



UNIVERSITÀ DI PAVIA

DIPARTIMENTI DI GIURISPRUDENZA, INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE, SCIENZE
ECONOMICHE E AZIENDALI, SCIENZE POLITICHE E SOCIALI, STUDI UMANISTICI

CORSO DI LAUREA INTERDIPARTIMENTALE IN

COMUNICAZIONE DIGITALE

LA SCIENZA DEI DATI E LA VALIDAZIONE DELL'INFORMAZIONE

NELL'ERA DEL CAOS INFORMATIVO

Relatore:

Chiar.mo Prof. Paolo Costa

Correlatore:

Chiar.mo Prof. Flavio Antonio Ceravolo

Tesi di laurea di

Lorenzo Tiari n° 542577

ANNO ACCADEMICO 2024/2025

Abstract

La tesi magistrale si inserisce nell'ambito della *Data Science* come paradigma epistemologico emergente in un'epoca di caos informativo. L'obiettivo principale è costruire un ponte teorico tra le scienze informatiche e quelle naturali, al fine di comprendere se la *Data Science* rappresenti solamente un semplice insieme di tecniche computazionali o una nuova teoria scientifica sui dati. Si è ripreso un modo di ragionare comune ai grandi scienziati del passato mettendo il dubbio come elemento centrale durante la trattazione e cercando un costante intreccio tra scienza e filosofia, rimanendo sempre in un quadro metodologico rigoroso. Partendo dalle teorie sull'informazione formalizzate da Claude Edward Shannon e Luciano Floridi, si affrontano i fondamenti teorici della disciplina definendo i concetti di dato, informazione, osservazione e metrica di validazione. Si è tentato di sviluppare una minima base logica sottostante a questa disciplina, desumendola dall'intelligenza artificiale, neuroscienze, neuroetologia e qualche rimando alla botanica. Si è dedicata molta attenzione al concetto di metrica di validazione tramite le idee del neuroscienziato Daniel Bor, secondo cui la "rilevanza" di un'informazione dipende dal contesto in cui viene elaborata riprendendo la teoria dell'evoluzione di Darwin.

Operando una simbolica traduzione linguistica, la tesi sostiene che la Scienza dei Dati non si basa soltanto sull'elaborare molti dati o calcolare probabilità, ma nel costruire metriche di validazione capaci di costruire "modelli del mondo" più rilevanti. I risultati mostrano che nell'impianto teorico attuale esistono definizioni vaghe e non ben definite, nonostante il mondo inorganico e quello organico presentino alla base gli stessi limiti universali nel costruire informazioni. In seguito ai recenti tentativi di fusione tra dispositivi meccanici e materiali organici, sarà necessario nei prossimi decenni porre molta attenzione a questa scienza e delineare un nuovo impianto teorico che sia coerente con l'attuale progresso scientifico.

The master's thesis is situated within the field of Data Science as an emerging epistemological paradigm in an age of informational chaos. Its main objective is to build a theoretical bridge between computer science and the natural sciences, in order to understand whether Data Science represents merely a set of computational techniques or a new scientific theory about data.

The thesis adopts a mode of reasoning common to the great scientists of the past, placing doubt at the center of the discussion and seeking a constant interplay between science and philosophy, while always remaining within a rigorous methodological framework. Starting from the theories of information formalized by Claude Shannon and Luciano Floridi, the theoretical foundations of the discipline are addressed by defining the concepts of data, information, observation, and validation metric. An attempt was made to develop a minimal logical foundation underlying this discipline, drawing from artificial intelligence, neuroscience, neuroethology, and with some references to botany. Particular attention was devoted to the concept of validation metrics through the ideas of neuroscientist Daniel Bor, according to whom the "relevance" of information depends on the context in which it is processed, drawing on Darwin's theory of evolution.

Through a symbolic linguistic translation, the thesis argues that Data Science is not based solely on processing large amounts of data or calculating probabilities, but on constructing validation metrics capable of building more relevant "models of the world". The results show that within the current theoretical framework there are vague and poorly defined concepts, despite the fact that both the inorganic and organic worlds share the same universal limits in constructing information. In light of recent attempts to merge mechanical devices with organic materials, it will be necessary in the coming decades to devote significant attention to this science and to outline a new theoretical framework consistent with current scientific progress.

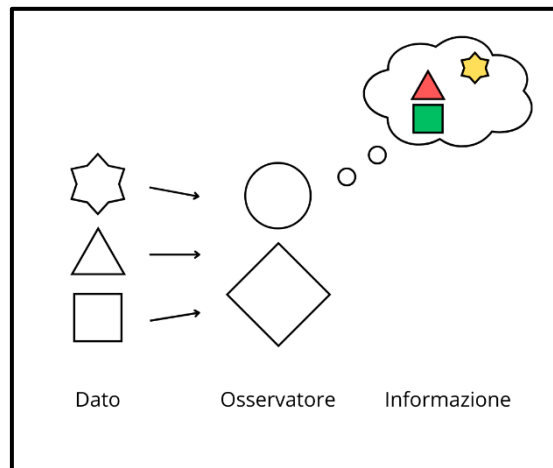
Indice

Introduzione	
Cap. 1 Il caos informativo nell'era dei dati	
1.1 L'evoluzione dell'informazione dal DNA fino all'intelligenza artificiale.....	
1.2 La sopravvivenza del più validato.....	
1.3 Data Science o Data Sapiens: un nuovo modo di conoscere?	
Cap. 2 Come costruire un modello del mondo	
2.1 Analisi di un atto di osservazione.....	
2.2 L'inversione dei valori nei dati	
2.3 La validazione dell'informazione.....	
Cap. 3 Prevedere il passato per capire il futuro	
3.1 Il bilanciamento tra comprensione e approssimazione.....	
3.2 Risorse fondamentali per gestire l'incertezza.....	
3.3 Allucinazioni e inspiegabilità come algoritmi naturali.....	
Cap. 4 Modellare il linguaggio scientifico	
4.1 Database e intelligenza artificiale applicati nella scienza.....	
4.2 Struttura dell'esperimento e metriche utilizzate.....	
4.3 Analisi del modello Undermind.....	
Conclusione	
Bibliografia	
Sitografia	
Ringraziamenti	

Introduzione

In un'epoca in cui tutto sembra essere sempre più mediato da ciò che può essere misurato, registrato e calcolato interrogarsi sul significato della realtà diventa un'esigenza non solo filosofica, ma anche scientifica. Ciò che non è standardizzabile rischia di apparire irrilevante mentre ciò che è computabile tende ad assumere lo statuto di verità. La computazione non si limita così a descrivere il mondo, ma contribuisce a definirne i confini influenzando il modo in cui osserviamo, interpretiamo e attribuiamo senso alle cose. In questo scenario, il nostro modello mondo non è mai neutro e diventa necessario lo studio della teoria che ne sostiene l'uso e il significato.

Figura 1-Rappresentazione concettuale dei termini utilizzati nella tesi.



Autore: Lorenzo Tiari

Di conseguenza, verrà analizzata la *Data Science* partendo dal significato originario del termine introdotto da Peter Naur nel 1974, che lo impiegò come alternativa alla *Computer Science* per indicare in maniera generica lo studio dei metodi di elaborazione dei dati tramite le macchine.¹ Da quei decenni di grande fermento tecnologico e dalla paura crescente dell'automazione in grado di sostituire il lavoro fisico si è arrivati all'intelligenza artificiale ridefinendo anche il rapporto tra intelligenza e vita stessa. Nello stesso momento storico si è avuto una fioritura delle scienze naturali come le neuroscienze, neuroetologia e "neurobiologia" vegetale in grado di mostrarci per la prima volta il funzionamento della natura analizzando aspetti come la creatività, la cognizione e la logica con cui operiamo ogni giorno. Si è scoperto che il tentativo di ridurre meccanicamente

¹ P. Naur, *Concise Survey of Computer Methods*, Petrocelli Books, 1974

la natura ad una serie di operazioni matematiche risulti molto più complesso del previsto, costringendo gli scienziati a adottare nuove strategie come la nascente neuroinformatica che si avvale direttamente di composti biologici prelevati dalla natura. Un processo naturale, una sorta di simbiosi avvenuta molte volte nel corso dell'evoluzione dalle prime cellule fino agli organismi pluricellulari. Telefoni personalizzabili con il tocco di un dito, la costruzione di complicati codici di programmazione tramite la voce fino ad arrivare ai dispositivi che sfruttano direttamente l'interazione con il cervello. Il naturale tentativo della tesi è operare una traduzione linguistica del termine *Data Science* tentando di costruire un ponte tra le discipline informatiche e quelle naturali in merito allo studio dei dati. Durante la trattazione si è notato che la Scienza dei Dati è sottoposta alle solite pressioni evolutive. I risultati prodotti da algoritmi molto complessi danno vita a fenomeni tipici dell'intelligenza come errori, inspiegabilità e allucinazioni. Si è osservato come un robot, nelle sue fasi iniziali, debba compiere i primi passi in modo analogo a un bambino condividendo un destino universale: l'interazione con la realtà presuppone sempre l'adozione di una logica di base, definita comunemente come morale. Allo stesso modo, gli attuali modelli generativi di intelligenza artificiale si ritrovano in una scomoda condizione simile a quella di uno studente, ovvero essere costantemente sottoposti a continue metriche di validazione che ne determinano le capacità.

Nel primo capitolo è stato ripercorso lo sviluppo storico ed economico di discipline quali la matematica, la statistica, la logica e l'informatica, considerate come pilastri fondamentali di questa scienza. L'analisi ha messo in luce come queste discipline non si siano evolute in modo isolato, ma in stretta relazione con i contesti storici, tecnologici e produttivi rispondendo a esigenze concrete di misurazione, previsione e controllo della realtà. In questo quadro, è stata delineata l'importanza crescente dell'analisi dei dati nel corso del tempo evidenziando come la capacità di raccogliere, organizzare e interpretare informazioni rappresenti una costante evolutiva che attraversa l'intera storia della vita organica. A partire dai primi organismi viventi in grado di reagire a segnali ambientali minimi fino ad arrivare ai modelli artificiali, è emerso come il trattamento dell'informazione costituisca un elemento centrale per l'adattamento e la sopravvivenza. Successivamente, il capitolo ha approfondito l'evoluzione della specie umana alla luce di un contesto storico caratterizzato da un'abbondanza senza precedenti di informazioni e tecnologie. Si è introdotta la suggestiva ipotesi dei cosiddetti *Data Sapiens*, intesi come una possibile traiettoria evolutiva in cui le capacità cognitive dell'essere umano risultano sempre più mediate dai dati e dai modelli computazionali. Infine, è stato evidenziato come questa evoluzione non coincida necessariamente con il concetto tradizionale di progresso. L'aumento della disponibilità informativa e della precisione tecnologica solleva nuove problematiche mettendo in

discussione l'idea che l'accrescimento quantitativo della conoscenza conduca automaticamente a un miglioramento della condizione umana.

Nel secondo capitolo si sono discussi alcuni fondamenti teorici di questa scienza: il concetto di osservazione, dato, informazione e metrica di validazione. Si è analizzata la Scienza dei Dati tramite una prospettiva scientifica e filosofica, cercando di riprendere il modo di ragionare tipico dei grandi scienziati come il pensatore cinese Shen Kuo, l'illuminista americano Benjamin Franklin e il tedesco Alexander Von Humboldt. La scienza mira alla formulazione di teorie sul mondo naturale attraverso l'analisi quantitativa dei fenomeni e l'applicazione del metodo sperimentale. La filosofia contribuisce allo sviluppo dei concetti e della logica necessaria per un'analisi qualitativa della realtà. Essa non funge solo da supporto, ma ne analizza i presupposti interrogandosi sul significato e sulla coerenza delle teorie formulate. L'indagine ha preso in considerazione le teorie sull'informazione proposte da scienziati quali: Daniel Bor, Giorgio Vallortigara, Donald Hoffman e Claude Shannon. La scelta risponde all'esigenza di adottare un approccio multidisciplinare alla materia al fine di valutare eventualmente possibili influenze teoriche alla luce dei più recenti progressi scientifici.² Inoltre, l'analisi ha esteso il pensiero di alcuni filosofi, dai più antichi come Lao Tzu, Immanuel Kant e Friedrich Nietzsche fino ai più contemporanei, in particolare Antoinette Rouvroy e Luciano Floridi, rispettivamente in merito ai concetti di *Big Data* e informazione.

Le definizioni adottate nella tesi sono riprese comunemente dalla semiotica utilizzata in informatica e nel calcolo computazionale, quindi un modello di realtà codificato e codificabile. Alla luce delle capacità scientifiche attuali non esistono definizioni esaustive e universalmente accettate poiché ciascun termine riflette prospettive teoriche differenti, quelle prese in esame le ho ritenute più comprensibili per chi legge.

Dato → Etimologia latina di *datus* (dare, ciò che è dato). Il dato è una rappresentazione formalizzata di un fenomeno osservabile. Può trattarsi di un valore numerico, un fotone e molto altro. In questo senso, è l'unità minima potenzialmente significativa, ma il suo significato dipende dall'osservatore che lo interpreta.³

Informazione → Etimologia latina di *informatio* (dare forma, rappresentazione). L'informazione nasce dall'organizzazione e l'interpretazione dei dati. Essa rappresenta il passaggio da una rilevazione a un contenuto dotato di senso. Per esempio, una serie di transazioni online o un colore.⁴

² La *Data Science* verrà analizzata dal punto di vista dell'intelligenza artificiale, neuroetologia, botanica e neuroscienze.

