synthesis Lectures on Advances in Automotive Technology* 3: 1–80.

Landgrebe, Jobst, and Barry Smith. 2019. "Making Ai Meaningful Again." *Synthese* DOI 10.1007/s11229-019-02192-y: 1–21.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge (MA): Harvard University Press.

Lovejoy, Arthur O. 1936. *The Great Chain of Being: A Study of the History of an Idea*. Cambridge (MA): Harvard University Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67: 1–84.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting Ai: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955."

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, 3111–9.

Nastase, Samuel A., Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, et al. 2021. "The 'Narratives' FMRI Dataset for Evaluating Models of Naturalistic Language Comprehension." *Scientific Data* 8: 250.

O'Keefe, Jhon, and Neil Burgess. 2005. "Dual Phase and Rate Coding in Hippocampal Place Cells: Theoretical Significance and Relationship to Entorhinal Grid Cells." *Hippocampus* 15: 825–989.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, et al. 2022. "In-Context Learning and Induction Heads." *Transformer Circuits Thread*.

Osherson, Daniel N., Edward E. Smith, Ormond Wilkie, Alejandro López, and Eldar Shafir. 1990. "Category-Based Induction." *Psychological Review* 97: 185–200.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why – the New Science of Cause and Effect*. New York: Hachette Book Group.

Penn, Derek C., Keith J. Holyoak, and Daniel J. Povinelli. 2008. "Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds." *Behavioral and Brain Science* 31: 109–78.

Pepperberg, Irene M. 1999. *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Cambridge (MA): Harvard University Press.

Perconti, Pietro, and Alessio Plebe. 2023. "Do Machines Really Understand Meaning? (Again)." *Journal of Artificial Intelligence and Consciousness* 10: 181–206.

Pinker, Steven. 1994. *The Language Instinct. How the Mind Creates Language*. New York: William Morrow.

Pinker, Steven, and Alan Prince. 1988. "On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition." *Cognition* 28: 73–193.

Plebe, Alessio, and Giorgio Grasso. 2019. "The Unbearable Shallow Understanding of Deep Learning." *Minds and Machines* 29: 515–53.

Preston, John, and Mark Bishop, eds. 2002. *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford (UK): Oxford University Press.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323: 533–36.

Rumelhart, David E., and James L. McClelland. 1986a. "On Learning the Past Tenses of English Verbs." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by David E. Rumelhart and James L. McClelland, 2:216–71. Cambridge (MA): MIT Press.

———, eds. 1986b. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge (MA): MIT Press.

Rivista di Storia delle Idee 12:2 (2023) pp. 168-187
ISSN. 2281-1532 http://www.intrasformazione.com
DOI 10.4474/RCR/12/02/RCR648/19
Patrocinata dall'Università degli Studi di Palermo

Russell, Bertrand. 1932. "In Praise of Idleness."

Savage-Rumbaugh, Sue, Stuart Shanker, and Talbot Taylor. 1998. *Apes, Language and the Human Mind*. Oxford (UK): Oxford University Press.

Searle, John R. 1980. "Mind, Brain and Programs." *Behavioral and Brain Science* 3: 417–24.

Srnicek, Nick, and Alex Williams. 2015. *Inventing the Future: Postcapitalism and a World Without Work*. London: Verso.

Tan, Kar-Han, and Boon Pang Lim. 2018. "The Artificial Intelligence Renaissance: Deep Learning and the Road to Human-Level Machine Intelligence." *APSIPA Transactions on Signal and Information Processing* 7: e6.

Trott, Sean, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. "Do Large Language Models Know What Humans Know?" *Cognitive Science* 47: e13309.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59: 433–60.

Ullman, Tomer D. 2023. "Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks." *arXiv* abs/2302.08399.

VanRullen, Rufin. 2017. "Perception Science in the Age of Deep Neural Networks." *Frontiers in Psychology* 8: 142.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 6000–6010.

Veselý, Karel, Arnab Ghoshal, Lukas Burget, and Daniel Povey. 2013. "Sequence-Discriminative Training of Deep Neural Networks." In *Conference of the International Speech Communication Association*, 2345–9.

Weizenbaum, Joseph. 1966. "Eliza – a Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the Association for Computing Machinery* 9: 36–45.

Wickelgren, Wayne A. 1969. "Context Sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior." *Psychological Review* 76: 1–15.

Winograd, Terry. 1972. *Understanding Natural Language*. New York: Academic Press.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*. London: Trench, Trubner & Co.

Wolfram, Stephen. 1988. *What Is Chatgpt Doing ...and Why Does It Work*. Champaign (IL): Wolfram Media.