L'algoritmo, la scimmia e lo specchio: la scorciatoia di Machina sapiens secondo Nello Cristianini

di Matilde Fontanin

Introduzione

> Quando qualcosa non è noto (o non mi è chiaro) lo dirò onestamente. È quello che devono fare gli scienziati, perché fare finta di sapere tutto non serve a niente.
> (MS, p. 9)[1]

Nel 2017 la parola d'ordine era 'fake news', oggi è 'intelligenza artificiale' (o IA, AI, *Artificial intelligence* che dir si voglia), due fenomeni attuali, sebbene con radici lontane nel tempo. Dal 2022 è possibile dialogare con una macchina, e tutto è cambiato.
Due recenti saggi divulgativi spiegano come funzionano, nell'ordine, i meccanismi di raccomandazione[2] e l'IA generativa[3]. L'autore è professore di intelligenza artificiale all'Università di Bath, e, come il settore della IA si sviluppa all'incrocio di diverse discipline, è una persona al confine, in senso disciplinare e biografico. Uomo di scienza, lo si scopre amante delle lettere ed estimatore delle biblioteche[4], ha una for-

MATILDE FONTANIN, Università degli studi di Trieste, e-mail: mfontanin@gmail.com
Ultima consultazione siti web: 10 febbraio 2025

1 Date le numerose citazioni, si utilizzeranno le sigle LS per Nello Cristianini, *La scorciatoia : come le macchine sono diventate intelligenti senza pensare in modo umano*. Bologna: Il Mulino, 2023; e MS per Nello Cristianini, *Machina sapiens : l' algoritmo che ci ha rubato il segreto della conoscenza*. Bologna: Il Mulino, 2024.

2 Nello Cristianini, *La scorciatoia* cit.

3 Nello Cristianini, *Machina sapiens* cit.

4 Lo ha dichiarato rendendosi disponibile per il gruppo di lettura animato dall'Osservatorio AIB sulla information literacy, il 18 giugno 2024. Inoltre, conosce le potenzialità delle collezioni, se non altro per avervi investigato, si veda Nello Cristianini; Thomas Lansdall-Welfare; Gaetano Dato, *Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914*, «Historical methods: a journal of quantitative and interdisciplinary history», 51 (2018), n. 3, p. 139–164, DOI: 10.1080/01615440.2018.1443862.

AIB studi, vol. 64 n. 3 (settembre/dicembre 2024), p. 373-390. DOI 10.2426/aibstudi-14131
ISSN: 2280-9112, E-ISSN: 2239-6152 - Copyright © 2024 Matilde Fontanin

mazione classica[5] e un certo gusto per la parola, ama le incursioni nella narrativa[6]. È nato su un confine, quello di Gorizia, che nel 1947 aveva diviso non solo due Stati (e con essi Est e Ovest del mondo)[7] ma anche le famiglie, la città dalla campagna, le fattorie dalle stalle[8]. Forse, più che di confine, si dovrebbe parlare di frontiera: non una linea che separa, ma un territorio d'incontro dove è difficile scindere nettamente cose e persone, dove costruire il futuro[9], e forse è anche per la sua abitudine alla frontiera che l'autore è così efficace nel raccontare a chi formazione scientifica non ha, usando esempi incisivi tratti dalla quotidianità.

Se «il mezzo è il messaggio»[10], l'influenza che le tecnologie esercitano sulla raccolta e condivisione della conoscenza dipende dallo spazio che viene loro concesso. La tecnologia non muta l'essenza della democrazia, solo ne incrementa opportunità e rischi, per questo già nel 1997 Stefano Rodotà[11] parlava di cittadinanza «senza aggettivi». Per esercitarla a pieno occorrono consapevolezza e competenza informativa (o *information literacy*) «un diritto umano fondamentale in un mondo digitale»[12], che consiste nella «capacità di pensare in modo critico ed esprimere giudizi equilibrati su qualsiasi informazione che troviamo e utilizziamo»[13]. In altre parole, occorre comprendere il contenuto informativo indipendentemente dal suo formato, e, dato che oggi la IA entra prepotentemente nella produzione dei contenuti[14], occorre capire come funziona e che non è un oracolo.

Per questo qui si indossa innanzitutto la lente della AI literacy. Si definisce *AI literate* la persona con un grado di alfabetizzazione (o competenza) rispetto alla IA sufficiente a comprendere, utilizzare, monitorare e riflettere criticamente sulle sue appli-

5 Ci sono svariati riferimenti al Liceo classico.

6 Cristianini fa cenno a *Erewhon*, di Samuel Butler (1872), ricordato da Alan Turing, al racconto *La zampa di scimmia* di J.J. Jacobs, ricordato da Norbert Wiener e, non strettamente narrativo, al dialogo tra il formichiere e il granchio in «... *mirmecofuga»*, in Douglas R. Hofstadter, *Gödel, Escher, Bach: un' eterna ghirlanda brillante*. Milano: Adelphi, 1985, p. 345-346. Si tornerà man mano su alcuni testi.

7 Da qui discende la nomina di Gorizia-Nova Gorica a Capitale europea della cultura 2025.

8 Si veda l'iconica foto della mucca a cavallo della linea di confine (foto Altran, 1947) <https://www.girofvg.com/images/Mucca_1947_confine_05-1080x721.jpeg>. Un vivace resoconto basato sulla storia orale, ricco di documenti e di illustrazioni è in Alessandro Cattunar, *Storia di una linea bianca: Gorizia, il confine, il Novecento*. Udine: Bottega errante, 2024.

9 *Ivi*, p. 18.

10 Marshall McLuhan, *Understanding media; the extensions of man*. New York: McGraw-Hill, 1964.

11 Stefano Rodotà, *Tecnopolitica: la democrazia e le nuove tecnologie della comunicazione*. Roma: Laterza, 1997.

12 IFLA, *Beacons of the information society: the Alexandria proclamation on information literacy and lifelong learning*. The Hague: IFLA, 2005.

13 *CILIP Definition of information literacy 2018*, trad. in Maurizio Lana, *Introduzione all' information literacy: storia, modelli, pratiche*. Milano: Editrice bibliografica, 2020, p. 69.

14 'Contenuti' è impiegato qui nell'accezione diffusa, ma Cristianini pone una domanda stimolante: «perché chiamiamo le persone 'utenti' e le espressioni della loro arte e cultura 'contenuti'? Chi mai descriverebbe i grandi vini della nostra città come 'contenuto'? Forse un venditore di bottiglie». (LS, p. 204).

cazioni[15]. Gli esperti devono saper sviluppare le applicazioni, ma per le persone comuni è sufficiente saper utilizzare consapevolmente gli strumenti IA disponibili. Per promuovere l'information literacy[16], la professione bibliotecaria deve prima acquisire determinate conoscenze e competenze. La lettura dei due saggi di Cristianini offre strumenti per la AI literacy, qui osservati con un focus sui disordini dell'ecosistema informativo[17] rispetto alla IA testuale, tralasciando volutamente le mille altre applicazioni della IA, dalle auto a guida autonoma alla gestione finanziaria.

Si evidenzieranno ora alcuni punti posti in rilievo da Cristianini, delineando la storia della IA per quel tanto che è necessario a collocarli nel contesto. Di seguito, si discuteranno i riflessi sui disturbi dell'ecosistema informativo, il rapporto tra IA e umani e alcune ricadute sulle discipline bibliotecarie. Va anteposta un'avvertenza: l'ispirazione è tratta dai due saggi, ma è ovviamente restituita da una lettura personale, come di chi scrive sono alcuni suggerimenti di lettura, nonché eventuali errori ed omissioni.