³ [https://www.treccani.it/enciclopedia/dato_\(Dizionario-delle-Scienze-Fisiche\)](https://www.treccani.it/enciclopedia/dato_(Dizionario-delle-Scienze-Fisiche)) (attivo)

⁴ <https://www.treccani.it/vocabolario/informazione/> (attivo)

Osservatore → Etimologia latina di *observator* (colui che osserva). In questa tesi per osservatore s'intende un'entità in grado di ricevere, selezionare e interpretare dati producendo informazioni complesse. Le caratteristiche minime che lo definiscono sono la limitatezza delle informazioni accessibili, l'esistenza di un sistema di conservazione di informazioni, la presenza di scambi energetici con l'ambiente e una struttura materiale di base che ne renda possibile il funzionamento. Ad esempio, un sistema computazionale o un apparato sensoriale come l'occhio. Nell'ambito dell'analisi dei dati rende possibile la trasformazione dal dato ad una informazione.⁵

Il terzo capitolo è stato dedicato all'analisi dei modelli predittivi applicati all'ambito sociale. In particolare, viene approfondita la differenza tra la previsione effettuata da un modello computazionale e quella operata dal sistema nervoso umano. A tal fine, si sono presi come riferimento alcuni testi classici della filosofia antica considerati quasi come “*token* semantici”, per mostrare come determinati pensatori, grazie all'esperienza e all'osservazione sistematica della realtà, fossero in grado di anticipare con una certa regolarità il verificarsi di specifici fenomeni sociali e comportamentali. L'analisi ha messo in relazione queste riflessioni filosofiche con i contributi più recenti provenienti dalla letteratura scientifica contemporanea, con l'obiettivo di evidenziare punti di continuità e convergenza tra approcci antichi e moderni alla previsione dei fenomeni complessi. Successivamente, il capitolo si è concentrato su un caso di studio concreto basato sull'analisi di un dataset fornito dall'ISTAT. I dati sono stati elaborati tramite il software di analisi KNIME, utilizzando l'algoritmo di *clustering* K-Means, al fine di individuare schemi latenti all'interno del campione osservato. Si è condotta un'analisi delle abitudini di lettura dei residenti in Italia con l'intento di formulare alcune predizioni sui comportamenti sociali emergenti.

Il quarto capitolo ha analizzato la piattaforma Undermind, un modello progettato come strumento di supporto alla ricerca scientifica. Undermind si configura come un assistente avanzato per gli scienziati in grado di facilitare l'esplorazione della letteratura, l'individuazione di connessioni concettuali tra studi differenti e la sintesi di grandi quantità di informazioni provenienti da fonti eterogenee. In questa sezione è stata analizzata la tipologia di database e di archivi scientifici da cui il modello attinge, soffermandosi sul ruolo centrale delle grandi collezioni di articoli accademici. L'analisi ha consentito di introdurre una riflessione critica sui principali problemi che caratterizzano la letteratura scientifica contemporanea come il plagio, la duplicazione dei risultati, la frammentazione delle conoscenze e la crescente difficoltà per i ricercatori di mantenere una visione d'insieme aggiornata del proprio ambito di studio. A partire da queste criticità, il capitolo ha messo in evidenza come l'affidarsi a un modello artificiale non sia concettualmente diverso

⁵ <https://www.treccani.it/vocabolario/osservatore/> (attivo)

dall'affidarsi a un altro collaboratore umano, con la differenza che il modello opera su scala molto più ampia e con tempi notevolmente ridotti. Successivamente, è stato presentato un semplice esperimento di tipo qualitativo finalizzato all'osservazione del comportamento del modello in un contesto di ricerca simulato. L'analisi si è concentrata sulle caratteristiche principali come la capacità di individuare relazioni non immediatamente evidenti tra studi, la chiarezza nella sintesi delle informazioni e la trasparenza delle fonti utilizzate. Infine, sono stati discussi i principali vantaggi e svantaggi del modello mettendo in luce il suo potenziale come strumento di accelerazione del progresso scientifico e, dall'altro, i rischi legati a un utilizzo acritico quali l'eccessiva delega cognitiva o la possibile amplificazione di *bias* presenti nei dati.

Nella conclusione sono stati ripercorsi in modo sintetico i principali temi teorici affrontati nel corso della tesi. Questa sintesi ha consentito di richiamare i risultati metodologici emersi nei diversi capitoli evidenziando il contributo complessivo del lavoro svolto. Successivamente, l'attenzione è stata rivolta ai limiti e alle problematiche future dell'approccio scientifico contemporaneo, in particolare di quelle discipline che raggiungendo livelli di precisione sempre più elevati, si trovano a confrontarsi con questioni tradizionalmente di natura filosofica sui nostri limiti conoscitivi. Si è evidenziato come l'accelerazione attuale non elimini la necessità di un confronto costante con le domande fondamentali sulla realtà e su un possibile futuro post-scienza nei prossimi millenni.

Capitolo 1

Il caos informativo nell'era dei dati

Negli ultimi decenni a livello mondiale si è assistito ad una esplosione di dati grazie al miglioramento costante delle nuove tecnologie e della cooperazione nella ricerca scientifica tra le nazioni. Prendendo il caso specifico italiano le opere librarie (compresi e-book) hanno raggiunto la cifra di 102.987 titoli nel 2022, una media di 282 titoli diffusi ogni giorno dell'anno.⁶ Nel mondo negli ultimi vent'anni si presume siano stati pubblicati almeno 64 milioni di articoli scientifici, di cui 5 milioni solo nel 2022, una media di 13.699 per ogni singolo giorno.⁷ Mai nella storia evolutiva della nostra specie siamo stati di fronte ad una tale quantità di dati. Dati che dobbiamo quotidianamente essere in grado di leggere, controllare e validare. Le fonti di aleatorietà sono innumerevoli e, quando si considerano nella realtà quotidiana, la loro analisi diviene troppo complessa per essere calcolata con precisione. I modelli di machine learning scartano cartelle mediche incomplete per non peggiorare un algoritmo, le persone potrebbero farlo con una notizia di un sito sconosciuto sull'andamento del mercato finanziario per non peggiorare le loro finanze. Si potrebbe controllare l'autore, il tipo di fonte, se è scritta con toni sensazionalistici, se è incoerente al suo interno, se sono presenti dei *bias* cognitivi, ma non basterebbe per trovare un sistema univoco nel validare ciò che vediamo davanti a noi tutti i giorni. Viviamo nell'era dei dati: un'epoca in cui, con questo termine, si intendono tutte le realtà che ci circondano e che in qualche modo possono essere misurate e calcolate. Se poi tali misurazioni vengono automatizzate, possono trasformarsi in informazioni preziose capaci di contribuire alla sopravvivenza e ampliare la nostra conoscenza. Resta allora una domanda: perché l'umanità ha scelto di costruire la propria civiltà su questo modello, invece di affidarsi alla pura casualità degli eventi, vivendo semplicemente giorno per giorno? Il nostro cervello si è evoluto nei milioni di anni prevedendo certi fenomeni e cercando di diminuire l'incertezza del mondo: quando gioco d'azzardo, maggiore è la mia abilità nel calcolare le probabilità, maggiore è la sicurezza di portarmi a casa dei soldi che potrò spendere. In secondo luogo, è la spinta alla curiosità dei singoli che premia chi, al costo di qualche rischio, trova nuove risorse nell'ambiente. Sono algoritmi del nostro cervello, dette abitudini, talmente tanto consolidate da guidare gran parte delle nostre scelte quotidiane spesso in maniera inconsapevole. Queste abitudini si formano per ottimizzare l'energia mentale e rendere più efficiente l'interazione col mondo. La raccolta e l'analisi dei dati, in fondo,

⁶ Istat, *Produzione e lettura di libri in Italia: Anno 2022*, AA.VV, 2023, p. 2

https://www.istat.it/it/files/2023/12/REPORT_PRODUZIONE_E_LETTURA_LIBRI_2022.pdf (attivo)

⁷ Consiglio nazionale delle ricerche, *L'editoria scientifica in (poche) parole chiave*, AA.VV, 2024, p. 5

https://www.cnr.it/sites/default/files/public/media/attivita/editoria/9788880806875_7018_testo.pdf (attivo)

non sono che un'estensione meccanica di questi meccanismi biologici. Immaginiamo, ad esempio, che all'improvviso salti la corrente. La nostra prima reazione sarebbe tentare di ripristinarla. Se non ci riuscissimo, usciremmo di casa per cercare qualcuno in grado di intervenire, magari un tecnico o un vicino informato. Una volta risolto il problema, cercheremmo di individuarne la causa del guasto per evitare che si ripresenti in futuro. In ognuna di queste fasi il nostro cervello raccoglie una grande quantità di dati, li seleziona, li interpreta e li trasforma in informazioni utili. Fa esattamente ciò che oggi chiediamo agli algoritmi, dare senso al caos informativo attraverso l'analisi. La svolta in questa epoca è stata nell'automatizzazione della raccolta di questi dati grazie all'informatica; l'era dei dati è anche l'era della Scienza Dei Dati. La data di partenza di questa disciplina non è chiara poiché emerse a seguito di grandi innovazioni, cambiamenti culturali ed eventi storici avvenuti dalla fine del 1800 in poi. Per prima cosa possiamo rintracciare la statistica come elemento chiave di sviluppo laddove gli stati nazionali necessitarono sempre di più di un'analisi dettagliata della situazione socioeconomica del paese per riuscire a gestire *polis* divenute ormai vere e proprie masse di cittadini. Non è un caso che le vittorie di Cavour durante il Risorgimento italiano non si dovettero unicamente ai suoi successi diplomatici, ma anche per un grande interessamento della reale situazione economica e sociale di un paese ancora diviso.⁸ Indici di produzione dell'agricoltura, debiti finanziari, bilanci e volumi di scambi con l'estero entrarono a far parte del linguaggio politico quotidiano richiedendo un calcolo matematico delle risorse.⁹ Il Regno di Sardegna fu l'esempio di un paese che necessitò costante efficienza e dinamismo, laddove gli scarti in termini di produzione, energia e tempo andavano ridotti al minimo, in maniera simile al nostro concetto di stato moderno. Una delle problematiche maggiori dell'Italia riunita nel 1861 fu la mancanza di conoscenza di una parte del proprio territorio. Gli intellettuali del Nord non ebbero a disposizione dati sufficienti sulla popolazione e pertanto non riuscirono a costruire quelle informazioni in grado di determinare politiche efficaci nei confronti del Sud. La scarsità di conoscenza incominciò a divenire sempre di più un problema serio nei decenni successivi, in particolar modo durante le due guerre mondiali. Lo sbarco in Normandia avvenuto nel 1944 fu l'episodio apicale di come le informazioni possano aver deciso la disfatta della Germania nazista nei confronti delle forze alleate. La mappatura del territorio su cui sbarcare, il grado di conoscenza della posizione del nemico, il suo stato psicologico e la disponibilità di soldati furono tutte informazioni che, una volta costruite, permisero di assicurarsi la vittoria. Dai campi di battaglia si passò alle battaglie di pensiero tra gli accademici in cui il tema centrale era la costruzione di elaboratori di calcolo che, nel corso della seconda metà del 1900, riunirono la statistica e la matematica nel calcolo computazionale facendo

⁸ A. Canavero, *Storia contemporanea*, Pearson, 2019, p. 44

⁹ Ivi, p. 45

emergere la figura dei programmatori e degli informatici. Il mondo accademico iniziò a interessarsi ai limiti logici delle macchine, osservando in parallelo i progressi negli studi sul cervello. Da questo intreccio di ricerche emerse un nuovo campo del sapere: l'intelligenza artificiale. Le grandi aziende, a loro volta, cominciarono a seguire da vicino gli sviluppi accademici, sfruttando questa nuova potenza informativa e una metodologia scientifica capace di ridurre l'incertezza nella gestione quotidiana. Sorsero i primi linguaggi di programmazione che permisero di automatizzare il calcolo in maniera sempre più complessa e fu a causa di questo sviluppo che Peter Naur (ricercatore e tra i più importanti pionieri dell'informatica) decise di coniare, in modo molto generico, il termine *Data Science* come quella scienza in grado di cogliere ed elaborare dati.¹⁰ Sebbene oggi la *Data Science* sia associata all'intelligenza artificiale, la connessione fra le due discipline non fu immediata. I modelli richiedevano quantità di dati attendibili, ma i dati potevano essere elaborati solo con tecnologie costose per l'epoca. L'informatica permise lo sviluppo del *software*, mentre per l'*hardware* fu necessario attendere economie di scala rispetto ai materiali necessari. Questo fu uno dei motivi per cui l'intelligenza artificiale ebbe i cosiddetti "inverni ed estati", periodi di grande sviluppo abbinati a successivi arresti nella ricerca nel corso della seconda metà del 1900. Il primo modello d'intelligenza artificiale applicato all'industria fu R1 della Digital Equipment in grado di automatizzare gli ordini per nuovi computer, questo modello dopo pochi anni permise all'azienda un risparmio di milioni di dollari all'anno.¹¹ L'efficacia industriale proiettò alla fine degli anni '90 lo sviluppo del *machine learning* e *data mining*, in cui l'estrazione del dato digitale e la sua elaborazione potevano portare un valore concreto. Si utilizzarono metodologie strettamente matematiche e scientifiche su banche dati in grado di estrarre informazioni utili, in cui le azioni tipiche furono la preparazione, filtraggio e analisi del dato. La *Data Science* incominciò ad avere significativi riconoscimenti dagli anni 2000 in poi quando l'utilizzo dei dati divenne massiccio (grazie anche alla diffusione di Internet) e avere un ritorno monetario. Si susseguirono numerosi tentativi di definirla come una disciplina autonoma con dei propri principi e numerosi dibattiti sorsero riguardo alla sua effettiva natura. Il dibattito si divise in chi la ritenne niente di più di un ramo della statistica, chi tentò di trovare modelli concettuali universali o ancora chi ritenne fosse semplicemente un nuovo paradigma del metodo scientifico basato sui dati. Il tempo e l'economia non diedero spazio a lunghi ragionamenti teorici avvenuti invece in altri campi scientifici come quelli della fisica o della biologia, i dati incominciarono a diventare sempre più numerosi sia grazie al miglioramento delle infrastrutture di comunicazione che grazie all'avvento dei *social media* e delle applicazioni o ancora al progresso costante dei computer. I dati portarono sempre

¹⁰ <https://di.ku.dk/english/news/previous-news/peter-naur/> (attivo). Il libro in questione è: P. Naur, *Concise Survey of Computer Methods*, Petrocelli Books, 1974

¹¹ Jhon McDermott, *R1: A rule-based configurer of computer systems*, Carnegie-Mellon University, 1982

più soldi nelle casse di chi li utilizzò efficacemente, in certi casi così tanti soldi che furono necessarie nuove figure lavorative come il *data analyst* o il *data scientist* in grado di trattare i famosi *big data*, grandi moli di dati spesso non strutturati. Nel primo decennio del nuovo millennio si ritrovò intorno a questa scienza una forte connotazione economica che venne utilizzata dai grandi gruppi industriali per una produzione più precisa e su misura dei clienti. Non fu più il consumatore a adattarsi alla produzione industriale, bensì fu egli stesso a fornire informazioni per personalizzare il bene che avrebbe consumato, offrendo all'industria numerosi vantaggi, come la riduzione degli scarti dei prodotti che non sarebbero andati bene sul mercato e un guadagno in termini di idee ricevute dagli stessi consumatori. Le aziende che basarono le proprie politiche sui dati divennero talmente tanto competitive che esplosero finanziariamente, logicamente gran parte di quelle del settore informatico o ad esso correlato. Come la scrittura, nel momento in cui una certa invenzione o sistema entra a far parte nei meccanismi economici della società incomincia a svilupparsi in modo progressivo prendendo poi conseguenze impreviste dal motivo per cui era nata.¹² La *Data Science* entrò progressivamente in tutti i settori come la medicina, biologia, sociologia per ritornare di nuovo alla politica. Le campagne politiche di Obama condotte da Jim Messina furono *data-driven*, in cui le decisioni vennero prese da metriche ben precise, non ci si affidò più all'intuizione umana.¹³ Un esempio particolare si verificò la sera del 30 agosto 2012, quando Clint Eastwood intervenne durante la *convention* repubblicana inscenando un dialogo con una sedia vuota, simbolicamente utilizzata per rappresentare l'operato presidenziale del presidente Barack Obama. Attraverso la retorica, Eastwood mise in discussione, in chiave ironica, la presunta mancanza di autorevolezza del presidente americano. L'episodio suscitò immediatamente un'ampia reazione sui social media, dove divenne rapidamente di tendenza, alimentando un acceso dibattito pubblico. L'interpretazione dominante sui diversi canali digitali evidenziò come l'intervento fosse percepito come un attacco diretto e mirato alla figura presidenziale. Tuttavia, in tempi brevissimi, Obama rispose efficacemente attraverso un post sui *social* accompagnato da una sua immagine al lavoro e da una frase d'effetto: "*This seat's taken*", riuscendo a ribaltare la provocazione iniziale. Il risultato fu l'immediato rovesciamento della narrativa repubblicana e più di 58.000 *retweet*.¹⁴ Nel giro di poco tempo si assistette a un'accelerazione senza precedenti: velocità, precisione e grandi volumi di dati furono in grado di mutare il destino di un'intera nazione.

¹² John Scales Avery, *The information explosion*, autopubblicato sul sito <https://www.johnavery.info>, 2018, p. 135. Ad esempio, la scrittura nei Sumeri aveva importanza negli scambi commerciali di oggetti come ciotole o argento permettendo di gestire la complessità politica e sociale delle prime civiltà. Dopo un lungo processo la scrittura venne impiegata per diffondere informazioni di varia natura come narrazioni, poesie o manifesti sociali.

¹³ Marco Cacciotto, *Il nuovo marketing politico: Vincere le elezioni e governare al tempo della politica veloce*, Il Mulino, 2019, p. 152

¹⁴ Ivi, p. 155

Fino ad arrivare ad oggi, laddove un mondo socialmente algoritmico necessita dei dati come fonte di potere in grado di controllare perfino il nostro concetto di verità. In pochissimo tempo, forse in poco meno di un secolo, si è potuto constatare che le metodologie utilizzate nella *Data Science* funzionano in modo molto efficiente. Grandi occasioni ed eventi ne hanno modificato un influsso teorico già accelerato consentendo di mostrare la sua intrinseca multidisciplinarietà.¹⁵ Difficilmente nei documenti scientifici, dalla geologia fino alla linguistica, non si trova qualche dato che è stato trattato e va a supporto di una teoria invece che di un'altra. Inoltre, sarebbe sbagliato non ammettere che la società ricerca la soddisfazione dei propri bisogni, sia che siano materiali o immateriali. La diffusione dei dati e la loro importanza sono determinati anche da un contesto sociale molto favorevole laddove la carenza di cultura, dei valori politici e della filosofia richiedono l'afflusso di tante e nuove idee per colmare questi vuoti, se poi arrivano da algoritmi, dalle nostre abitudini quotidiane o ancora dall'unione fra le due cose poco importa. Bisogna fare attenzione, però: l'informatica unita alle scienze cognitive, statistica, filosofia e politica torneranno rilevanti nel secondo capitolo quando si costruirà un "modello del mondo", mostrando come i dati, le informazioni e la stessa *Data Science* utilizzati in contesti anomali presentano problemi logici di una scienza che si sta sviluppando troppo velocemente. Partiamo però dal principio, quando si sono diffusi i primi dati?

1.1 L'evoluzione dell'informazione dal DNA fino all'intelligenza artificiale

Nel corso della storia dell'universo l'informazione ha sempre avuto un ruolo centrale nei processi di organizzazione, adattamento e sopravvivenza. Molto prima dell'invenzione della scrittura gli atomi si aggregarono fra di loro formando le galassie e la Terra, in cui nel caos primordiale l'informazione si trasmise in modo efficace tramite il codice genetico. RNA e DNA rappresentarono i primi linguaggi complessi con cui la natura riuscì a codificare, trasmettere e replicare informazioni vitali, affidando alla biologia il compito di conservare e rinnovare la memoria delle specie. Che sia sorto prima il dato o l'informazione non è ancora possibile saperlo, di certo sappiamo che la diffusione di dati aggregati fece esplodere la vita organica velocizzando il processo di evoluzione. Con la nascita di organismi sempre più complessi, come piante e animali, l'ambiente si diversificò migliorando in maniera bidirezionale sia il modo in cui venivano costruite informazioni utili alla sopravvivenza sia come venivano diffuse. Comportamenti sociali e riproduttivi complessi alterarono in continuazione le probabilità di vittoria in un ambiente da una specie all'altra, richiedendo adattabilità nel tempo e un sistema efficace di apprendimento da

¹⁵ La *Data Science* è convenzionalmente accettata come disciplina che si avvale al suo interno della matematica, statistica, informatica e logica. Esternamente viene ormai utilizzata in tutte le altre discipline sia da quelle quantitative che sociali. A causa di questa sua adattabilità è estremamente vitale e soggetta a continue influenze e mutamenti, infatti, di modelli concettuali su come trattare dati se ne trovano moltissimi online.

trasmettere alle generazioni successive. I microrganismi sfruttarono fenomeni di adattabilità ad ambienti diversi e talvolta a condizioni estreme, le piante una chimica efficace e il movimento meccanico dell'ecosistema come vento e acqua per riprodursi su grandi distanze. Infine, gli animali svilupparono sistemi nervosi arrivando a sistemi centralizzati come il cervello. Il sistema nervoso permise una rapida comunicazione degli stimoli ricevuti e una reazione immediata, dapprima difficoltosa come il movimento in un dato spazio o il dolore, successivamente informazioni sensoriali più complesse che portarono a comportamenti sociali. Nel tempo tramite un costante bilanciamento tra informazioni acquisite e informazioni consolidate si formarono altri comportamenti come la curiosità, premiando chi esplorava l'ambiente con nuove informazioni utili alla sopravvivenza. Curiosità, adattamento e innovazione non sono parole richieste per una società divenuta industriale con un lavoro altamente specializzato, bensì sono caratteristiche che si sono sviluppate nei milioni di anni in natura.

In questo contesto la nostra specie trovò vita fertile riuscendo a sviluppare sempre di più un'altra caratteristica già presente: la capacità di prevedere eventi futuri e di comprenderne le conseguenze. L' *Homo Habilis* possedeva all'incirca 650 centimetri cubi di cervello mentre l'*Homo Sapiens* circa 1400 centimetri cubi.¹⁶ L'espansione del cervello permise l'emergere di comportamenti sempre più complessi, quali la creazione di strumenti, la costruzione di intricate relazioni sociali e, infine, lo sviluppo del linguaggio, che consentì di trasmettere informazioni e di modificare la rappresentazione del mondo. Si svilupparono sistemi di credenze, miti e convenzioni sociali. Si assistette all'elaborazione di nuovi codici di trasmissione del sapere, tra i quali la scrittura occupò un posto privilegiato. Quest'ultima fu in grado di ridurre drasticamente i tempi di circolazione delle informazioni, permettendo loro di viaggiare anche per millenni con un livello minimo di perdita informativa.

Nel corso dell'età moderna l'umanità si è abituata a scrivere in misura sempre maggiore, soprattutto in rete, favorita dall'economicità e dalla rapidità di diffusione delle telecomunicazioni, ma ha finito per perdere di vista la centralità della scrittura nel progresso umano. La scrittura costituì una delle spinte fondamentali al perfezionamento continuo del linguaggio, ma soprattutto si rivelò un mezzo di straordinaria versatilità. Così come il DNA risultò capace di mantenersi stabile e al tempo stesso sufficientemente adattabile, come nel caso dei fenomeni epigenetici o delle mutazioni, la scrittura, a prezzo di una minore conservabilità, si mostrò estremamente variabile, adattandosi costantemente nel corso delle epoche.

Dall'invenzione della scrittura questa tecnica è presente in tutti i passaggi storici più importanti: dall'epoca antica con i papiri, all'epoca medievale con i testi fino al miglioramento progressivo della stampa moderna. Questo è dovuto anche al fatto che la scrittura e il linguaggio presentano

¹⁶ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, p. 20

una base matematica e fisica notevole. Il linguaggio può essere definito come un insieme di onde sonore che producono vibrazione meccanica dell'aria recepite dall'orecchio dando un certo tipo di rappresentazione del mondo in maniera veloce. La scrittura, al contrario, essendo ragionata durante la sua composizione, almeno fino in epoca passata, permette al mittente di costruire un'informazione in maniera più complessa e dettagliata. Il parlato e lo scritto, rispetto al visivo, hanno il vantaggio di rappresentare il mondo in maniera sequenziale. Ad esempio, la frase: “La mela di Giovanni è stata comprata al mercato”, descrive uno stato della realtà in cui la parola successiva si combina con quella precedente costruendo man mano il significato. Oltretutto, Il linguaggio scritto pur essendo più lento di quello visivo, ha una capacità informativa esponenziale. Vedere una mela e rappresentare il suo stesso concetto divengono nella nostra era due cose completamente diverse, la parola “mela” è in grado di rappresentare tutte le mele del mondo, un'intera classe di oggetti. L'era del caos informativo non è dovuta solamente ad un progressivo miglioramento delle tecnologie informatiche, bensì anche al progressivo miglioramento dei testi scritti. In merito a questo basterebbe confrontare un testo delle medie di oggi contornato da immagini, testo curato e link di approfondimento online rispetto ai testi delle medie di soli 50 anni fa.

In seguito, si assistette ad una svolta quando, dal 1500 in poi, invece di conservare le proprie credenze, seppur sbagliate, gli intellettuali incominciarono a dubitare delle teorie antiche cercando di colmare il vuoto di conoscenza con degli esperimenti.¹⁷ La rivoluzione scientifica permise, con sempre maggiore convinzione, di aumentare la quantità di investimenti da parte dei governi nella ricerca per accrescere le informazioni sul mondo e sfruttarle a proprio vantaggio. Gli umani furono disposti, attraverso numerosi errori e prove empiriche, ad ammettere la propria ignoranza nei confronti del mondo che li circondava. Questo passaggio logico permise di creare tecnologie in grado di osservare e manipolare al meglio la materia. Fa riflettere come in soli cinque secoli il paradigma del metodo scientifico sia progressivamente cambiato. L'osservazione diretta è stata sostituita da algoritmi e *software* di simulazione, le prove empiriche da milioni di documenti scientifici diffusi online. Come sostiene lo storico Harari, questa fu una rivoluzione dell'ignoranza, ovvero la consapevolezza del bisogno di accedere ad una nuova conoscenza in cui l'elemento centrale è il dubbio.¹⁸ Grandi scienziati iniziarono ad indagare la realtà per la passione di scoprire nuove cose, per il “gusto della scoperta” come direbbe il fisico Feynman.¹⁹ La scienza concepita in questo senso si ritrova in personaggi come Benjamin Franklin, Marie Curie e

¹⁷ Yuval Noah Harari, *Sapiens: Da animali a dèi. Breve storia dell'umanità*, Giunti editore, 2017, p. 310

¹⁸ Ivi, p. 314

¹⁹ Feynman già nel corso della seconda metà del 1900 intravede una tendenza alla degenerazione nell'etica degli scienziati rispetto al modo di fare scienza. Lui stesso vi si ritrova implicitamente nella critica quando ripensa alla costruzione e al lancio della bomba atomica in una guerra mondiale, di fatto, praticamente finita. Descritto in modo esaustivo nel libro “Richard P. Feynman, *Il piacere di scoprire*, Adelphi edizioni, a cura di Jeffrey Robbins, 2002.”

Alexander Von Humboldt, forse uno degli ultimi scienziati in grado di produrre teorie generali dal particolare per esperienza diretta. Alexander Von Humboldt, eclettico nel sapere rispetto all'iperspecializzazione odierna, alternò lo studio letterario alle esplorazioni svolte nell'ambiente naturale del Sudamerica. Grazie ai suoi viaggi, oggi viene associato ai movimenti ambientalisti nati un secolo dopo la sua morte per la sua idea di una natura che non andava sezionata, bensì vista come un sistema interconnesso in cui ogni elemento contribuiva a modificare tutti gli altri. Il paradigma scientifico ha portato a nuove tecnologie negli ultimi decenni, tra cui Internet, in grado di distruggere spazio e tempo creando un ecosistema multimediale in grado di diffondere informazioni rilevanti in modo massiccio e in qualsiasi parte del mondo. Il Novecento rappresenta un secolo fondamentale laddove non ci si limita più a trasmettere informazioni, bensì avviene uno studio sistematico delle stesse. Scienziati come Alan Turing indagano sui limiti di logica del nostro mondo, altri come Peter Naur indagano sulla natura dei dati, altri ancora tentano di formulare impianti teorici veri e propri. In merito a questo è importante la teoria dell'informazione formulata nel 1948 da Claude Elwood Shannon rispetto all'applicazione pratica della matematica sulla trasmissione di un messaggio.²⁰ Ci sono numerose teorie che andrebbero citate, ma la teoria dell'informazione si presenta come un'equazione elegante in senso matematico: sintetica, semplice e generalizzabile rispetto a un modello di realtà. L'equazione ha lo scopo di fornire una descrizione quantitativa del concetto d'informazione, oltre che a studiare i fenomeni che avvengono nei canali di comunicazione. Il primo passaggio è la definizione di entropia (H) come misura della quantità di informazione o incertezza presente nel messaggio, maggiore sarà l'entropia, maggiore è l'imprevedibilità di un evento. Nell'equazione in questione la realtà che si vuole rappresentare è declinata all'ambito informatico in cui la costante ($-K$) e il logaritmo in base 2 vanno a rappresentare i bit. Successivamente per ogni evento (n) si va a calcolare la probabilità dell'evento (P_i) in cui il logaritmo (\log) va a trasformare la probabilità dell'evento nella quantità d'informazione.

$$H = -K \sum_{i=1}^n p_i \log_2 p_i$$

Il concetto diventa immediato tramite degli esempi. Tenendo conto un lancio di una moneta in cui i due possibili eventi testa e croce hanno il 50% di probabilità, se si fanno i calcoli presupponendo la costante ($-K = -1$) e il logaritmo (\log) abbia base 2, l'entropia (H) sarà uguale a 1 bit d'informazione. Ripetendo nuovamente lo stesso calcolo, se prendiamo una moneta truccata in

²⁰ C. E. Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal e ristampato da <https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>, 1948, pp. 10-11

grado di dare sempre testa al 100% e croce allo 0% in cui gli eventi si presentano prevedibili e ordinati, l'entropia sarà uguale a 0 bit. L'equazione di Shannon permette un calcolo matematico preciso dell'informazione e dell'incertezza contenuta in un evento, una formula che va a rappresentare un modello di realtà. Divenuta talmente influente che ad oggi rappresenta un punto di riferimento sul concetto di trasmissione delle informazioni in numerose discipline.