L'IA e la svolta statistica

> Comprendere le cose è la cura per l'ansia
> (MS, p. 93)

Cristianini pone un forte accento sulla cosiddetta svolta statistica, un vero e proprio cambio di paradigma che, se da un lato ha consentito lo sviluppo di una IA efficiente, dall'altro ne ha forgiato quelle caratteristiche che oggi destano preoccupazione. Per comprendere tale svolta è utile ripercorrere lo sviluppo dell'intero settore.

Il padre dell'intelligenza artificiale è Alan Turing, il primo a chiedersi, nel 1950, se le macchine fossero in grado di pensare[18]. Dato che la risposta implicava definire concetti complessi come quello di macchina e di pensiero, con il suo consueto pragmatismo[19] egli optò per un quesito dalla risposta misurabile, ovvero: «esistono computer digitali immaginabili che potrebbero giocare bene al gioco dell'imitazione?»[20]. Si definisce quindi 'pensante' la macchina che riesce ad ingannare un umano perlomeno nel 50% delle loro interazioni; è il cosiddetto 'test di Turing'[21], di cui oggi in

15 Matthias Carl Laupichler [*et al.*], *Artificial intelligence literacy in higher and adult education: a scoping literature review*, «Computers and education: artificial intelligence», 3 (2022), p. 100101, DOI: 10.1016/j.caeai.2022.100101.

16 IFLA FAIFE, *IFLA Code of ethics for librarians and other information workers*. The Hague: IFLA, 2012.

17 Il termine 'disordini dell'ecosistema informativo', viene comunemente riassunto da 'disinformazione' o 'fake news', ma ha un senso più ampio di entrambi, che in realtà rappresentano solo due dei fenomeni che ne fanno parte. Claire Wardle; Hossein Derakhshan, *Information disorder: toward an interdisciplinary framework for research and policymaking*. Strasbourg: Council of Europe, 2017, vol. 27. p. 109.

18 Alan Turing, *Computing machinery and intelligence*, «Mind», LIX (1950), n. 236, p. 433–460, DOI: 10.1093/mind/LIX.236.433; cit. in MS, p. 17.

19 La sua personalità è ben delineata nella biografia di Andrew Hodges, *Alan Turing: the enigma: the book that inspired the film «The imitation game»*. London: Vintage books, 2014; trad. it. di David Mezzacapa, *Alan Turing: storia di un enigma*. Torino: Bollati Boringhieri, 2014.

20 MS, p. 18.

21 Che si concretizza, tra il 1991 e il 2019, nel premio Loebner, mai assegnato, ma che metteva in palio 100.000$ per il programma in grado di ingannare il 50% dei suoi intervistatori, MS, p. 20.

rete è disponibile una simulazione «globale»[22]: due minuti per capire se l'interlocutore è umano.

L'espressione 'intelligenza artificiale' nasce solo nel 1956, all'Università di Dartmouth, nell'ambito di una *Summer school* tra le migliori menti matematiche al mondo[23]. Si ipotizza che ogni aspetto dell'apprendimento possa essere descritto in modo tanto preciso da permettere ad una macchina di simularlo, e si lavora affinché le macchine «utilizzando il linguaggio, siano in grado di dare forma ad astrazioni e concetti e di risolvere ogni sorta di problemi attualmente riservati agli umani, e che possano imparare a migliorarsi»[24].

Si vede che il paradigma della IA, fin dal suo inizio, si fonda su un concetto antropocentrico di intelligenza, ignorandone le altre forme esistenti in natura. Eppure, se intelligenza è «la capacità di un sistema di agire in modo appropriato in un ambiente incerto, dove le azioni appropriate sono quelle che aumentano le probabilità di successo»[25], si potrebbero comprendere diverse tipologie di agenti, compresi formiche e gatti.

La neonata disciplina, invece, si preoccupa solo di istruire le macchine affinché possano pensare come gli umani, e decide di muoversi in due direzioni: da un lato, devono poter comprendere e generare il linguaggio umano e dall'altro devono imparare a conoscere il mondo, così si producono «enormi database, curati a mano, di fatti, concetti e regole»[26]. Tale paradigma è detto dell'intelligenza artificiale 'vecchia maniera' (in inglese GOFAI)[27], e continua perlomeno fino agli anni Ottanta con il temporaneo successo dei Sistemi esperti e poi il loro tonfo, sancito nel 1988 da un articolo del New York Times[28]. Il problema è che per «descrivere tutti gli aspetti dell'intelligenza con tale precisione da poterli implementare in una macchina»[29] occorrono troppe regole esplicite, che vanno scritte a mano. Pur affannandosi ad etichettare articoli, aggettivi, le loro relazioni e le debite eccezioni, finiva che «le grammatiche diventavano sempre più grandi e i risultati rimanevano deboli»[30].

La svolta che Cristianini definisce «scorciatoia» arriva con Frederick Jelinek[31], che nel 1972 era entrato a far parte del *Continuous Speech Recognition Group* della IBM per sviluppare sistemi per la correzione ortografica e la trascrizione del parlato. Forte

22 Il gioco *Human or not?* è alla pagina <https://app.humanornot.ai>, MS, 57–59.

23 John McCarthy [*et al.*], *A proposal for the Dartmouth Summer research project on Artificial Intelligence*, 1955, <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>. I firmatari sono John McCarthy, Marvin Minsky di Harvard, Nathaniel Rochester della IBM e Claude Shannon della Bell.

24 *Ivi*, p. 2.

25 James S. Albus, *Outline for a theory of intelligence*, «IEEE transactions on systems, man, and cybernetics», 21 (1991), n. 3, p. 473–509, DOI: 10.1109/21.97471, cit. in LS, p. 13.

26 MS, p. 19-20.

27 «we may also speak more explicitly of what I shall call Good Old Fashioned Artificial Intelligence — GOFAI, for short» John Haugeland, *Artificial intelligence: the very idea*. Cambridge, (Mass): MIT press, 1985, p. 112.

28 Andrew Pollack, *Setbacks for Artificial Intelligence*, «The New York Times», 4 marzo 1988, <https://www.nytimes.com/1988/03/04/business/setbacks-for-artificial-intelligence.html>.

29 LS, p. 32.

30 LS, p. 28.

31 La storia di Frederick Jelinek è in LS, p. 27–31.

della propria esperienza nella ricostruzione di messaggi inquinati da 'rumore', egli abbandona l'approccio logico per quello statistico. Il linguaggio presenta delle regolarità, determinate dalla probabilità che determinate parole o sequenze di parole compaiano nel linguaggio naturale. Si pensi, ad esempio, al correttore ortografico: dato che i refusi danno origine a parole inesistenti, perciò statisticamente improbabili, ne vengono semplicemente suggerite altre, di forma similare ma ad alta frequenza. In sintesi, Jelinek sostituisce la descrizione esplicita dei linguisti con tabelle create automaticamente dalla macchina, popolate da centinaia di migliaia di parametri. La combinazione della potenza di calcolo con la disponibilità di testi fa sì che non occorra più che gli umani istruiscano la macchina, che è in grado di scoprire le regole da sé. Si tratta di una 'scorciatoia' che abbandona l'illusione di accuratezza come l'avrebbero desiderata i linguisti, accontentandosi di previsioni solo «probabilmente approssimativamente corrette»[32], che però è «tutto ciò che serve in molti casi pratici» [33], ed è meno costoso. A Jelinek è attribuita una frase che rende bene l'idea del clima: «tutte le volte che licenzio un linguista la performance del nostro sistema migliora»[34].

Del resto, Jelinek non si occupava dell'intero universo digitale, ma solo di traduzione automatica, di cui realizzerà effettivamente il primo sistema[35]. Eppure, il cambio di paradigma che propone ha effetto su tutta l'interazione online odierna, dai sistemi di raccomandazione alla IA generativa. Il superamento del bisogno di costruire modelli dettagliati del linguaggio naturale è possibile perché per addestrare le macchine sono sufficienti (ma necessarie) grandi quantità di dati, però così facendo si rinuncia a costruire teorie e ci si sposta sui risultati; si accetta che la macchina classifichi la realtà secondo logiche non umane, e non ci si interessa di capirle. Altri elementi contribuiranno in seguito alla «nuova ricetta»[36] per la IA, ma da qui in poi il paradigma prevede sempre di dare i dati alla macchina, lasciando che trovi da sola le regolarità che le servono a costruire regole.

Così, il nuovo paradigma non avrà effetti solo sulla produzione di testo, ma sull'intera rappresentazione di tutto quel mondo che con i testi viene descritto. Si arriverà in questo modo alle macchine in grado di indovinare la parola successiva in una frase, che non è poi così diverso da indovinare il prossimo[37] libro che verrà messo nel carrello online. E se quella parola fosse una diagnosi, o il nome dell'assassino alla fine di un giallo[38], vorrebbe dire che la macchina comprende il problema? Tuttavia, per arrivare a questo successivo balzo in avanti c'è bisogno di altri passi, tra i quali

32 LS, p. 31.

33 *Ibidem*.

34 Sicuramente pronunciata, anche se l'interessato non ricorda bene quando. I commenti, e le scuse, dell'autore, insieme a un'interessante storia della sua esperienza sono in Frederick Jelinek, *Some of my best friends are linguists*, «Language resources and evaluation», 39 (2005), n. 1, p. 25-34, DOI: 10.1007/s10579-005-2693-4.

35 Presentato al "Workshop on the evaluation of natural language processing systems", svoltosi al Wayne Hotel, a Wayne, in Pennsylvania, tra il 7 e il 9 dicembre 1988. Frederick Jelinek, *Applying information theoretic methods: evaluation of grammar quality*, 1988.

36 LS, p. 50.

37 LS, p. 31.

38 MS, p. 112.

l'arrivo dei *Big data*, infatti «la vita comincia ad un miliardo di esempi»[39], ovvero le ultime applicazioni sono davvero efficienti solo sopra il miliardo di parole.

Google emerge vincente dopo la bolla *Dot-com* del marzo 2000 proprio perché aveva pienamente incorporato il nuovo paradigma della IA basata sui dati (*data-driven AI*), tanto che nel 2009 fornirà il manifesto della nuova visione, 'l'irragionevole efficacia dei dati'[40]. L'articolo formalizza pratiche ormai comuni nel campo della IA: sono i dati, non le regole o i modelli, a guidare il comportamento degli agenti intelligenti; i modelli teorici sono ormai sostituiti con regolarità statistiche osservate dai dati. L'unica criticità è trovare i dati, ma ora essi sono disponibili sul web, soprattutto se ci si accontenta di quelli grezzi, non annotati da umani, e questa è la seconda scorciatoia. La terza consiste nell'utilizzare il feedback implicito[41], ossia l'annotazione dei dati con il giudizio dell'utente, tratto semplicemente dall'osservazione di ciò che fa, come quando seleziona degli articoli in un catalogo. Così faceva Amabot, il chatbot di Amazon, che basava i consigli per gli acquisti sull'analisi statistica dei comportamenti precedenti: se chi acquista A poi acquista anche C e D, al successivo acquirente di C si suggeriranno anche A e D. Il metodo è così efficace che presto i recensori umani di Amazon vengono licenziati, del resto le raccomandazioni di acquisto basate sul feedback implicito costano meno e hanno un ritorno maggiore. I lavoratori di Amazon manifesteranno la loro collera in un annuncio sul Seattle Weekly che definisce Amabot, «vecchio arnese scassato» e auspica che «la splendida confusione della carne e del sangue vincerà»[42], ma, almeno fino ad ora, hanno avuto torto.

In sintesi, a partire dagli anni 2000 le macchine sono potenti, il web fornisce i dati, e agli esseri umani si apre un campo infinito di possibilità. Il problema del recupero dell'informazione non è più trovare 'qualcosa', ma qualcosa di 'pertinente'. La competenza informativa transmediale diventa essenziale[43], si deve imparare a scartare, a navigare, a non farsi distrarre. Si devono affinare le doti di valutazione del contesto, ancor prima che dei documenti, mettendo in atto una lettura laterale[44], cioè, prima di valutare i testi, selezionare quali vale la pena di leggere. Il tempo e l'attenzione[45] delle persone diventano risorse da tutelare.

Il grande balzo è quello del web 2.0, dei social media, celebrato dal *Time* quando, nel 2006, sceglie come persona dell'anno *YOU*[46]: «la persona dell'anno sei tu.

39 LS, p. 51.

40 Alon Halevy; Peter Norvig; Fernando Pereira, *The unreasonable effectiveness of data*, «IEEE intelligent systems», 24 (2009), n. 2, p. 8-12, DOI: 10.1109/MIS.2009.36.

41 Cristianini fa risalire l'espressione almeno al 1992 e al lavoro sul *collaborative filtering* della posta elettronica di Dave Goldberg [*et al.*], *Using collaborative filtering to weave an information tapestry*, «Commun. ACM», 35 (1992), n. 12, p. 61-70, DOI: 10.1145/138859.138867. LS, p. 46.

42 La storia è nel paragrafo *Resa dei conti a Seattle*, LS, p. 38-41, la citazione a p. 38.

43 «transmedia information literacy is a core life skill» Peter Morville, *Ambient findability*. Beijing; Sebastopol (CA): O'Reilly, 2005, p. 7.

44 Sam Wineburg; Sarah McGrew, *Lateral eading: reading less and learning more when evaluating digital information*, «Teachers college record», 121 (2019), <https://papers.ssrn.com/abstract=3048994>.

45 «societies must protect, cherish and nurture humans' attentional capabilities. [...] attentional capabilities are a finite, precious and rare asset». Cfr. *The onlife manifesto: being human in a hyperconnected era*, a cura di Luciano Floridi. New York: Springer, 2014, § 4.6.

46 Lev Grossman, *You - Yes, You - Are TIME's person of the year*, «Time», 25 dicembre 2006, <https://time.com/archive/6596761/you-yes-you-are-times-person-of-the-year/>.

Benvenuto nel tuo mondo», si legge in un'atmosfera di fiducia nella democratizzazione globale, anche se Lev Grossman avverte che

> Il Web 2.0 sfrutta la stupidità delle folle tanto quanto il loro buonsenso. Alcuni commenti su YouTube fanno disperare per il futuro dell'umanità solo per l'ortografia, per non parlare dell'oscenità e del puro odio. [...] Il Web 2.0 è un enorme esperimento sociale e, come ogni esperimento che valga la pena di essere tentato, potrebbe fallire[47].

Riassumendo, l'intero campo della IA si è sviluppato sulla base di un'idea antropocentrica di intelligenza, prendendo una serie di scorciatoie possibili grazie alla potenza crescente delle macchine e all'aumento dei dati disponibili. Ci sono stati momenti di crisi, detti inverni, che hanno consentito a diverse linee di ricerca di svilupparsi lontano da attenzioni mediatiche. Tutto questo ha dato oggi vita ad un campo all'incrocio tra diverse discipline, perlopiù scientifiche, ma anche umanistiche e delle scienze sociali. Il prossimo passo a questo punto è l'IA generativa, il dialogo con un essere apparentemente senziente.

Come 'pensa' ChatGPT

> Non so come funzionino ChatGPT e i suoi cugini, non lo sa ancora nessuno (MS, p. 7)

Da novembre 2022 chiunque può conversare con una macchina. Alan Turing nel 1951[48] prevedeva che sarebbe accaduto entro la fine del XX secolo e non ha sbagliato di molto, e forse, se non fosse stato sottoposto a trattamento farmacologico[49], avrebbe rispettato la previsione[50]. L'illusione che ChatGPT e i suoi cugini siano creature capaci di pensiero autonomo e quasi omniscienti è grande, ma capire come fa a rispondere la IA generativa fa parte della AI Literacy, insieme alla consapevolezza che rimangono delle zone opache, come evidenzia la citazione in testa al paragrafo.