Da qui in poi l'umanità ha subito un processo di relativa accelerazione. In primo luogo, la capacità di analizzare grandi quantità di dati diversificati tramite strumentazioni apposite in grado di ampliare i nostri sensi in maniera meccanica. In secondo luogo, la velocità di propagazione delle informazioni grazie a sistemi di comunicazioni sempre più veloci. L'accelerazione dello sviluppo deriva dalla nostra capacità di aggregare dati e costruire informazioni in modo veloce, efficiente e distribuibile nello spazio con poca perdita informativa. Inoltre, il potenziamento dell'informatica e i progressi delle neuroscienze hanno reso possibile la creazione dei primi modelli di intelligenza artificiale, traducendo molte delle nostre abitudini quotidiane in algoritmi stabili e riproducibili nel tempo. Questo processo ha segnato un vero e proprio passaggio da un paradigma organico a uno meccanico, in cui il comportamento umano viene formalizzato e simulato da sistemi artificiali. Tuttavia, la realtà è molto più complessa di quanto possa sembrare, come si può intuire confrontando un neurone biologico con un neurone artificiale. I nostri neuroni biologici sono composti da dendriti ovvero prolungamenti ramificati da cui ricevono input di centinaia o migliaia di altri neuroni, un corpo cellulare che elabora gli input ricevuti e un assone che corrisponde ad un prolungamento che si ramifica alla fine trasmettendo degli output nervosi.²¹ La trasmissione dell'informazione da un neurone all'altro può avvenire sia in forma elettrica, rapida ed efficiente grazie al movimento di ioni attraverso la membrana, sia in forma chimica, lenta e versatile, grazie al rilascio di neurotrasmettitori. L'informazione nel sistema nervoso viene modulata in maniera esponenzialmente molto complessa, da neurone a neurone, da neurone ad un altro tipo di cellula o ancora dalla cellula ad un neurone.²² Per esempio, un neurone dopo aver ricevuto un input lo elabora e potrebbe comunicare con una cellula muscolare tramite un output come il neurotrasmettitore acetilcolina.²³ Un neurone biologico ha modi differenti di comunicare con altri neuroni e con altre cellule rendendo più complesso il comportamento etologico di un animale. Un percorso durato milioni di anni in grado di adattarsi finemente all'ambiente circostante, accrescendo di volta in volta le competenze e ottimizzandole a livello calorico. Analisi dei dati, elaborazione e miglioramento energetico degli organi di senso sono stati parti cruciali per rendere il nostro sistema nervoso adattabile come lo riscontriamo oggi. Al contrario, i ricercatori hanno avuto pochissimo tempo per sviluppare l'intelligenza artificiale cercando di

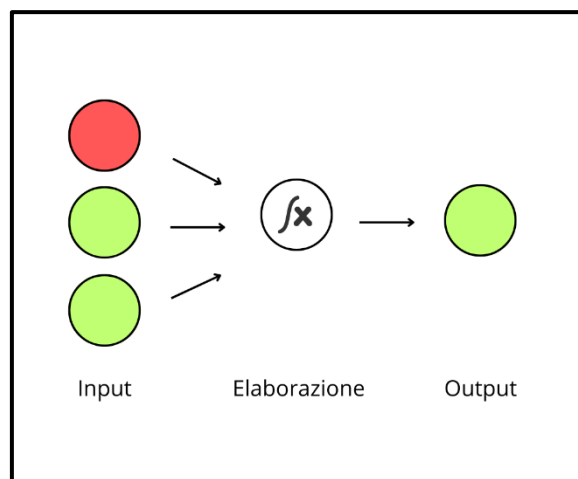
²¹ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, p. 91

²² Ivi, p. 85

²³ Ivi, p. 97

formalizzare nella maniera più semplice il comportamento dei neuroni biologici. Parliamo quindi dei primi neuroni artificiali programmati per ricevere input (dati grezzi), elaborarli tramite funzioni di attivazione in grado di emulare matematicamente l'inibizione (0) o l'eccitazione (1) di un neurone biologico, producendo infine un output.²⁴

Figura 2-Modello concettuale di un neurone artificiale in cui gli input rossi e verdi vengono analizzati da una funzione simbolica, producendo un output.



Autore: Lorenzo Tiari

In un mondo informatico primitivo i limiti dei neuroni artificiali sottostavano ai limiti di logica dei primi linguaggi di programmazione. Un neurone artificiale, quindi, divenne una rappresentazione scheletrica di quello biologico ponendo numerose barriere come quelle di computabilità o di costi elettrici in termini efficienza-efficacia; infatti, la tendenza negli ultimi anni da parte dei ricercatori è di fondere vantaggi e svantaggi dell'una e dell'altra parte. Tentativi sono le *spiking neural network*, la costruzione di chip neuromorfici e la nascente bioinformatica.²⁵ Sarebbe semplicistico ridurre la questione in questi termini. Nel tempo si è compreso che ogni tipo di realtà che può essere quantificata e quantificabile viene piegata non molto tempo dopo da leggi ferree della logica matematica. Prendiamo come esempio il DNA. Il DNA possiede quattro basi fondamentali: adenina, guanina, citosina, timina e ognuna di queste basi si lega solamente a un'altra in uno schema elegantemente ripetuto nella storia biologica. Una volta formalizzato il concetto, per quanto inizialmente sintetico, diviene assolutamente replicabile e migliorabile in

²⁴ Un esempio di funzione di attivazione è quella di Tanh, detta tangente iperbolica, in grado di riunire i valori di input tra -1 e 1. Questo tipo di funzione d'attivazione mostra come l'intelligenza artificiale necessiti sempre di più una logica probabilistica invece di una deterministica per arrivare all'efficacia di una cellula biologica.

²⁵ Le *spiking neural network* sono reti ad impulso profondamente ispirate alla biologia. Per esempio, viene incorporato il concetto di tempo in una cellula come il periodo di riposo tra un impulso da un neurone ad un altro neurone. Oltretutto l'estrema sincronizzazione è il motivo per cui le reti neurali classiche sono dispendiose, quando pensiamo è raro che tutte le zone cerebrali vadano alla massima potenza poiché richiederebbe troppa energia calorica.

modo iterativo nel tempo. D'altronde, è esattamente ciò che fanno le cellule biologiche in natura da sempre, trovata la regola che funziona la si applica finché un'altra regola dimostra di funzionare meglio. Ridurre il campo dell'intelligenza artificiale ad un mero studio di macchine che simulano scheletricamente le cellule è sbagliato, come è sbagliato pensare che il mondo biologico sia inefficiente e superato. In merito a questo, nel secondo capitolo si rileverà come l'errore sia parte essenziale nel miglioramento dell'informazione, tant'è che un tipo apprendimento molto di moda nell'ambito dei modelli artificiali è l'apprendimento per rinforzo, ovvero un apprendimento in cui si programma il modello a sbagliare per potersi migliorare costantemente.

Lo sviluppo di tutte queste discipline ha migliorato gli algoritmi e la nostra comprensione del mondo facendo sorgere diverse domande, le quali sono di grande interesse per lo sviluppo di questa disciplina. La più cruciale di tutte è se le macchine possano sviluppare una coscienza? Una domanda particolarmente complessa e dibattuta tra i più grandi scienziati cognitivi del mondo, ma assolutamente necessaria per il trattamento dei dati. In un futuro non molto lontano la ricezione di dati aggregati e la trasformazione in informazioni sarà spesso sottoposta da parte di agenti artificiali e se le scienze cognitive saranno in grado di rispondere a questa domanda potremo capire se esiste una differenza fra l'intenzione di un'azione oppure una sua simulazione, con ricadute evidenti anche nel diritto.²⁶ Proviamo a pensare ad un *data analyst* di un futuro lontano che, utilizzando un modello artificiale avanzato per il trattamento di dati biometrici, produce errori evidenti andando a ledere la vita delle persone. Per difendersi davanti alla giustizia potrà imputare la colpa al modello utilizzato? Attualmente nelle scienze cognitive sono stati fatti enormi passi in avanti, sappiamo per esempio che l'esperienza del mondo è strettamente connessa con il nostro cervello fisico, ma la scienza non è ancora in grado di spiegare come da una scarica elettrica si possa generare la sensazione del profumo di una rosa.²⁷

1.2 La sopravvivenza del più validato

Un elemento cruciale per la sopravvivenza in un dato ambiente è capire se ciò che l'organismo sta facendo sia giusto o meno. Una volta costruite le informazioni persone, animali, piante, virus, batteri e macchine condividono tutti lo stesso destino, ovvero l'errore. Non tutto ciò che viene diffuso è vero, affidabile o utile e l'obiettivo diviene raggiungere l'ordine nell'era del caos informativo. Nel mondo tutti i fenomeni, in maniera diversa, sono sottoposti a metriche di validazione. Nella vita quotidiana si possono pensare ai *like* sui *social media* rispetto ad un certo tipo di contenuto oppure al numero di ripubblicazioni di un articolo scientifico.

²⁶ Per agente artificiale si intende un modello in grado di ricevere dati e compiere azioni senza l'intervento umano.

²⁷ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, p. 23

L'aggettivo valido deriva dal latino *validus*, godere di buona salute o essere forte.²⁸ Il duplice significato porta a riflettere su due modelli di pensiero diversi: un'informazione gode di buona salute se nel tempo non mostra particolari segni di invecchiamento. Per esempio, si può ritenere valida l'informazione in etologia che il comportamento degli animali può essere modificato tramite chemiotassi, ovvero la modifica del comportamento rispetto a sostanze chimiche. Quest'informazione gode tuttora di buona salute. Secondariamente l'informazione si può definire forte quando s'impone rispetto alle altre. Per esempio, l'informazione secondo cui sia sbagliato entrare nella proprietà altrui è forte, ma non gode di buona salute. La forza della sua validazione è dettata dal fatto che la maggioranza può imporla come rappresentazione del mondo, una rappresentazione che viene a meno se consideriamo che il denaro e la proprietà sono concetti utili, ma unici nella nostra specie rispetto a tutte le altre. La popolazione dei Liguri, prima dell'espansione di Roma nell'Italia antica, possedeva una debole gerarchizzazione della società in cui gran parte dei beni erano in comune e vigeva uno spirito egualitario. Solo in un momento successivo con il mutamento politico e il miglioramento dei beni si formerà una classe aristocratica. Molto probabilmente prima di quel mutamento l'informazione non sarebbe stata valida all'interno dei rapporti sociali dei Liguri. Il concetto di *validus* assume in modo alternato valenza positiva o negativa sulla base della momentanea rappresentazione del mondo. Talvolta si possono presentare casi più complessi in cui l'informazione costruita diviene valida solo nel momento successivo a quando la approviamo. Caso tipico in medicina è l'effetto placebo, ovvero quando una piccola bugia diviene verità certa. Per esempio, sappiamo che mentire su un farmaco scambiandolo con acqua e sale oppure zucchero dato ai bimbi che, in mancanza di certi tipi di conoscenze mediche, può portare a reazioni neurologiche e muscolari in grado di migliorare la situazione interna.

Delineato uno scenario del genere la validazione dell'informazione assume un'importanza vitale da cui può dipendere la vita o la morte, l'utilizzo del modello o la cancellazione del suo codice. I sistemi di validazione concettualizzati e ancora oggi utilizzati hanno dovuto essere forti rispetto alla dura prova empirica e di buona salute per resistere nel tempo. Il problema principale, però, è che di modelli ne esistono moltissimi e affondano in svariati campi del sapere rappresentando un malessere costante per i *data scientist* che cercano di trovare un modo comune per determinare un dato. Di conseguenza per questione di sintesi ne verrà considerato qualcuno per capire quanto possa essere complicato trovare un sistema univoco che possa mettere d'accordo tutti.

Proviamo a immaginare come una persona o una macchina potrebbe validare un'informazione, partendo da esempi concreti. Si può immaginare un contesto di 30.000 anni fa, prima della nascita delle grandi civiltà. Un gruppo di antenati si ritrova ad esplorare una zona sconosciuta alla ricerca

²⁸ <https://www.treccani.it/vocabolario/valido/> (attivo)

di cibo. Addentrandosi in un bosco vengono scoperte lunghe tracce curve nel fango e proseguendo vengono avvistati alcuni animali distesi a terra con evidenti segni di morsi. A questo punto, la mente di un membro del gruppo collega i segni osservati a un possibile predatore: probabilmente un serpente. Pochi minuti dopo, l'ipotesi viene confermata dall'apparizione dell'animale nelle zone vicine. Questo tipo di validazione possiamo definirla come "comprensione del contesto", un modello etologico (basato su esperienze passate nell'ambiente naturale) e statistico (si valuta la probabilità dell'evento sulla frequenza normale con cui si verifica) in cui un serpente rappresenta un elemento perfettamente coerente nel contesto bosco, anche prima della presenza effettiva del predatore. L'informazione, anche in assenza della prova definitiva, si sarebbe potuta ritenere plausibile. Si può immaginare ora una situazione diversa. Durante l'esplorazione nel bosco un membro vede certamente il serpente che striscia nelle vicinanze e decide di comunicarlo agli altri membri del gruppo che, in assenza di prove, devono credergli sulla parola. In questo caso il sistema di validazione non si basa sull'esperienza diretta, ma dalla "credibilità della fonte". Interviene il concetto di reputazione, definendola come l'insieme delle informazioni che producono un modello predittivo sociale, in cui l'informazione ottenuta nell'ambiente può essere validata solamente con la probabilità che qualcuno stia dicendo il vero. Un ultimo metodo potrebbe essere una "verità socialmente accettata". In questo caso, il gruppo forte della memoria di tutte le esplorazioni fatte in precedenza, analizza le prove del passaggio del serpente attraverso un approccio consolidato nel tempo. Anche in assenza di una prova diretta, se combaciano i segni avviene un confronto sistematico tra i membri e solo allora si può validare l'informazione serpente.