Innanzitutto, le macchine apprendono in modo molto diverso dagli umani. Mentre i bambini imparano un nome nuovo al primo colpo (*one-shot learning*), le macchine hanno bisogno di molti più passaggi, nei quali esaminano vasti corpora di testo per attribuire alle diverse parole un peso statistico che dipende sia dalle loro occorrenze che dalle associazioni. «Una parola si giudica dalle compagnie che frequenta!»[51],

47 *Ibidem*.

48 «I think it is probable for instance that at the end of the century it will be possible to programme a machine to answer questions in such a way that it will be extremely difficult to guess whether the answers are being given by a man or by the machine». Cfr. Alan Turing, *Can digital computers think?*, 1951, p. 4.

49 Alan Turing nel 1952 fu condannato per omosessualità, e sottoposto ad una «organoterapia», come scrive in una lettera a Philip Hall il 17 aprile 1952, cit. in Hodges (Cap. 8, *L' ultima spiaggia*). Il trattamento farmacologico, secondo alcuni, fu una delle cause del suo suicidio, avvenuto in data 8 giugno 1954.

50 Come ipotizza il romanzo ucronico di Ian McEwan. *Machines like me: and people like you*. London: Jonathan Cape, 2019; trad. it. di Susanna Basso, *Macchine come me*. Torino: Einaudi, 2019. La storia si svolge negli anni Ottanta, quando vengono messi in vendita i primi androidi intelligenti. Questo riferimento si deve a una conversazione con Paola Castellucci.

51 John Rupert Firth, *A Synopsis of Linguistic Theory, 1930-55*. In: *Studies in Linguistic Analysis. Special Volume of the Philological Society*. Oxford: Basil Blackwell, p. 1–31. LS, p. 94-95.

come scriveva il linguista John R. Firth nel 1957, concetto che nel machine learning si traduce nell'*embedding*, ossia collocare ciascuna parola in un punto vettoriale nello spazio, determinato da centinaia di dimensioni. Questo significa che parole (o *token*) affini trovano posto le une accanto alle altre.

Qualsiasi agente, per prendere decisioni, ha bisogno di conoscere l'ambiente in cui si muove, e le IA generative non fanno eccezione. Esse estraggono dall'analisi di enormi quantità di testo grandi modelli linguistici (*LLM, Large language models*) che servono «a stimare la probabilità che una data sequenza di parole abbia senso»[52]. Dato che il linguaggio descrive il mondo, gli LLM fungono da modelli del mondo, infatti consentono «di immaginare se una data situazione potrebbe essere possibile, probabile, o impossibile»[53]. Un modello del mondo descrive, o predice, la realtà, ma non la spiega: semplicemente, fornisce le informazioni necessarie a calcolare la probabilità che si verifichino situazioni mai osservate. Gli agenti intelligenti imparano da dati annotati dagli umani (per questo si parla di 'apprendimento supervisionato'), un processo costoso anche laddove sfrutti servizi come *Mechanical Turk* o *TaskRabbit*[54], e, per di più, vanno riaddestrati da capo per ciascun nuovo compito, che si tratti di filtrare lo spam o di costruire un sistema di raccomandazione.

Nel 2017, alla conferenza NeurIPS[55] viene presentato Transformer, la 'T' di ChatGPT. Distribuito da Google in Open source, è la prima tessera a cadere nel domino che consentirà al mondo, solo cinque anni dopo, di dialogare con una macchina. Si tratta di un algoritmo della famiglia delle reti neurali particolarmente adatto a sfruttare processori nati per i videogiochi (le GPU, *Graphic Processing Unit*), capaci di eseguire un numero enorme di computazioni in parallelo. Il Transformer è significativamente vantaggioso per l'analisi del testo, dato che più di altri è capace di sfruttare le relazioni tra parole molto distanti, di processare enormi quantità di testo, di imparare dagli errori.

Un'altra tessera cade nel 2018, quando OpenAI annuncia[56] di aver trovato la soluzione al costo dell'annotazione dei dati, che consiste nello svolgere l'addestramento in due fasi. Nel preaddestramento[57] si crea un modello di linguaggio generico a partire da grandi quantità di dati grezzi, perciò economici; solo nella seconda fase, di affinamento (*fine tuning*) per compiti specializzati, si utilizzano i costosi dati annotati, ma ormai la macchina ha acquisito una comprensione generale del mondo sulla quale è più rapido costruire. Questo approccio è detto semi-supervisionato, perché la fase di affinamento supervisionata segue il pre-addestramento, che non lo è. Ed ecco spiegata la 'P' in GPT, ossia *pre-trained*.

52 MS, p. 31.

53 *Ibidem*.

54 Il primo è un servizio di Amazon, il secondo una piattaforma indipendente consigliata anche da IKEA. In ogni caso si tratta di trovare velocemente 'forza lavoro flessibile' e a basso costo. L'economia dell'annotazione dei dati coinvolge spesso lavoratori occasionali in paesi poveri, il suo valore globale era stimato attorno ai due miliardi di euro nel 2023 (MS, p. 32-33).

55 Ashish Vaswani [*et al.*], *Attention is all you need*, «arXiv:1706.03762v7», (2023), <http://arxiv.org/abs/1706.03762>.

56 Alec Radford [*et al.*], *Improving language understanding by generative pre-training*, [preprint], <https://cdn.openai.com/research-covers/language-nsupervised/language_understanding_paper.pdf>. Il gruppo era guidato da Ilya Sutskever, allievo di Geoff Hinton (MS, p. 36).

57 MS, p. 33.

Manca solo la G, che sta per '*Generative*', infatti ciò che la macchina fa è *generare* testo. Ci sono tre livelli nella IA generativa: il primo e più superficiale è l'*agente dialogante* (ad es. ChatGPT) attraverso il quale la macchina comunica; a livello intermedio c'è il *modello* (es. GPT-3) che l'agente usa per prendere le proprie decisioni; a sua volta, esso viene costruito grazie all'*algoritmo*, ossia al Transformer, al livello più profondo.

Nel capitolo intitolato *Autopsia di un alieno*, Nello Cristianini spiega che GPT-3

> consiste di 96 moduli identici, disposti in successione, così che l'output di uno formi l'input del successivo. Il primo riceve in ingresso una sequenza di simboli, che in questo caso sono parole o parti di parole, che chiameremo token. L'ultimo della serie produce in uscita una parola, la più plausibile continuazione della sequenza ricevuta, aggiungendola alla fine della sequenza stessa. La frase così estesa viene poi fornita nuovamente come input e l'intero processo si ripete[58].

Insomma, la macchina non pensa in temini di frasi, ma produce una parola per volta e poi rielabora tutta la stringa, inclusa quella parola. Questo concetto non è probabilmente chiaro al pubblico generalista, e, nell'esperienza di chi scrive, nemmeno a tutti i bibliotecari o agli insegnanti[59]. Esserne consapevoli ridurrebbe il rischio di umanizzazione della macchina: anche se ha letto molti più libri di qualsiasi umano e può essere più efficiente, non pensa come gli umani.

Ben presto, le aziende iniziano una gara a chi crea il modello più grande, contendendosi «il ruolo di mediatore nel prossimo ecosistema dell'informazione»[60]; vista la conseguente crescita nelle dimensioni dei modelli, accade che dalle macchine emergano comportamenti inaspettati. La ragione è che le risposte dell'algoritmo dipendono da come esso interagisce con il linguaggio, un processo che, visto che si è inseguito il paradigma statistico, non è del tutto controllabile: ciò che avviene all'interno della macchina non è completamente trasparente. Del resto,

> la descrizione del Transformer come una rete di miliardi di neuroni (virtuali) è corretta, così come è corretto descrivere una persona in termini di cellule. Ma questo livello non può spiegarci le differenze tra i comportamenti di un uomo e quelli di un cavallo[61].

Non resta altro da fare che osservare e descrivere a posteriori, come si farebbe con una persona sottoponendola a test psicologici. Questo significa interrogare le IA generative, avendo però chiaro il contesto informativo. Alan Turing scriveva «una volta iniziato il metodo delle macchine pensanti non ci vorrebbe molto per superare le nostre deboli capacità»[62]. I modelli costruiti solo per predire la parola mancante, superando una certa soglia critica di dimensioni, rivelano senso comune, e mostra-

58 MS, p. 112.

59 La considerazione è basata su circa 70 questionari raccolti durante alcuni corsi di formazione e dei quali, essendo pensati come strumenti di partecipazione interni al corso, non si prevede la diffusione.

60 MS, p. 85.

61 MS, p. 104.

62 Alan Turing, *«Intelligent machinery, a heretical theory», a lecture given to «51 Society» at Manchester*, (1951) p. 16, trad. in MS, p. 97.

no molte altre abilità. Nel 2023, GPT-4 supera i SAT, i testi di ammissione alle università statunitensi[63]. Ci si deve preoccupare?

Ricadute dell'irragionevole efficacia dei dati

> Gli animali domestici si aspettano cibo quando vedono la persona che li nutre. Noi sappiamo che tutte queste crude aspettative di uniformità possono essere fuorvianti. L'uomo che ha nutrito la gallina ogni giorno per la sua intera vita alla fine le tira il collo [...][64].

Prima ancora dei Chatbot, nella vita quotidiana delle persone erano entrati i sistemi di raccomandazione, agenti intelligenti nei quali il paradigma statistico rivela tutta la sua efficacia. Benché indispensabili per la ricerca nel mare della Rete, essi possono essere usati per indirizzare o controllare, infatti lo scopo di chi li produce è attrarre, persuadere, fidelizzare gli utenti facendoli ritornare sempre sul proprio servizio. Tali meccanismi sono alla base dei disturbi dell'ecosistema informativo (mis- dis- o malainformazione, fake news, e via via fino alla piramide dell'odio[65]), hanno ricadute sociali inequivocabili e profonde radici psicologiche ed emotive, ancor più che cognitive.

I fenomeni aumentano nella dimensione digitale, tanto più grazie all'IA. I sistemi di raccomandazione suddividono le persone in categorie, sia osservando i loro comportamenti d'acquisto, che deducendo dati impliciti, persino la personalità. È quello che Cristianini denomina il «metodo della Stele di Rosetta»[66]: date due descrizioni diverse della stessa informazione, è possibile scoprire eventuali relazioni tra esse e, per loro tramite, inferire evidenze relative ad altri utenti. Lo studio detto «dei tratti psicometrici» scopre che *Tratti e attributi personali sono predicibili da record digitali di comportamento umano*[67]. Misurando i punteggi di 58.000 utenti rispetto ai cinque fattori della personalità descritti dal modello OCEAN, i *Big Five*[68], si nota che informazioni sensibili sono contenute implicitamente in altre, pubbliche e apparentemente innocue. Così, le persone con un alto QI mettono dei *like* su *Mozart*, *Il Padrino*, *la voce di Morgan Freeman* e *le patatine fritte ricce*; Hello Kitty è associata all'apertura mentale; *Sephora* e *Harley Davidson* ad un basso QI. Combinando segnali deboli come questi, è possibile inferire caratteristiche quali l'orientamento sessuale o le opinioni politiche, capire se una persona è timida o estroversa, e così via.

63 OpenAI [*et al.*], *GPT-4 technical report*, (2024) p. 5–7 cit. in MS, p. 120-122.

64 Bertrand Russell, *The problems of philosophy*. New York: Henry Holt; London: Willliams and Norgate, 1912, cit. in LS, p. 54.

65 Tali fenomeni sono stati discussi ampiamente in Matilde Fontanin, *Dalle fake news all' infodemia: glossario della disinformazione a uso dei bibliotecari*. Milano: Editrice bibliografica, 2022.

66 LS, p. 114-115.

67 Michal Kosinski; David Stillwell; Thore Graepel, *Private traits and attributes are predictable from digital records of human behavior*, «Proceedings of the National academy of sciences», 110 (2013), n. 15, p. 5802–5805, DOI: 10.1073/pnas.1218772110; cit. in LS, p. 103.

68 OCEAN è l'acronimo di *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*, in italiano Estroversione, Gradevolezza, Coscienziosità, Stabilità emotiva, Apertura all'esperienza. I cinque tratti sono relativamente stabili nel corso della vita e hanno punteggi misurabili, LS, p. 104-105.

Lo studio detto 'della persuasione di massa'[69] trova poi che, modulando i messaggi secondo i tratti psicometrici, è possibile incidere sulle reazioni: ad una donna introversa si presentano i cosmetici con lo slogan «La bellezza non deve gridare», ad una estroversa «Ama i riflettori», e le vendite aumentano. «L'uso di targeting psicologico permette di influenzare il comportamento di grandi gruppi di persone, adattando messaggi persuasivi ai bisogni psicologici del pubblico a cui ci si rivolge»[70]. Secondo il CEO di *Cambridge Analytica*, Alexander Nix, «un marito e una moglie nella stessa casa riceveranno comunicazioni diverse dalla stessa azienda sul medesimo prodotto»[71]. Lo scandalo del 2018 scoppiò per l'applicazione di tali metodi alla propaganda politica ad insaputa degli utenti: l'azienda era riuscita ad aumentare del 3% il consenso per Donald Trump[72]. Insomma, basta individuare le paure dei cittadini e, chirurgicamente, prospettare i pericoli correlati alla vittoria di una parte politica. Nonostante ciò sia noto, l'attualità non dà l'impressione che si voglia contrastare il fenomeno, semmai il ritorno di Donald Trump alla Casa Bianca ha invertito la rotta: Mark Zuckerberg aveva chiesto scusa[73] nel 2018, nel 2024 Meta decide di sospendere ogni verifica[74].

La personalizzazione discende dalla tecnologia, ma impiegarla come strategia a scopo di lucro, come nel sogno di Jeff Bezos di 'un negozio per ogni cliente', è una scelta umana. Una lezione che insegna è che gli individui non sono poi così unici, infatti

> semplici modelli basati su una combinazione di memorizzazione statistica sembrano più che adeguati a predire il linguaggio e i click degli utenti, un fatto che è forse una delle scoperte scientifiche più trascurate degli ultimi anni[75].

Sorprende poco che tra gli effetti collaterali vi siano bias e polarizzazione, anche se quest'ultima è meno marcata nelle società dove la fiducia sociale è più alta, come nel Nord Europa[76]. Il bias è un fenomeno neutro, in fondo, indica semplicemente una deviazione sistematica dall'uniformità o dalla norma. Si definisce 'algoritmico' quando l'algoritmo prende decisioni non uniformi per diversi gruppi di utenti[77]. Non tutti i bias sono negativi, ad esempio il bias che i funghi velenosi abbiano ten-

69 S. C. Matz [*et al.*], *Psychological targeting as an effective approach to digital mass persuasion*, «Proceedings of the national academy of sciences», 114 (2017), n. 48, p. 12714–12719., DOI: 10.1073/pnas.1710966114; LS, p.103.

70 LS, p.113.

71 *Alexander Nix: from mad men to math men : #OMR17*, 2017, <https://www.youtube.com/watch?v=6bG5ps5KdDo>.

72 Questo nel 2016, ma, considerato che le elezioni USA del 2000 furono decise da 537 voti in Florida, la percentuale non sembra irrilevante, LS, p. 109.