Ora facciamo un salto avanti nel tempo di 20.000 anni nel funzionamento di un algoritmo in una macchina. Il compito del modello è verificare se i dati relativi ai bambini di un certo asilo siano stati registrati correttamente da parte del servizio nazionale di un paese. Il primo passaggio consiste nel valutare se l'informazione sia coerente rispetto al dataset. Per esempio, se tutti i tipi di dati espressi sono numerici e sono presenti alcune stringhe di testo allora l'algoritmo rileverà un errore rispetto al tipo di dato inserito. L'informazione viene invalidata tramite un "controllo semantico" producendo errore nel codice e rimuovendolo dal dataset. Un secondo metodo è quella che in statistica si definiscono "valori fuori scala", ovvero quei valori che sono anomali rispetto alla media.²⁹ Se l'età dei bambini si attesta mediamente sui 6 anni e troviamo un'informazione pari a 70 come età anagrafica, l'algoritmo la riconoscerà come anomala rispetto alla distribuzione prevista e verrà rimosso il valore. Un altro modo è la validazione per "integrità e provenienza" dei dati. La macchina può verificare se il dato ricevuto proviene da una fonte attendibile o se risulta alterato. Per esempio, l'algoritmo potrebbe esaminare i metadati del pacchetto informativo

²⁹ S. Borra e A. Di Ciaccio, *Statistica: metodologie per le scienze economiche e sociali*, Graw Hill, 2021, p. 60

e rilevare che l'indirizzo IP del mittente appartiene a un paese estero non autorizzato. In questo caso l'informazione verrà segnalata o scartata poiché ritenuta potenzialmente compromessa. Attualmente gli algoritmi d'intelligenza artificiale utilizzano sistemi di validazione integrati che si attivano prima di elaborare dati, durante il processo e dopo aver inviato il risultato. Questi sistemi hanno nomi molto raffinati come *pre-processing*, *cross-validation*, *feedback*. Di certo un processo comune anche negli organismi biologici, ad esempio il nostro sistema integrato a cui gli scienziati hanno dato invece un nome meccanico nella pronuncia: il sistema neuroendocrino immunitario, un sistema che richiede un discreto dispendio di energia.³⁰

Nella tesi verrà presa in esame la validazione intesa nel senso preposto dal neuroscienziato Daniel Bor che tenta di avvicinare l'informatica alle scienze cognitive, quella che definirò semplicemente come una validazione per "rilevanza". Questo tipo di validazione non pone come assunto una valenza benevola o malevola bensì come l'evoluzione plasmata e portata avanti "idee" del mondo che facciano sopravvivere nel tempo.³¹ Tutti gli organismi in grado di rappresentare il mondo in maniera dettagliata hanno più possibilità di andare avanti mentre chi ha "idee antiquate" viene lasciato indietro. Il mondo è in continua trasformazione e informazioni non aggiornate potrebbero rappresentare la morte; pertanto, un certo organismo deve vivere tra il raggiungimento di un equilibrio di conoscenza e una sufficiente creatività per migliorarle.³² Nel corso della storia dati aggregati trasformati in informazioni hanno reso più complessa la nostra rappresentazione del mondo alterando il modo di ricevere altri dati. Se nei primi animali i sensi erano inefficienti e inaffidabili, via via nel tempo c'è stato un progressivo miglioramento. L'occhio può immagazzinare pacchetti di luce e il cervello è in grado di trasformarli nella sensazione di colori, permettendoci di associarli a rappresentazioni negative o positive del mondo. Se quella struttura tridimensionale striscia nel fango e ha colori diversi, probabilmente non è una buona idea mangiarlo o avvicinarsi. Pertanto, l'evoluzione ha reso la natura sempre più complessa e gerarchica, laddove in ogni singolo livello si possono trovare modelli informativi.³³ L'organismo umano ha un cervello, il cervello è composto da tessuti, i tessuti sono composti da cellule, le cellule sono composte da molecole, le molecole da atomi e così via. Per "rilevanza" allora s'intende la capacità di costruire informazioni rilevanti in un dato ambiente che permettano di

³⁰ Il sistema neuroendocrino immunitario si riferisce ad un modello concettuale nuovo in cui l'elemento cardine è l'integrazione delle informazioni tra il sistema nervoso, endocrino e immunitario che operano al mantenimento dell'organismo. Un caso lampante è l'intervento di questo sistema durante la febbre, obbligando l'organismo ad alzare la sua energia termica.

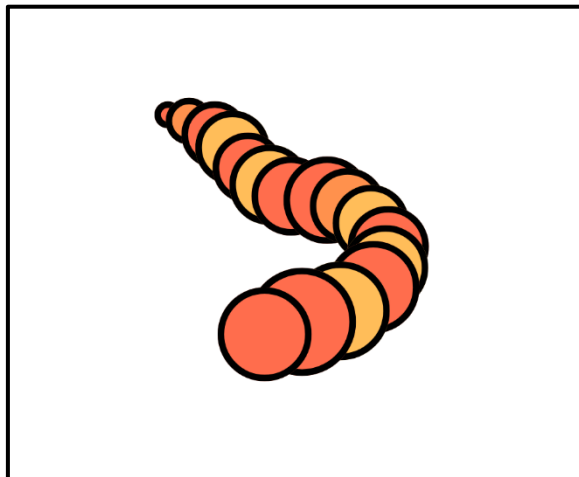
³¹ Daniel Bor, *La voracità del cervello: cosa spinge la nostra coscienza verso un'insaziabile ricerca di significato*, Castelveccchi, 2012, pp. 49-50

³² Ivi, p.51

³³ Ivi, p. 83

rappresentarlo, laddove il concetto di “giusto” o “sbagliato” è relativo e sottoposto a continua analisi.

Figura 3- Applicazione dell’effetto stereocinetico, dimostrando come un semplice disegno a due dimensioni sforzi il cervello ad associare e completare le informazioni.



Autore: Lorenzo Tiari.

1.3 Data Science o Data Sapiens: un nuovo modo di conoscere?

Una domanda cruciale per la *Data Science* è indagare la natura di un dato e strutturarla di modo da diffondere informazioni in grado di aumentare la nostra conoscenza del mondo. La crescente complessità degli organismi ha permesso una rappresentazione del mondo sempre più dettagliata trasponendo all’informatica il dilemma di Kant: sono le informazioni che si conformano ai dati che riceviamo dal mondo o sono i dati del mondo a conformarsi alle informazioni che abbiamo costruito in precedenza?

La ragione umana ha questo singolare destino in una specie delle sue conoscenze: che essa viene tormentata da questioni che non può evitare, perché le sono imposte per natura, ma che neppure può risolvere, perché superano ogni suo potere. In questo imbarazzo essa cade senza sua colpa. Comincia con principi, il cui uso è inevitabile nel corso dell’esperienza, e che anche da questa è sufficientemente garantito. Con essi la ragione sale sempre più in alto, verso condizioni sempre più lontane [...] ma con ciò essa cade in oscurità e contraddizioni, da cui può bene arguire che alla base debbano esservi errori nascosti; senza peraltro riuscire a scoprirli, perché i principi di cui si serve, sorpassando i limiti di ogni esperienza, non riconoscono più la pietra di paragone dell’esperienza.³⁴

Una domanda alla quale questa scienza non può più sfuggirvi se deve essere impiegata in modo massiccio a supporto della ricerca scientifica e nel mantenere la struttura di uno stato sociale

³⁴ Immanuel Kant, *Critica della ragion pura*, Editrice Morcelliana, a cura di Gustavo Bontadini, 2020, pp. 3-4

divenuto sempre più algoritmico. Non può più sfuggirvi se si decidesse di cambiare il paradigma del nostro modo di conoscere: non più una conoscenza dettata dall'esperienza qualitativa, ma una conoscenza dettata da enormi quantità di dati prelevati da fonti di natura diversa e poi aggregate insieme. L'umanità ha influenzato inevitabilmente il modo con cui gli algoritmi elaborano trasponendo gli stessi limiti di logica, ora gli algoritmi stanno influenzando inevitabilmente il modo con cui gli individui pensano le idee del mondo.

La *Data Science* trasformerà la nostra specie in *Data Sapiens*? Bisogna considerare che tutti i fenomeni apparentemente caotici a cui la società sta assistendo negli ultimi decenni in realtà sono perfettamente normali nel corso dell'evoluzione. Partiamo dal fenomeno dei *big data*, dati che hanno certi tipi di caratteristiche come il volume, velocità e varietà in grado di portare valore economico. L'assunto che solamente nell'ultimo decennio l'umanità sia sottoposta a grandissimi quantitativi di dati è sbagliato, perché è sbagliato il modo di pensare ad un dato. I nostri occhi tramite la retina devono acquisire un enorme quantità di pacchetti di luce, detti fotoni, la cui energia elettromagnetica determina se le cellule andranno a segnalare un certo tipo di impulso al cervello.³⁵ Se l'energia è troppo alta, tipica della luce ultravioletta, le cellule potrebbero danneggiarsi e provocare una serie di errori nell'organismo. Se l'energia è troppo bassa, tipica degli infrarossi, l'occhio non può generare un impulso affidabile. In questo processo si parla di uno strumento in grado di ricevere grandi quantità di dati, elaborarli e filtrare ciò che non è rilevante, infine trasformare questi dati in informazioni come: "Questa cosa che vedo è di colore verde". Se poi si richiede di elaborare enormi quantitativi di dati sequenziali (quindi non più simultanei come le immagini) di un testo di una notizia online in cui ogni parola e numero precedente determina di significato quella successiva, in quel momento semplicemente si sta richiedendo un compito diverso in cui gli organi dovranno imparare a adattarsi nel corso del tempo. Un'altra questione è l'automatizzazione della conoscenza, ovvero ottenere dati da fonti certe e produrre informazioni in modo costante. Il DNA e l'RNA in questo compito eccellono da molto tempo in cui, tramite abitudini consolidate, si prelevano dati per produrre le proteine delle cellule. I problemi nella conservazione e distribuzione delle informazioni potrebbero essere il degradamento delle strutture chimiche ed elettriche oppure un virus biologico che va ad alterare l'informazione. Nulla di diverso dall'ossidazione che potrebbe subire un data center a causa di vento e pioggia o di un virus, stavolta informatico, che vada ad alterare il contenuto dei byte. Un'ultima questione di rilievo è quello che i sociologi chiamano "accelerazionismo" o "singolarità tecnologica", fenomeni sociali che prevedono un aumento improvviso delle tecnologie e quindi dell'evoluzione umana. La capacità di gestire dati e l'introduzione di algoritmi come una serie di passaggi logici, è nient'altro che una costante dell'universo in cui l'unica

³⁵ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, pp. 191-192

certezza è l'imprevedibilità. L'espansione degli atomi e la loro aggregazione in pianeti, l'esplosione della vita animale, l'estinzione dei dinosauri come battuta di arresto nel lungo periodo della vita organica sono fenomeni perfettamente normali. L'evoluzione è un processo di progresso, regresso, espansione o costanza che non segue una linea dritta e spesso esegue curvature inimmaginabili. Sostenere che l'umanità sia di fronte ad una singolarità poiché nella rappresentazione odierna del mondo l'informazione rilevante è che all'aumentare del tempo si velocizzeranno le tecnologie rappresenta un errore sistematico prelevato da un campione in scala di qualche migliaio di anni a questa parte. Se arbitrariamente tra le varie ipotesi si decidesse di fissare la nascita della vita organica circa 3.8 miliardi di anni fa, tenendo conto che l'umanità ha avuto l'ultimo punto in comune con gli scimpanzè prima di differenziarsi circa 6 milioni di anni fa, indicativamente da quel momento è passato lo 0.16% della storia complessiva.³⁶ Si sarebbero dovuti occupare altri 32 milioni di anni (ovvero ripetere per poco più di 5.3 volte la nostra durata come specie) per riuscire ad arrivare ad un misero 1%. Qualsiasi *data scientist* che lavora nel sociale in grado di proporre un'idea con certezza quasi massima su un'informazione di questo genere non verrebbe neanche considerato, eppure la teoria dell'accelerazionismo tecnologico sembra essere molto allettante. Le barriere che andrebbero a limitare questa teoria sono numerose. In primo luogo, la scarsità delle risorse naturali sul nostro pianeta. Negli ultimi anni sono diventate evidenti le tensioni geopolitiche in merito al prelevamento di minerali preziosi e la prospettiva di recuperarli su pianeti molto distanti per ora non è ancora allettante a livello economico. In secondo luogo, lo spreco di risorse a livello industriale e l'ossessione per l'innovazione costante stanno rendendo un sistema economico che sembrava efficiente molto inefficace sul piano degli obiettivi di lunga durata. Come già detto in precedenza rispetto alla popolazione italica dei Liguri, quando una popolazione aumenta progressivamente il benessere gran parte di quest'ultimo finisce in maniera caotica nelle mani di chi detiene il potere o chi ha il possesso degli strumenti più adatti. Un fenomeno che si presenta come una delle poche costanti dalla nascita delle grandi civiltà, da Babilonia fino agli stati moderni. Ulteriore problema che si potrebbe presentare nei prossimi decenni risiede proprio nella scienza, laddove la crisi internazionale con relative guerre e povertà potrebbe bloccare o rallentare la cooperazione scientifica mondiale, con il rischio di manipolare l'editoria scientifica nascondendo documentazioni rilevanti o consentire la diffusione di false notizie. Altro caso sempre in merito alla scienza potrebbe essere di natura tecnica, la complessità degli esperimenti scientifici sta diventando sempre più elevata sfidando a volte i nostri stessi limiti di logica. In informatica all'aumentare della complessità di un algoritmo aumenta la difficoltà per il ricercatore di riuscire

³⁶ Yuval Noah Harari, *Sapiens: Da animali a dèi. Breve storia dell'umanità*, Giunti editore, 2017, p. 7

a seguire i passaggi effettuati dalla macchina su un certo calcolo numerico. In fisica quantistica quando si trattano particelle subatomiche la difficoltà maggiore sta nel non interagire con la materia, sostanzialmente ciò che si richiede è di fare esperimenti sulla realtà senza interagire con la realtà. Non ci vorrà molto affinché in numerose discipline gli esperimenti diverranno così tanto complessi che ci saranno barriere all'entrata come costi di ricerca o tecnologie sempre più costose e avanzatissime, sempre che quest'ultime siano in grado effettivamente di reggere la complessità. Di certo "l'accelerazionismo tecnologico" ad oggi non rappresenta un accelerazionismo della conoscenza, basterebbe fare un piccolo esperimento sui modelli linguistici artificiali più in voga scrivendo una domanda come: "Scrivimi dieci domande a cui la scienza non è ancora in grado di rispondere". Il risultato mostrerà come ogni singola domanda non risolta sia in grado di distruggere e alterare interi campi del sapere. Questi sono solo alcune variabili che potrebbero rallentare questa "accelerazione", tenendo conto di aspetti psicologici come il rifiuto della popolazione e dei politici ad innovazioni che non comprendono oppure aspetti economici, laddove si crea una filiera industriale se ne distrugge un'altra con ricadute sull'occupazione. Se dovessimo immaginare un grafico di funzione in merito a questa teoria di certo non sarebbe né esponenziale e neanche logaritmico, probabilmente sarebbero un insieme di curvature che rappresenterebbero un miscuglio tra espansione e arresti rispetto ai balzi tecnologici. Una semplice dimostrazione di come le informazioni costruite non rappresentano il bene o il male, bensì fungono per dare forma alla rappresentazione del mondo e vanno avanti se vengono ritenute rilevanti, anche se per un piccolo periodo.