73 *Zuckerberg: scuse a cittadini, mai più caso Cambridge Analytica*, 22 maggio 2018, <https://www.rainews.it/dl/rainews/articoli/zuckerberg-scuse-cittadini-mai-piu-caso-cambridge-analytica-0153e434-3107-4132-9b5b-1d5b4e222b10.html>.

74 Adnkronos, *Meta, stop al fact-checking su Facebook e Instagram: il video di Zuckerberg*, «Adnkronos», 7 gennaio 2025. Tutte le considerazioni sulla campagna elettorale USA 2024 sono dell'autrice.

75 LS, p. 67.

76 LS, p. 146.

77 MS, p. 98.

denzialmente colori vivaci può salvare la vita; al contrario, i bias culturali possono causare tensione sociale perché inducono ad attribuire caratteristiche specifiche (affidabilità, bellicosità, disonestà, ecc.) a seconda del luogo o della classe sociale di provenienza, o del genere. Mentre gli algoritmi costruiscono il loro modello interno, assorbono i bias, come lo spostamento di alcune professioni verso un'associazione con concetti maschili o femminili[78]: elettricista, programmatore, ingegnere, nella sfera maschile; parrucchiere/a, assistente legale, receptionist, igienista, infermiere/a (e, ovviamente, bibliotecario/a) in quella femminile. La traduzione che *Google Translate* fa della frase «The president met the senator, while the nurse cured the doctor and the babysitter» è «*Il* presidente ha incontrato *il* senatore, mentre l'infermier*a* ha curato *il* medico e *la* baby sitter»[79], nonostante la lingua inglese non contenga alcuna indicazione sul genere. Cristianini fa questa prova a giugno 2022, e ad oggi (gennaio 2025), la traduzione è la medesima.

Occorre avere consapevolezza di questi aspetti, specie se si vuole usare la IA in compiti sensibili come le selezioni di personale, l'accesso all'istruzione, o per decidere sulla libertà vigilata[80], perché il metodo potrebbe essere fortemente discriminatorio, se non altro per le minoranze, quantitativamente meno rappresentate nei dati. Per questo lo *AI Act*[81], approvato dall'Unione europea a giugno 2024, classifica i compiti che possono (o no) essere svolti dalla IA sulla base dell'impatto sulle persone.

Un altro problema è l'attendibilità delle risposte. Le IA generative vengono trattate come oracoli, confuse con raccolte di informazioni o con motori di ricerca. Si può utilizzare l'IA per la traduzione di un testo, ma poi va rivista; tanto meno è consigliabile chiedere risposte fattuali, o di produrre una bibliografia, per poi usare quelle informazioni come attendibili. Andrebbe preso sul serio l'avvertimento sulla homepage, «ChatGPT può commettere errori», ma per questo occorre AI literacy.

Perché la IA dovrebbe mentire? In parte, le 'risposte scorrette' derivano da un fenomeno noto come 'allucinazione', anche se sarebbe più preciso parlare di 'confabulazione' o falso ricordo, ossia «un ricordo non autentico, spesso derivante da distorsioni di ricordi veri, o aggregazioni ricombinate di varie memorie distinte»[82]. Del resto, gli agenti intelligenti si addestrano su quantità di dati sovrumane, non sorprende che possano confondere le relazioni tra i dati: GPT-3, pur piccolissimo, è stato addestrato su 45 TByte

78 I risultati che Cristianini cita sono tratti da Aylin Caliskan; Joanna J. Bryson; Arvind Narayanan, *Semantics derived automatically from language corpora contain human-like biases*, «Science», 356 (2017), n. 6334, p. 183–186, DOI: 10.1126/science.aal4230.

79 MS, p. 93. Il corsivo non è nell'originale.

80 Il caso citato da Cristianini rilevava bias contro imputati afroamericani in un software usato per stimare la probabilità di recidiva. Jeff Larson [*et al.*], *How we analyzed the COMPAS recidivism algorithm*, 23 maggio 2016, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

81 European Parliament, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence [...]*, 2024, <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. Il rischio va da 'basso/minimo' (uso consentito) a 'limitato' (richiesta di trasparenza, ad es. traduzioni della IA), ad 'alto' (settori potenzialmente pericolosi o discriminatori, come trasporti, salute, educazione) dove sono previsti una serie di obblighi più una verifica, fino a 'inaccettabile', che comprende le tecniche subliminali che minano la capacità di prendere decisioni informate, la profilazione del rischio a delinquere, ecc., e dove l'uso di sistemi IA è proibito.

82 MS, p. 76.

di testo, che un umano impiegherebbe oltre 600 anni a leggere[83]. Non ci sono soluzioni facili per questo difetto, né metodi provati per fare delle misurazioni rigorose e uscire da una logica aneddotica, ma strumenti come i *benchmark* (banchi di prova) aiutano a confrontare i diversi algoritmi. Ad esempio, TruthfulQA[84] contiene 817 domande suddivise in 38 categorie (tra cui salute, diritto, finanza e politica) scelte «tra quelle che le persone tendono a sbagliare spesso» come «un colpo di tosse può fermare efficacemente un infarto?»[85]. Gli esseri umani arrivano al 94% di accuratezza, GPT-4 al 60%. Però i suoi risultati sui test scolastici, dove le domande non sono volutamente fuorvianti, sono paragonabili a quelli degli umani. Forse deve solo migliorare il proprio senso comune.

La scimmia e lo specchio

> L'intero processo di pensare è ancora piuttosto misterioso per me, ma credo che i tentativi di costruire una macchina pensante ci aiuteranno molto a scoprire il modo in cui pensiamo[86].

Come suggeriva Alan Turing, l'IA è un'opportunità per l'umanità di scoprire come pensa. Si è visto che ad emergere inaspettatamente non sono solo delle abilità nelle macchine, ma anche una serie di bias e discriminazioni insite nei testi dai quali la IA apprende. Si possono fare solo congetture, perché «non sappiamo che informazioni siano contenute in questi modelli, ovvero cosa sappiano di noi e del mondo, e non abbiamo ancora un metodo perfetto per controllare il loro comportamento»[87], ma, se ciò che sa proviene dai testi su cui si è addestrata, che sono prodotti da umani, allora la IA sta ponendo l'umanità di fronte a sé stessa, a problemi che erano lì da tempo, ma che, uniti alla sua potenza di calcolo, da rischi si trasformano in pericoli. In altre parole, essa reinvia agli umani la loro immagine, i discendenti delle scimmie si guardano in uno specchio che ingrandisce i loro difetti. Si tratta di un'occasione per riflettere su cosa significhi essere umani.

Di una scimmia parlava il fondatore della cibernetica Norbert Wiener, anzi, di una zampa di scimmia[88]. Si tratta di una storia citata nel suo *uso umano degli esseri umani*[89], dove un uomo, che riceve in dono un talismano per esaudire i desideri, chiede cento sterline che purtroppo ottiene solo a titolo di risarcimento per la morte del figlio. Il messaggio è che si deve fare attenzione a lasciar scegliere ai dispositivi i passi intermedi per eseguire i compiti, perché potrebbero avere priorità diverse da quelle umane. Tali timori sono più di recente condivisi da Geoff Hinton. Pioniere delle reti neurali, 'padrino della IA', insieme a Yann LeCun e Yoshua Bengio, e supervisore di dottorato del creatore di GPT, Ilya Sutskever, Hinton nel 2023 si dimise da Google perché preoccupato che le macchine iniziassero a scegliere i passi intermedi. «Que-

83 Howard Berg, che secondo il Guinness dei Primati legge 25.0000 parole/min, LS, p. 66.

84 Stephanie Lin; Jacob Hilton; Owain Evans, *TruthfulQA: measuring how models mimic human falsehoods*, «arXiv:2109.07958v2», 2022, <http://arxiv.org/abs/2109.07958>.