Sarebbe incorretto dire che non stia avvenendo nulla, poiché il salto evolutivo attuale che l'umanità sta facendo rimane comunque evidente. Per *Data Sapiens* si potrebbe intendere il cambiamento stesso della società e del nostro corpo. La nostra cognizione è sempre più affidata agli strumenti ormai divenuti protesi digitali che vengono chiamati telefono e computer. L'uomo e la macchina divengono co-ideatori del dato che viene creato, elaborato e inviato da una parte all'altra del mondo. L'educazione stessa nella modernità si basa sempre di più su algoritmi di raccomandazione dei siti internet, social o piattaforme varie che, sulla base di modelli predittivi basati sui dati, permettono di consigliare libri e video diversi rispetto al risultato matematico ottenuto in merito alle nostre inclinazioni. In questo senso già esistono politiche educative che vanno a modificare radicalmente la scuola o i percorsi universitari degli studenti basandosi su algoritmi che vanno a predire se uno studente sia in grado di affrontare o meno una materia. I test OCSE-PISA non sono nient'altro che *big data* cartacei e poi digitalizzati che indicano immediatamente su quali discipline gli studenti devono rafforzare le proprie competenze.

L'evoluzione non ha una base morale, non indica se la direzione sia bene o male. Evidenti sono tutti quei fenomeni distorsivi che si vanno a creare in merito al concetto di dato. Immaginiamo,

ad esempio, un ipotetico “*Data Sapiens Moderno*”: un individuo che assimila ed elabora soprattutto le informazioni che gli sono più congeniali o attraenti, come quelle di carattere politico o sociale. Col tempo, questo soggetto rischierebbe di sviluppare una vera e propria dipendenza dai dati stessi. Nel caso non trovasse strumenti o applicazioni in grado di soddisfare costantemente questa sete informativa, potrebbe generarsi in lui una sensazione di frustrazione, un vuoto che lo porterebbe a sentirsi disconnesso rispetto al mondo circostante. Questa disconnessione potrebbe alterare tutto il suo comportamento etologico. Al contrario, all’interno di questa stessa specie potrebbero emergere individui in grado di sviluppare forme di disagio del tutto nuove. Tra queste potremmo immaginare la *datofobia*: una condizione psicologica legata all’eccesso di informazioni e alla disinformazione continua. In un simile scenario, l’individuo si troverebbe sommerso da dati contraddittori o ridondanti, al punto da non riuscire più a elaborare con chiarezza ciò che lo circonda. La realtà diventerebbe confusa e sfuggente, in cui la capacità di comprenderla in modo coerente andrebbe progressivamente compromessa. Inoltre, un *Data Sapiens* poco vigile potrebbe cadere nell’ingenuità di considerare affidabili in modo assoluto i modelli artificiali, i quali spesso trasmettono l’illusione di possedere una conoscenza universale. Questo atteggiamento rischierebbe di generare una sorta di patologia: un’incapacità di accettare l’errore e, di conseguenza, una visione distorta della realtà e dei suoi inevitabili limiti. In futuro l’espressione: “La malattia è certa perché l’ha detto il medico più bravo che abbiamo” potrebbe trasformarsi in: “L’informazione è sicura perché è il miglior modello di cui disponiamo”. In merito a quest’ultima osservazione delle ricadute evidenti sono già presenti nella sociolinguistica da molti anni, aumentati con il fenomeno dell’italiano artificiale. I correttori di Word correggono il nostro italiano alterandolo di conseguenza, ma chi corregge i correttori?³⁷ D’altronde, il linguaggio non è che vibrazione meccanica di onde sonore nell’aria, mentre i testi scritti possono essere analizzati e descritti attraverso rapporti e variazioni tipiche della statistica. In futuro sarà necessario decidere quale tipo di matematica adottare: una logica umana, una logica artificiale, o forse entrambe?

In ultimo, si può intendere la persona come generatore di flussi di dati costanti comprensibili grazie a tecnologie indossabili o dispositivi *IoT*. Un esempio sono i moderni bracciali da fitness, capaci non solo di segnare l’ora, ma anche di monitorare la frequenza cardiaca e rilevare possibili anomalie o disturbi durante il sonno. Oppure la possibilità di registrare la lunghezza di un percorso di corsa, calcolare la velocità in base al tempo impiegato e conservare in memoria tutte le prestazioni precedenti. Forse l’intelligenza artificiale sta spingendo le persone a rivalutare la

³⁷ Un esempio è una varietà linguistica dell’italiano detta italiano artificiale: l’italiano su cui vengono addestrati i modelli linguistici. A causa dell’abuso degli algoritmi sulla scrittura dei libri e degli articoli si sta andando ad alterare il nostro stesso italiano. Pensiamo ad inglesismi non necessari come *feedback* e *ICT*, o ancora l’utilizzo di un italiano stereotipato scolastico tipico dei testi accademici oppure colloquiale elaborato a partire da conversazioni online.

matematica delle grandi misure, considerandola una componente essenziale per comprendere e sopravvivere nel mondo contemporaneo. Un giorno non molto lontano potrebbe essere necessario sviluppare certe facoltà intellettive a scapito di altre. Nulla di nuovo, se pensiamo a quando, in epoche remote, i nostri antenati abbandonarono la postura quadrupede per adottare quella bipede così da poter vedere più lontano.

Capitolo 2

Come costruire un modello del mondo

In ogni epoca per garantire la propria sopravvivenza, gli organismi hanno cercato di elaborare modelli capaci di rappresentare il mondo circostante. Il modo più efficace per farlo è sempre stato quello di ridurre la complessità della realtà traducendola in modi che ne permettessero la manipolazione. Dai primi batteri in grado di sfruttare la luce convertendola in energia chimica fino agli scienziati moderni che si avvalgono di modelli formali (fisici, economici o statistici) per produrre nuovi strumenti o rappresentazioni operative.

La *Data Science* rappresenta oggi una delle forme più avanzate di questo processo: essa si occupa di analizzare e interpretare i dati attraverso strumenti computazionali, fondandosi su teorie quantitative e seguendo spesso un approccio probabilistico. Il presente capitolo propone di indagare questa disciplina attraverso una prospettiva filosofica e scientifica. Tale approccio nasce dall'esigenza di recuperare un metodo d'indagine che affonda le sue radici nell'Antica Grecia, successivamente trasformato con la rivoluzione scientifica e, in larga parte, smarrito a causa della crescente meccanizzazione e specializzazione della scienza contemporanea. Ripercorrere il contributo di alcuni grandi scienziati e filosofi del passato consente infatti di mettere in luce l'importanza di un approccio integrato, capace di coniugare riflessione teorica e pratica scientifica.

Tra le figure importanti di un sapere capace di tenere insieme osservazione scientifica e riflessione filosofica si colloca innanzitutto Alexander von Humboldt, considerato uno dei fondatori della geografia moderna e dell'ecologia. Nelle sue opere egli sviluppò una concezione della natura come dinamica e interconnessa, in cui fenomeni geologici, biologici e climatici sono legati da relazioni reciproche. Questa visione sistemica, fondata sull'osservazione empirica ma guidata da una profonda intuizione teorica, pone una sorte di estetica e l'esigenza di comprendere il mondo naturale come una rete complessa di interdipendenze.

Un analogo intreccio tra scienza e filosofia emerge nella figura di Benjamin Franklin, scienziato, inventore e pensatore centrale dell'Illuminismo americano. I suoi celebri esperimenti sull'elettricità non si limitavano a un interesse tecnico, ma erano inseriti in una più ampia indagine sulle leggi universali della natura. Per Franklin la conoscenza scientifica aveva una chiara valenza etica e comprendere i meccanismi naturali significava anche mettere il sapere al servizio del progresso umano e del bene comune. In questa prospettiva scienza e morale costituivano due dimensioni inseparabili dello stesso progetto illuministico, tendendo a valorizzare esperimenti che fosse concreti e reali diminuendo l'eccessivo formalismo.

René Descartes rappresenta un passaggio cruciale nella costruzione della scienza moderna, poiché tentò di fondare la conoscenza su basi razionali e certe, unificando indagine filosofica e calcolo matematico. Il celebre *cogito ergo sum* inaugura un percorso metodologico fondato sul dubbio sistematico e sull'analisi che influenzerà profondamente il pensiero scientifico successivo. La sua filosofia naturale mirava a comprendere il rapporto tra mente e mondo fisico, tra *res cogitans* e *res extensa*, delineando un quadro unitario della realtà.

Infine, Charles Darwin segnò una svolta decisiva nel modo di concepire la natura e il posto dell'uomo al suo interno. Con l'elaborazione della teoria dell'evoluzione, egli mise in crisi l'idea di una natura statica e finalisticamente ordinata ridimensionando tanto l'antropocentrismo quanto le interpretazioni rigidamente creazionistiche. Attraverso osservazioni empiriche ed esperimenti relativamente semplici, Darwin formulò un nuovo paradigma teorico in grado di inserire l'essere umano nel continuum dei processi naturali e ridefinire radicalmente il rapporto tra scienza, filosofia e visione del mondo.

La scienza mira alla formulazione di teorie sul mondo naturale attraverso l'analisi quantitativa dei fenomeni e l'applicazione del metodo sperimentale. Quest'ultimo si articola in una sequenza rigorosa che comprende l'osservazione empirica, la formulazione di ipotesi, la loro verifica mediante esperimenti controllati e replicabili e, infine, la costruzione di modelli teorici capaci di spiegare e prevedere i fenomeni osservati.³⁸ La filosofia contribuisce allo sviluppo degli strumenti logici e concettuali necessari a un'analisi qualitativa della realtà. Essa non si limita a supportare l'indagine scientifica, ma ne analizza i presupposti interrogandosi sul significato, sulla coerenza e sulla portata interpretativa delle teorie formulate. In questo senso, la riflessione filosofica sollecita lo scienziato non solo a produrre risultati sperimentalmente verificabili, ma anche a chiarire la comprensione dei fenomeni indagati e dei concetti impiegati.

Un esempio emblematico di questa distinzione emerge nell'ambito della *Data Science* con la teoria dell'informazione di Claude Shannon, che stabilisce una relazione formale tra probabilità e quantità di informazione. Eventi altamente probabili veicolano una minore quantità di informazione, mentre eventi rari o improbabili ne veicolano una maggiore. La teoria fornisce una misura rigorosa dell'informazione in termini di riduzione dell'incertezza, ma non affronta il problema della sua natura o del suo significato. Per comprendere questo, un'analogia significativa può essere tratta dalla biologia molecolare. È possibile analizzare quantitativamente la struttura chimica del DNA, così come misurare la frequenza e la distribuzione delle sue componenti, tuttavia, una comprensione più profonda richiede anche un'analisi qualitativa del significato

³⁸ Il metodo sperimentale si sta evolvendo poiché gli esperimenti empirici vengono sostituiti da *software* di simulazione. Per esempio, si potrebbe modellizzare matematicamente comportamenti tipici in biologia evolutiva facendo variare continuamente parametri e iperparametri. I vantaggi sono la velocità dell'esperimento e l'abbattimento dei costi con l'evidente problema che non si sta più osservando la realtà.

funzionale delle sequenze genetiche. Il DNA costituisce infatti il principale mezzo di trasmissione dell'informazione biologica, e attorno a esso si sviluppano teorie fondamentali sull'evoluzione e sull'organizzazione della vita. In modo analogo, nella *Data Science* il concetto di informazione rimane intrinsecamente complesso: pur essendo formalizzabile e quantificabile, esso continua a sollevare questioni teoriche e filosofiche che non trovano una definizione univoca.

Da questo punto in avanti il termine verrà tradotto in italiano con l'intento di distanziarlo dal suo significato originario evoluto dalla *Computer Science* cercando di evidenziare come, all'interno di questa disciplina, si intreccino diversi livelli di relazione che spaziano dalla fisica fino alle scienze cognitive, in particolar modo dati i recenti tentativi di fusione tra computer e sistemi biologici.³⁹ La Scienza dei Dati si occupa del processo di trasformazione dei dati in informazioni rilevanti rispetto ad un atto di osservazione. Questa definizione, per quanto sintetica, mette in luce la questione centrale della disciplina, la costruzione di un modello del mondo comporta limiti di logica, poiché ogni rappresentazione implica una riduzione della complessità. Pertanto, le parole utilizzate in questa definizione dovranno essere prese con la dovuta attenzione poiché consentiranno durante il capitolo di "aprire" il modello logico di questa scienza evidenziando i vantaggi e gli svantaggi, dando vita a numerose domande e mettere il dubbio come elemento centrale durante tutta la trattazione.

2.1 Analisi di un atto di osservazione

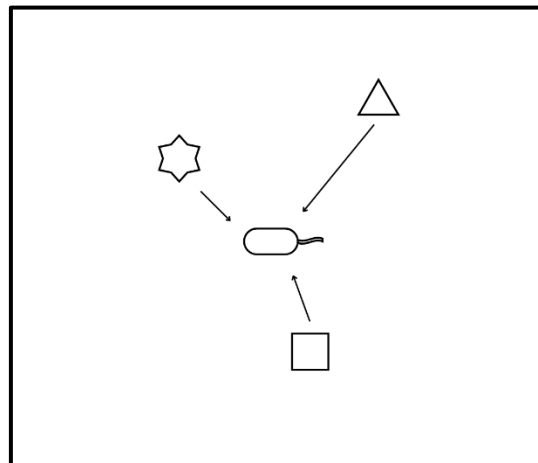
In primo luogo, che cosa significa osservare? Per tentare di trovare dei punti di riferimento bisogna partire da un mondo passato altamente caotico e pericoloso, in cui microrganismi per sopravvivere hanno dovuto sviluppare dei sistemi capaci di interagire con l'ambiente.⁴⁰ Da queste interazioni emerge una forma embrionale di ciò che può essere definito come "osservazione", intesa non in senso cosciente, ma come capacità funzionale di rilevare e rispondere a stimoli ambientali. Si consideri, ad esempio, un microrganismo dotato di capacità di autoreplicazione. La sua sopravvivenza dipende dalla possibilità di accedere a risorse adeguate nel proprio ambiente. Nel caso in cui tali risorse siano disponibili, l'organismo dispone dei processi chimici necessari alla replicazione; in caso contrario, la sua linea evolutiva è destinata a interrompersi. Il processo di selezione naturale favorisce pertanto quei microrganismi capaci di mettere in atto comportamenti che aumentano la probabilità di sopravvivenza, come la capacità di reagire in modo differente alla distribuzione delle molecole utili nell'ambiente. In questo senso, la

³⁹ L'esempio più noto è CL1 della Cortical Labs che combina neuroni coltivati in laboratorio da cellule staminali facendoli crescere su un chip di silicio. Questi esperimenti porteranno a pericolosissime derive etiche. <https://corticallabs.com/cl1.html> (attivo)

⁴⁰ Daniel Bor, *La voracità del cervello: cosa spinge la nostra coscienza verso un'insaziabile ricerca di significato*, Castelvechi, 2012, pp. 51-52

possibilità di “inferire” che determinate risorse siano concentrate in specifiche regioni dello spazio circostante conferisce un vantaggio adattivo. Il microrganismo può decidere tra più “idee” come stabilizzarsi in certi luoghi oppure muoversi ed esplorare l’ambiente.

Figura 4- Rappresentazione concettuale di un microrganismo che ricerca molecole chimiche in un ambiente caotico.



Autore: Lorenzo Tiari

Sebbene nel corso del tempo il microrganismo possa sviluppare una “rappresentazione” dell’ambiente, la natura intrinsecamente dinamica del mondo richiede ulteriori capacità adattive, in particolare la possibilità di rispondere in modo efficace ai cambiamenti ambientali. Una ridotta flessibilità comporta infatti il rischio di non riuscire a individuare nuove risorse chimiche e, di conseguenza, di perdere la capacità di autoreplicazione. In una fase evolutiva primordiale il “cambiamento di idea” non si configura come un atto cognitivo o chimicamente molto complesso, ma come la produzione di copie leggermente differenti di sé. Un processo che realizza un equilibrio tra la conservazione delle strutture funzionali e l’introduzione di variazioni innovative.⁴¹ Queste variazioni comportano inevitabilmente una serie di conseguenze. Le nuove copie possono presentare alterazioni chimiche che, se da un lato introducono potenziali vantaggi adattivi, dall’altro possono generare inefficienze nei processi energetici o la perdita di caratteristiche fondamentali per la sopravvivenza, quello che si definisce “errore”. Non tutte le variazioni risultano adeguate a un ambiente complesso e mutevole. Questo continuo confronto con il mondo esterno dà origine a una tensione strutturale tra stabilità e cambiamento, tra conservazione e innovazione, che rappresenta uno dei problemi fondamentali dell’evoluzione biologica e di ciò che viene definito come informazione. Nei primi microrganismi, questa tensione

⁴¹ Ibidem

risultava particolarmente accentuata a causa dell'elevata instabilità e della scarsa affidabilità dei meccanismi di conservazione e trasmissione dell'informazione.

Con l'emergere del DNA, inteso come sistema altamente efficiente di immagazzinamento e trasmissione dell'informazione biologica, questo equilibrio è diventato progressivamente più gestibile. Le lunghe sequenze genetiche hanno permesso ai microrganismi, sempre più evoluti, di trasmettere "idee positive" o "negative" del mondo con pochissima perdita informativa.

Con l'aumento della complessità dell'interazione con l'ambiente, il paradigma fondamentale rimane la tensione tra conservazione e innovazione. In questo quadro evolutivo, alcuni microrganismi hanno compiuto un significativo salto di complessità sviluppando capacità sempre più sofisticate di interazione e di "analisi" dell'ambiente che hanno consentito un reperimento più efficiente delle risorse chimiche disponibili, sviluppando strutture biologiche più articolate come il sistema nervoso. La capacità del sistema nervoso di coordinare il movimento nello spazio ha favorito lo sviluppo di forme sempre più complesse di rappresentazione e organizzazione dell'esperienza spaziale, rendendo possibile un'interazione più flessibile e adattiva con l'ambiente.

Ciò non implica che questa traiettoria evolutiva sia stata l'unica via al successo. Altri organismi, sulla base di differenti "rappresentazioni" del mondo, hanno adottato strategie alternative prevalentemente sedentarie, come nel caso delle piante e dei funghi, sviluppando soluzioni radicalmente diverse, ma ugualmente efficaci.⁴² Diventa necessario soffermarsi su questo punto, poiché per lungo tempo sia nella *Data Science* che in altre discipline, si è spesso tentato di formalizzare a livello logico caratteristiche chiave dell'intelligenza partendo da organismi con un sistema nervoso, in particolar modo se fortemente centralizzati, quando anche altri regni multicellulari adottano strategie complesse. Per esempio, la pianta erbacea *Arabidopsis thaliana* ha undici tipi di fotorecettori per "osservare" i colori del mondo, mentre l'*Homo Sapiens*, considerando i fotopigmenti, ne possiede solamente quattro.⁴³ Strategie evolutive diverse non portano ai concetti comunemente intesi come "bene" o "male", "superiore" o "inferiore", portano a modelli del mondo diversi.

Possiamo distinguere il processo di percezione sensoriale degli organismi in diverse fasi: la ricezione di dati dal mondo, l'elaborazione dopo averli osservati e infine la costruzione di informazioni interne atte a migliorare le successive azioni.⁴⁴ Gli organismi come piante e animali, in maniera sempre più complessa, hanno incominciato a reperire quantità sempre maggiori di dati con la conseguenza di dover adottare sistemi di filtraggio in grado di rilevare solamente gli stimoli

⁴² Ivi, p. 23

⁴³ D. Hoffman, *L'illusione della realtà: come l'evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020, p 14.

⁴⁴ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, p. 155

più adattivi.⁴⁵ Per gli animali con un sistema nervoso vengono classificati due tipi di filtraggio: la filtrazione periferica e lo stimolo chiave.⁴⁶

Il primo filtro sono i dati che l'animale è effettivamente in grado di ricevere. Questo implica che, se un dato animale possiede un organo sensoriale come l'occhio possiede il vantaggio di utilizzare fotoni e la consistenza fisica dell'oggetto per costruire l'informazione "colore", dall'altra parte un organo come un'elevata percezione sensoriale richiede un discreto dispendio di energia calorica.

Questo è il primo limite logico: un organismo deve decidere se utilizzare organi in grado di aumentare la capacità di ricevere ed elaborare stimoli dal mondo con un determinato costo energetico, oppure deve scegliere di essere meno forte "computazionalmente" risparmiando calorie utili per la sopravvivenza.⁴⁷

Questo limite è lo stesso che si pone uno scienziato dell'intelligenza artificiale quando costruisce modelli artificiali: all'aumentare degli algoritmi, aumenta il dispendio energetico della macchina. Se diventa troppo elevata, l'energia termica rovina i componenti. Questo dilemma va risolto in qualche modo. In genere la tendenza è un controbilanciamento che costringe l'organismo a rimuovere stimoli eccessivi dal mondo, decidendo di dare delle rappresentazioni delle cose che non tendono ad una sorta di "verità", bensì generalizzazioni che servono per interagire con l'ambiente.⁴⁸

Per lungo tempo l'*Homo Sapiens* ha utilizzato la complessità del proprio cervello senza sapere precisamente le sue caratteristiche come se la storia evolutiva avesse preferito un sistema nervoso efficace nell'uso, ma poco trasparente nel suo funzionamento. Questa *black box* o inspiegabilità del sistema nervoso rappresenta per forza un "errore"? Dal punto di vista evolutivo il fatto che un organismo non capisca bene il suo funzionamento interno può portare a vantaggi come rendere le azioni più fluide e non bloccarsi davanti ai pericoli. Quante volte in una singola vita il cervello prova a fare predizioni sul futuro, talvolta sbagliando, bloccando o alterando comportamenti che avrebbero portato ad altre alternative. Questo non deve essere considerato un problema banale della vita quotidiana moderna, sta diventando un punto di svolta per capire se in un mondo abbondantemente informativo la "troppa intelligenza" possa deviarne completamente la sua analisi.

⁴⁵ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, p. 29

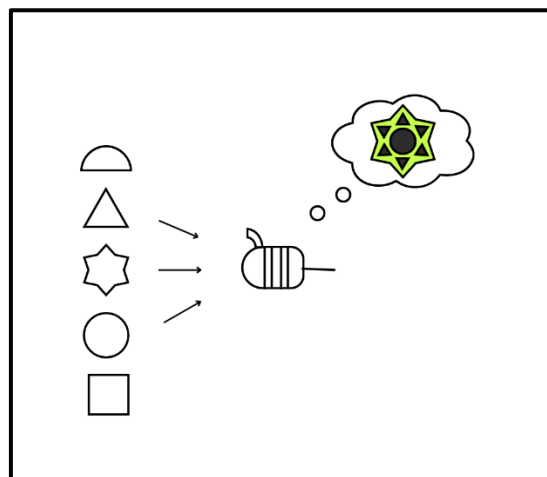
⁴⁶ *Ibidem*

⁴⁷ Il ragionamento è lo stesso che viene applicato ai centri di elaborazione dei dati quando si decide il tipo di raffreddamento, il consumo energetico e così via. La costruzione di un'informazione non è "gratis", richiede un qualche trasferimento di energia.

⁴⁸ D. Hoffman, *L'illusione della realtà: come l'evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020, pp 11-12-13

Il secondo livello di filtraggio dell'informazione è costituito dagli stimoli chiave, ovvero da specifiche caratteristiche dell'ambiente che vengono selettivamente percepite dall'organismo e che innescano risposte comportamentali altrettanto specifiche. Un esempio particolarmente significativo è fornito dallo studio della percezione visiva delle api. Questi insetti presentano capacità percettive che differiscono in modo sostanziale da quelle umane. Il loro sistema visivo è sensibile non solo alle lunghezze d'onda del visibile, ma si estende anche alla regione ultravioletta dello spettro elettromagnetico.⁴⁹ La luce solare contiene abbondante radiazione ultravioletta e molti fiori presentano motivi particolari che si possono notare solo in questo modo, guidando il comportamento da insetti impollinatori e contribuendo alla fertilità delle piante.

Figura 5- Rappresentazione concettuale di un'ape che va verso un fiore illuminato dalla luce solare. La radiazione ultravioletta è una guida per gli insetti impollinatori.



Autore: Lorenzo Tiari

Si pone dunque la questione se un aumento della quantità di stimoli che un sistema è in grado di ricevere dal mondo si traduca automaticamente in un miglioramento della qualità dell'informazione elaborata. L'osservazione dei sistemi biologici suggerisce che la risposta all'interrogativo diventa complessa. Essa dipende dalle modalità attraverso cui gli stimoli vengono selezionati, filtrati e interpretati. Da un lato, il continuo processo di filtraggio sensoriale consente la costruzione di modelli funzionali del mondo capaci di orientare l'azione in modo efficace. Dall'altro lato, tuttavia, l'evoluzione dei sistemi percettivi può introdurre distorsioni sistematiche dando origine a ripetuti "errori". Gli errori non sono necessariamente disfunzionali, ma emergono come conseguenza dell'adattamento a specifici contesti ambientali e a determinate pressioni selettive.

⁴⁹ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, pp. 160-161

Un esempio emblematico è rappresentato dai cosiddetti zimbelli naturali, ovvero caratteristiche di un organismo che fungono da stimoli chiave per un'altra specie. In questi casi, il sistema percettivo dell'osservatore viene sfruttato attraverso segnali che attivano risposte comportamentali automatiche, pur non corrispondendo a una rappresentazione accurata della realtà sottostante. Ciò mostra come un incremento della quantità di stimoli percepiti non garantisca di per sé una migliore informazione, ma richieda strutture interpretative adeguate a evitare fraintendimenti sistematici.⁵⁰ Generalmente il fenomeno più conosciuto è quello del mimetismo sfruttato sia dai predatori che dalle prede. Un caso d'esempio sono le falene *Automeris io* che nascondono sotto il secondo paio di ali delle macchie ocellari, simili ad occhi, che vengono esibite nel caso in cui l'animale venga disturbato.⁵¹ La funzione di queste macchie è provocare una risposta di fuga da parte del predatore grazie a una sorta di "somiglianza" negli stimoli, cercando di mimare gli occhi di un gufo o una civetta. D'altro canto, il predatore ha motivi evolutivi per esibire una risposta di fuga che in seguito sarà dannosa, rispondere allo stimolo chiave o a tutto ciò che gli assomiglia è necessario poiché funziona o è "vero" nella gran parte dei casi. Un altro caso più specifico è il cosiddetto mimetismo batesiano, una strategia di difesa evolutiva innocua che mima comportamenti di specie estremamente pericolose. Le mosche che appartengono a una vasta famiglia d'insetti, detti *Syrphidae*, vanno ad imitare completamente i colori delle api e vespe.

Come si può notare l'evoluzione costringe gli organismi ad una continua analisi di stimoli ricevuti, e li induce a mettere in atto operazioni di filtraggio mantenendo un costante bilanciamento tra conservazione e innovazione delle informazioni costruite in precedenza. L'evoluzione non ha consentito agli animali di vedere un'assoluta "verità"; anzi, in certi contesti, il fatto di vederla diventa perfino pericoloso.⁵² I sensi si sono evoluti affinché gli organismi potessero interfacciarsi in modo veloce nei confronti del mondo e porre azioni necessarie per la sopravvivenza, in caso di pericolo la "verità" risulta poco interessante. Per un *Homo Sapiens* risulta evolutivamente più rilevante la capacità di rappresentare un oggetto come una mela in termini tridimensionali, collocandolo nello spazio e associandone le proprietà percettive come il colore e informazioni quali la commestibilità, piuttosto che la comprensione del comportamento degli elettroni che costituiscono gli atomi della mela stessa. I sistemi percettivi e cognitivi umani si sono evoluti per operare efficacemente nel dominio macroscopico fornendo rappresentazioni del mondo orientate all'azione e alla sopravvivenza, non alla descrizione completa dei livelli fondamentali della realtà fisica.