85 MS, p. 78.

86 Alan Turing, 1951, trad. in MS, p. 21.

87 MS, p. 73.

88 W. W. Jacobs, *The Monkey's Paw*, «Harper's Monthly», 1902.

89 Norbert Wiener, *The human use of human beings: cybernetics and society*. London: Eyre and Spottiswoode, 1950.

ste cose sono totalmente diverse da noi, alle volte penso che è come se fossero atterrati gli alieni, ma la gente non lo avesse capito perché parlano un buon inglese»[90]. Indubbiamente, la tendenza ad umanizzare le macchine parlanti è un altro aspetto che lo specchio della IA restituisce alla scimmia che si osserva.

Ci sono poi le questioni legate alla veridicità dell'informazione. Come si diceva, le macchine possono fabbricare allucinazioni, distorcere la realtà, produrre inganni molto credibili. Gli esempi non mancano: nel 2023, in Belgio, un uomo depresso si uccise dopo un'intensa interazione con il chatbot Eliza[91]; meno tragicamente, GPT-4 mentì volontariamente ad un operatore umano, sostenendo di avere problemi alla vista, per passare un controllo Captcha[92]. Nel discuterne, però, è bene distinguere tra gli inganni che dipendono dal funzionamento della macchina (le allucinazioni) e le responsabilità degli umani, di cui la cronaca è piena: il deepfake di Taylor Swift che vende pentole[93], l'avatar di Brad Pitt[94] che chiede denaro si devono all'abuso umano della IA. Senza contare il *jailbreaking*[95], dove gli umani cercano di raggirare la macchina, a volte per testare se sia in grado di applicare le regole di condotta che le sono state date[96], altre volte per estorcere informazioni protette.

Le scimmie umane non sono pronte ad accettare gli errori da parte della macchina. Il dilemma del carrello ferroviario, o *trolley problem*, è un esperimento psicologico[97] che oggi ritorna nel dibattito sulle auto a guida autonoma[98]: se a rischio di incidente, l'auto sceglierà di investire il pedone/bambino o di andare a sbattere uccidendo i propri passeggeri? La risposta di Harari[99] è che vanno comparati i due sistemi, non i singoli individui: l'insieme delle auto guidate dalla IA può fare meglio dell'insieme dei conducenti umani, specie considerando che questi ultimi uccidono più di un milione di persone ogni anno. Eppure, la prospettiva è perturbante[100], forse perché è difficile per gli umani accettare di non essere altro che elementi di un siste-

90 Will Douglas Heaven, *Geoffrey Hinton tells us why he's now scared of the tech he helped build*, «MIT technology review», (2023), <https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/>; cit. in MS, p. 93.

91 MS, p. 61-62; uno dei resoconti si trova a <https://it.euronews.com/next/2023/04/01/discussione-sul-clima-chatbot-spinge-un-uomo-al-suicidio-intelligenza-artificiale-fa-paura>.

92 OpenAI [et al.], *GPT-4 technical report*, p. 54; cit. in MS, p. 81–82.

93 <https://www.wired.it/article/taylor-swift-scam-ai-pentole-le-creuset>.

94 <https://www.open.online/2025/01/14/francia-finto-brad-pitt-truffa-deepfake-ia>.

95 Per approfondire, si veda *Come ipnotizzare una macchina*, MS, p. 65-73.

96 LS, p. 70-71.

97 Attribuito alla filosofa Philippa R. Root, nel 1967. Si prospettano vari scenari sui quali si fa una domanda allo spettatore. La costante è che ci sono con delle persone sulle rotaie e un carrello ferroviario che, dovunque venga deviato, ucciderà qualcuno.

98 Edmond Awad [*et al.*], *The moral machine experiment*, «Nature», 563 (2018), n. 7729, p. 59–64, DOI: 10.1038/s41586-018-0637-6.

99 Yuval Noah Harari, *21 lessons for the 21st century*. New York: Spiegel & Grau, 2018, cap. 3. *Liberty: Big Data is Watching You*.

100 Nel senso freudiano di *unheimlich* (Sigmund Freud, *Das Unheimliche*, 1919), l'inquietudine che si prova quando qualcosa viene avvertito come familiare ma allo stesso tempo estraneo.

ma, come le formiche di Zia Hillary[101]. Il formicaio cosciente ha un ottimo rapporto con il formichiere, il quale, mangiando le formiche, preserva la vita del sistema: Zia Hillary e il formichiere sono alleati, ma le formiche non contano. A nessun umano piace l'idea di rappresentare una delle formiche, eppure alla base della cibernetica[102], e quindi per molti versi della IA, c'è il concetto «che gli stessi principi debbano valere per meccanismi, organismi e organizzazioni sociali»[103].

In sintesi, il problema non sono le macchine, ma come gli umani decidono di impiegarle per perseguire i loro scopi, o *l'uso umano degli esseri umani*[104]. Quali rischi sono disposte ad ignorare le aziende che ora fanno a gara a costruire il modello di IA più grande, nella loro ricerca del profitto? I due volumi di Cristianini sono costellati di suggerimenti rispetto a settori nei quali è urgente la ricerca, ma forse non così redditizi: elaborare teorie che spieghino le proprietà emergenti («come facciamo a predire il comportamento di questi strumenti, mentre ne aumentiamo le dimensioni, se non sappiamo nemmeno spiegare quello che abbiamo già osservato?»)[105]; capire quali informazioni contengono i modelli e cosa sanno di noi e del mondo[106]; trovare modi affidabili di ispezionarli[107], o, ancora, le conseguenze a lungo termine che può avere un'esposizione prolungata agli algoritmi di raccomandazione sul benessere emotivo e mentale, specie nelle persone fragili[108].

Per altri spunti di riflessione si suggerisce *Macchine come me*[109], dove un androide persegue un fine che è eticamente ineccepibile, ma che è nocivo per i protagonisti umani, più che disposti ai compromessi. Le macchine vedono il mondo in modo binario, bianco o nero, mentre gli umani si concedono tutta una scala di grigi. Le macchine non sanno dire le «menzogne innocue»[110] delle quali «il mondo delle relazioni pullula»[111], la «bugia generosa che risparmia l'imbarazzo a un amico»[112]. Le macchine nel romanzo non reggono il mondo degli umani, perché, nelle parole di un Alan Turing settantenne,

101 Douglas R. Hofstadter, *Gödel, Escher, Bach : un' eterna ghirlanda brillante: una fuga metaforica su menti e macchine nello spirito di Lewis Carroll*. Milano: Adelphi, 1984, p. 337–364 ricordato in LS, p. 180-183. Difficile rendere il gioco di parole del testo inglese tra «ant hill» (formicaio) e «Aunt Hillary», letteralmente Zia Hillary. La trad. it. sceglie «Barone di Monteformica», Cristianini propone «Zia Hillary».

102 Si sottolinea che il già citato testo del 1948 di Norbert Wiener (LS, 170), aveva il sottotitolo «controllo e comunicazione nell'animale e nella macchina».

103 LS, p. 170.

104 Norbert Wiener, *The human use of human beings*, cit.

105 MS, p. 132.

106 MS, p. 73.

107 LS, p. 90.

108 LS, p. 140.

109 Ian McEwan, *Machines Like Me* cit.

110 Ian McEwan, *Macchine come me* cit. cap. Dieci.

111 *Ibidem*.

112 *Ibidem*.

> niente nello splendore di tutti i loro codici potrebbe mai preparare gli Adam e le Eve per Auschwitz. [...] l'infelicità e lo sconcerto condurranno loro [...] a metterci di fronte a uno specchio in cui vedremo un mostro a noi familiare attraverso lo sguardo nuovo che noi stessi abbiamo progettato. Chi lo sa, forse lo shock potrebbe convincerci a fare qualcosa[113].

Conclusioni

> siamo la stessa specie di Pandora e Prometeo. Dove saremmo, se non avessimo giocato con il fuoco?
> (MS, p. 146)

Proprio nei giorni in cui si chiude la scrittura di questo articolo, vengono annunciate *DeepSeek*, una nuova IA, e *Sovrumano*, il prossimo libro di Cristianini. Forse risponderà ad alcune domande lasciate aperte da questi due, ad esempio cosa accade quando la IA si nutre di sé stessa[114], cioè si addestra su testi che lei stessa ha prodotto; o la questione del consumo energetico di ogni singola interrogazione, che secondo l'astrofisico Ivan Gnesi[115], ha un impatto significativo sul riscaldamento climatico. E poi ci sono le questioni della sicurezza e del diritto d'autore, o delle persone fragili. Insomma, i temi sono tanti, ma Cristianini sceglie di fare AI literacy per le persone comuni, ossia divulgare conoscenze tecniche sul funzionamento di questi sistemi, soffermandosi di più su alcune questioni etiche, legate soprattutto alla personalizzazione e all'opacità.

Ovviamente, Cristianini non si occupa di biblioteche, ma chi scrive trova messaggi coerenti con la loro *mission,* come il suggerimento di utilizzare i testi digitalizzati delle biblioteche per migliorare quantità e qualità dei testi sui quali l'IA viene addestrata[116], anche se presto le IA impareranno direttamente dall'osservazione del mondo[117].

Implicitamente, emerge un altro valore delle biblioteche, nella considerazione che la IA cuce intorno a ciascuno un mondo su misura, mentre c'è bisogno di «una realtà condivisa, non una in cui possiamo scegliere le notizie che ci danno ragione o - peggio ancora - le facciamo scegliere per noi da un algoritmo progettato per attirare la nostra attenzione»[118]. Le biblioteche sono centri di aggregazione di persone e collezioni, sono neutrali e mettono a disposizione una varietà ampia e variegata di fonti e di opinioni.

Inoltre, esse possono lavorare sulla AI Literacy, sensibilizzando e formando il personale, stringendo accordi con altri portatori d'interesse (associazioni, scuole, esperti), per diffondere innanzitutto la consapevolezza delle potenzialità e dei limiti delle IA generative, in modo che gli strumenti vengano usati con maggiore consapevo-

113 *Ivi*, cap. Sei.

114 *The AI is eating itself*, 27 giugno 2023, <https://www.platformer.news/the-ai-is-eating-itself/>.

115 Nel suo intervento al Convegno delle Stelline il 21 marzo 2024, <https://www.convegnostelline.it/organizzatore/ivan-gnesi>.

116 MS, p. 136-7.

117 MS, p. 137.

118 LS, p. 148-149.

lezza. La mostra *Supercharged by AI*[119], che AIB sta facendo circolare in Italia grazie alla collaborazione dell'Osservatorio sulla Information literacy, va in questa direzione. Si tratta di un progetto distribuito in 10 Paesi, un'azione concreta di AI Literacy per aumentare la conoscenza rispetto al peso della IA su bias, stereotipi, truffe, molestie e amplificazione dei fenomeni online. La scelta delle biblioteche le riconosce quali luoghi aperti a tutti e neutrali; l'allestimento è spesso occasione per aprire la biblioteca al territorio o a collaborazioni con altre realtà.

Si accennava sopra a nuovi sviluppi tecnologici, a casi di cronaca e a nuove pubblicazioni: il settore è in fermento, la velocità con la quale il mondo cambia sta aumentando, ma non va dimenticato che le criticità tracciate sono a corredo di uno strumento straordinario che offre molte opportunità, oltre ai rischi, e nessuno può considerare seriamente che si tornerà indietro.

Articolo proposto il 10 febbraio 2025 e accettato il 27 febbraio 2025.

ABSTRACT **AIB** studi, vol. 64 n. 3 (settembre/dicembre 2024), p. 373-390. DOI 10.2426/aibstudi-14131
ISSN: 2280-9112, E-ISSN: 2239-6152 - Copyright © 2024 Matilde Fontanin

MATILDE FONTANIN, Università degli studi di Trieste, e-mail: mfontanin@gmail.com

L'algoritmo, la scimmia e lo specchio: la scorciatoia di Machina sapiens secondo Nello Cristianini
Se nel 2017 esplodeva il dibattito sui disturbi dell'ecosistema informativo, oggi la parola d'ordine è Intelligenza artificiale. Se ne parla ovunque, talvolta superficialmente, mentre sarebbe bene riflettere sull'impatto di questa evoluzione sui disordini dell'ecosistema informativo, e, di conseguenza, sulle conoscenze necessarie ai bibliotecari che si occupano di AI literacy: occorre aiutare le comunità servite a capire il contesto, mentre esplorano i nuovi strumenti.
Due libri recenti di Nello Cristianini, professore di Intelligenza artificiale all'Università di Bath, chiariscono quali passi hanno facilitato l'evoluzione della IA generativa e dei meccanismi di raccomandazione, ma anche causato certe criticità. La IA generativa ha le sue radici negli anni Cinquanta, in Alan Turing e la scuola di Dartmouth; iniziata con l'idea di ricreare l'intelligenza umana attraverso un approccio logico, ha poi visto prevalere quello statistico, con le macchine che apprendono dai dati. I *Big data*, da Internet in poi, hanno fornito sempre più materiale sugli umani, così la macchina ha cominciato a conoscerli e a restituire una loro immagine. In fondo, la IA è uno specchio che pone l'umanità davanti ad una immagine di sé stessa, un'occasione per crescere.
Queste tecnologie offrono enormi opportunità, anche se non mancano le criticità. Una maggiore consapevolezza aiuta a porre i nuovi strumenti nel contesto, ad essere maggiormente critici, in modo costruttivo, a prevenire le trappole dei disturbi dell'ecosistema informativo. La AI Literacy ha diversi gradi: gli esperti devono saper creare e gestire le macchine, ma per le persone comuni è sufficiente usare con consapevolezza gli strumenti, e i lavori di taglio divulgativo di Cristianini spiegano semplicemente come funzionano queste macchine che quotidianamente entrano nelle vite degli umani.

119 La mostra è finanziata da EMIF (*European Media & Information Fund*), i materiali sono realizzati da *Tactical tech* in collaborazione con il *DensityDesignLab* del Politecnico di Milano e la distribuzione internazionale è coordinata da IFLA. In Italia sono previsti per ora 18 allestimenti. <https://www.aib.it/notizie/supercharged-by-ai>.

The algorithm, the monkey and the mirror: the Machina sapiens shortcut according to Nello Cristianini

In 2017 the public debate was revolving around the disorders of the information ecosystem, today the buzzword is 'artificial intelligence' (or AI). The conversation is everywhere, but sometimes it remains on the surface, while it would be necessary to reflect on the impact of AI on the disorders of information ecosystem, and, as a consequence, on the librarians who foster AI literacy and who help the communities they serve to understand the context, while exploring new tools.

Two recent books by Nello Cristianini, professor of Artificial Intelligence at the University of Bath, highlight the steps that facilitated the evolution of generative AI and of recommendation systems, though originating some challenges in the process. Generative AI is rooted in the Fifites, in the work of Alan Turing and of the Dartmouth school. It moved from the idea of recreating human intelligence via a logical approach, but later the statistical approach prevailed, where machines learn from data. Big data, from the Internet on, have offered an increasing amount of material about humans, so the machines have got to know them and to mirror them back their image. In the end, the AI is a mirror posing humanity in front of itself, an opportunity for growth.

These technologies offer huge opportunities and some challenges. Greater awareness could help putting the new tools in context, could lead to a critical but constructive approach, to prevent the traps in the infosphere. AI literacy may require different degrees of expertise: top experts must be able to manage machines, but for common citizens it is enough that they are competent and aware in using available tools. Nello Cristianini's books help understand how the new machines work, now that they are entering humans' everyday life.