⁵⁰ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, p. 37

⁵¹ *Ivi*, p. 38

⁵² D. Hoffman, *L'illusione della realtà: come l'evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020, pp. 9-10

Questa asimmetria evolutiva può contribuire a spiegare perché alcune branche della fisica moderna, in particolare la meccanica quantistica e lo sviluppo dell'elettrodinamica quantistica, presentino fenomeni che appaiono profondamente controintuitivi rispetto all'esperienza quotidiana, come il dualismo onda-particella o il comportamento dei fotoni. Tali concetti entrano in tensione con i modelli cognitivi costruiti nel corso dell'evoluzione, i quali presuppongono una realtà composta da oggetti localizzati, dotati di proprietà determinate e governati da relazioni causali lineari.⁵³ Quando si osserva il mondo subatomico, le nostre informazioni costruite lungo l'evoluzione incominciano ad andare in difficoltà intaccando un principio di coerenza e significato di un modello del mondo. Lo stesso problema si ritrova quando si guarda la totalità del mondo macroscopico. La vista non è in grado di cogliere minimamente la complessità dell'universo. Tutta la matematica e i formalismi logici applicati allo spazio e al tempo perdono di significato. Si possono utilizzare convenzioni come gli anni luce per calcolare distanze galattiche, ma a livello percettivo sono misure talmente elevate da risultare "inconsistenti". Il dubbio sollevato rappresenta una questione cruciale nella Scienza dei Dati: fino a che punto possiamo formalizzare un modello del mondo e suddividerlo in dati? Se la suddivisione in dati altera il contenuto stesso dell'informazione, è possibile che debbano coesistere diversi metodi di elaborazione e di conseguenza più teorie?

L'atto di osservare pone questioni che sono state affrontate da svariate filosofie. Per primo si rileva la dialettica opposta tra Illuminismo e Romanticismo, tra Empirismo e Razionalismo. Di conseguenza se si possa intendere a priori o a posteriori ciò che vediamo. L'Illuminismo pone come assunto che è attraverso una fiducia nella ragione umana che si può sconfiggere l'ignoranza nei confronti del mondo, scoprendo il vero significato delle cose.⁵⁴ La realtà è un'esplorazione sociale che possiede regole chiare che vanno scoperte tramite la scienza. Per esempio, Cesare Beccaria rapporta costantemente la matematica con la giustizia morale.

L'esito, dunque, della tortura è un affare di temperamento e di calcolo, che varia in ciascun uomo in proporzione della sua robustezza e della sua sensibilità; tanto che con questo metodo un matematico scioglierebbe meglio che un giudice questo problema: data la forza dei muscoli e la sensibilità delle fibre di un'innocente, trovare il grado di dolore che lo farà confessar reo di un dato delitto.⁵⁵

L'Illuminismo applicato al diritto può essere interpretato come una continua ricerca di proporzionalità morale e razionalità normativa fondata sull'adozione di principi universali e misurabili, analoghi a quelli propri della logica. Questa impostazione filosofica promuove l'idea

⁵³ S. Atzeni e F. Michelotti, *Fisica: meccanica e termodinamica. III edizione*, EdiSES edizioni S.r.l., 2023, p. 377

⁵⁴ C. Beccaria, *Dei delitti e delle pene*, prefazione di S. Rodotà, a cura di A. Burgio, Feltrinelli editore, 2024, p. 6

⁵⁵ Ivi, p. 63

che le informazioni disponibili possano essere ordinate, classificate e sistematizzate secondo criteri formali riducendo l'arbitrarietà e l'opacità delle decisioni.

In continuità con questa visione il collegamento con la Scienza dei Dati risulta particolarmente significativo. Essa presuppone che il mondo sia governato da regolarità logiche e strutture che possono essere individuate attraverso procedure rigorose di raccolta, pulizia, filtraggio e analisi dei dati. Mediante un'analisi metodicamente controllata, diventa possibile "illuminare" modelli del mondo rendendoli progressivamente più chiari, trasparenti e interpretabili. In questo senso, l'approccio illuminista si traduce nel rifiuto di informazioni preordinate o imposte dall'alto favorendo una continua messa in discussione di errori sistematici, distorsioni e assunzioni implicite. Questo ottimismo trova un'espressione concreta nell'adozione dei metodi quantitativi e del metodo scientifico nell'elaborazione dei dati. L'Illuminismo si lega all'Empirismo, secondo cui l'esperienza sensibile costituisce il criterio fondamentale di validità della conoscenza. In questa prospettiva, il sistema nervoso umano viene concepito come una *tabula rasa* progressivamente riempita attraverso l'interazione con il mondo esterno.

In contrapposizione a questo paradigma si collocano altre correnti di pensiero, come il Romanticismo e il Razionalismo, che, pur divergendo tra loro, mettono in discussione l'idea che la conoscenza possa essere ridotta esclusivamente a principi misurabili o da informazioni derivanti dall'esperienza empirica sottolineando rispettivamente il ruolo dell'esperienza soggettiva, dell'immaginazione o delle strutture razionali innate. Il Romanticismo rifiuta l'idea di una realtà dettata dalla ragione ponendo l'accento non più su un collettivismo, ma sull'individualismo e sulla soggettività, mettendo in risalto l'infinito irraggiungibile della natura. In realtà, il Romanticismo si rileva fondamentale poiché viene posto il problema se l'effettivo raggiungimento della conoscenza sia un modo per arrivare alla verità e ad un miglioramento. Per Rousseau la risposta è no.⁵⁶ Rousseau unisce la critica politica ad una sorta di antropologia cercando di mostrare come l'apparente progresso della tecnica e dei costumi non stia provocando un miglioramento, ma al contrario un peggioramento. La storia evolutiva ha costretto gli *Homo Sapiens* a costruire un modello del mondo secondo cui l'informazione "società" ha una valenza "positiva" in termini di adattamento e prosperità provocando errori sistemici. Rousseau identifica nel selvaggio antico una persona più libera, meno soggetta a *bias* e influenze rispetto al mondo naturale.⁵⁷ Rousseau non identifica l'adattamento nel tempo come un miglioramento verso la "verità".

Il primo che, avendo cinto un terreno, pensò di dire "questo è mio" e trovò delle persone abbastanza stupide da credergli, fu il vero fondatore della società civile. Quanti delitti, quante guerre, quanti

⁵⁶ J. J. Rousseau, *Origine della disuguaglianza*, a cura di G. Preti, Feltrinelli editore, 2024, pp. 106-107

⁵⁷ Ivi, p. 72-73

assassini, quante miserie ed errori avrebbe risparmiato al genere umano chi, strappando i piuoli o colmando il fossato, avesse gridato ai suoi simili: “guardatevi dal dare ascolto a questo impostore! Se dimenticate che i frutti sono di tutti e la terra non è di nessuno, siete perduti!” [...] Questa idea [...] non nacque improvvisamente nello spirito umano: fu necessario fare molti progressi, che si acquistassero molte capacità e molti lumi, e questi fossero trasmessi e aumentati da un’epoca all’altra prima che si arrivasse a quest’ultimo confine dello stato di natura.⁵⁸

Per mostrare il pensiero di Rousseau ci si può rivolgere all’etologia in cui un risultato particolare è stato ottenuto con gli esperimenti messi in evidenza da Tinbergen rispetto alla specie d’uccello *Haematopus ostralegus*, definita come beccaccia di mare. Gli esperimenti utilizzavano elementi sostitutivi artificiali simili a quelli ambientali, utilizzando degli stimoli supernormali. La femmina di questa specie, se messa di fronte a poche uova a grandezza naturale oppure ad un singolo uovo artificiale di grandezza notevolmente superiore, tenderà a covare quest’ultimo.⁵⁹ Gli esperimenti con gli stimoli supernormali sono stati ripetuti con varie specie ottenendo di volta in volta risultati simili. Un altro caso d’esempio è la lucciola *Lampryrus noctiluca*, che tende a preferire uno zimbello con una superficie più luminosa e vicina al colore giallo rispetto a quella prodotta dalle femmine conspecifiche.⁶⁰ Questi stimoli supernormali probabilmente hanno valore adattivo, poiché le condizioni che si verificano artificialmente negli esperimenti potrebbe essere viste come “ideali” per i singoli organismi, indicando che le strutture evolutive su cui poggia l’organismo costringono a pressioni che possono indurre a *bias*, quindi ad errori sistemici. Gli animali idealizzano un modello del mondo ottimale che porta a costruire informazioni sbagliate. Peraltro, gli strumenti in grado di creare stimoli supernormali sono onnipresenti nelle società di *Homo Sapiens*, alcuni chiari esempi sono proprio nella comunicazione. Per esempio, nei contenuti pubblicati su alcune piattaforme come *Tik Tok* si fa un notevole uso di uno strumento definito *ring light*, una luce ad anello che si può trovare con diverse dimensioni e caratteristiche sul mercato. Si posiziona davanti al soggetto e la luce circolare crea riflessi luminosi sugli occhi, alterando nel video la percezione del ricevente facendo costruire l’informazione “luccicante” oppure “intenso”. Questo strumento è stato riprogettato più volte consentendo di cambiare il colore della luce oppure la sua intensità per adattarsi alle diverse esigenze comunicative. L’espansione nell’utilizzo di questo strumento è dovuta ad un fenomeno adattivo, ovvero quello che viene comunemente definito col termine “bellezza”, una sorta di metrica di valutazione costante tra organismi. L’occhio, insieme ad altre caratteristiche fisiche, rappresenta un buon indicatore di giovinezza e salute.⁶¹ Gli anelli limbari, il colore dell’iride e la distanza tra i due occhi sul volto consentono al

⁵⁸ J. J. Rousseau, *Origine della disuguaglianza*, a cura di G. Preti, Feltrinelli editore, 2024, p. 72

⁵⁹ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, pp. 38-39

⁶⁰ *Ibidem*

⁶¹ D. Hoffman, *L’illusione della realtà: come l’evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020, pp. 54-55-56

sistema nervoso di costruire l'informazione "attraente". Il Romanticismo pensato da Rousseau si pone in netto contrasto con l'Illuminismo ed è evidente il tentativo di togliere una proprietà assoluta e indipendente nell'osservazione delle cose del mondo.

Il Razionalismo è una corrente filosofica che sostiene, in contrasto all'Empirismo, che la realtà sia governata da principi razionali che il pensiero umano può scoprire tramite deduzioni e idee innate. L'esperienza serve a confermare queste idee preesistenti tramite un ragionamento deduttivo e utilizzando la logica e la matematica. La ragione umana possiede la capacità di arrivare alla verità indipendentemente dai sensi o dall'esperienza partendo da logiche generali per trarre conclusioni particolari. Se un'empirista potrebbe dire una frase come: "la somma di due pecore con altre due pecore forma un gruppo di quattro, ed è vero perché l'esperimento è stato visionato più volte in giornate diverse", un razionalista affermerebbe altre frasi come: "due più due fa quattro per logica, non serve provarlo con l'esperienza".

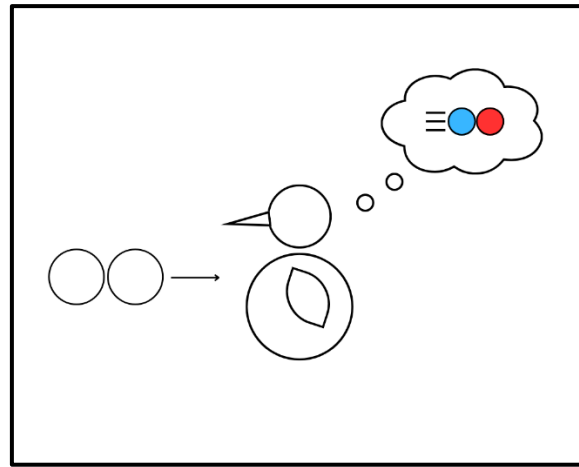
La tensione tra Empirismo e Razionalismo porta alla conseguente domanda: si può parlare di oggettività nell'osservazione in presenza di errori sistemici? La questione diventa capire se l'atto di osservare è semplicemente una scoperta di schemi stabiliti a priori oppure l'analisi del mondo tramite l'esperienza porta ad informazioni nuove. Il neuroscienziato Giorgio Vallortigara, tramite i suoi studi sui pulcini appena nati, consente di effettuare una riflessione su una questione cruciale quando si fa un trattamento di dati.⁶² Tramite i suoi esperimenti, Vallortigara ha scoperto che dopo la schiusa delle uova i pulcini possiedono fin da subito capacità notevoli, tra cui nozioni basilari di fisica, difficilmente giustificabili dall'esperienza ottenuta nei pochi istanti di vita. Secondo il principio d'inerzia di Galileo Galilei, un corpo non soggetto a forza non subisce variazioni di velocità restando in uno stato di quiete, o nel caso di movimento, in un moto rettilineo a velocità costante.⁶³ La distinzione tra oggetti inanimati e animati generalmente è che un potenziale predatore o un membro della propria specie agisce in modo semovente, accelera e decelera con continue variazioni di velocità. I sassi e le altre cose del mondo inanimate, invece, devono essere spinti da qualche forza. I pulcini, come i bambini appena nati, mostrano sensibilità a queste variazioni di velocità nel tempo.⁶⁴

⁶² G. Vallortigara, *Il pulcino di Kant*, Adelphi, 2023, p. 18

⁶³ S. Atzeni e F. Michelotti, *Fisica: meccanica e termodinamica. III edizione*, EdiSES edizioni S.r.l., 2023, p. 37

⁶⁴ G. Vallortigara, *Il pulcino di Kant*, Adelphi, 2023, p. 27

Figura 6- Rappresentazione concettuale di un pulcino che identifica un'entità animata da una inanimata tramite un cambiamento di velocità.



Autore: Lorenzo Tiari

Se i pulcini e i bambini devono decidere quale tra gli oggetti che vedono sia plausibilmente un qualcosa di animato, scelgono l'oggetto che muovendosi in velocità urta il secondo muovendolo. Una possibilità potrebbe essere la ricerca della novità nelle variazioni di velocità anche se meccanismi che sostengono questo genere di sensibilità sono individuati anche nel comportamento dei pesci.⁶⁵ Nel caso ipotetico che vi sia un principio di novità sarebbe necessario andare ad indagare perché questo avviene, poiché il concetto di "novità" per un organismo è comunque un *bias* strutturale perché richiede che il resto del mondo sia "normale". Quando si valuta uno stimolo nuovo un organismo deve eseguire due tipi di analisi. Per primo deve stimare il grado di novità dello stimolo richiamando la memoria, per secondo deve assegnare allo stimolo qualche proprietà per poterlo categorizzare.

Rimane difficile pensare che un sistema nervoso appena nasce sia carta bianca, ovvero che tutto ciò che vediamo avviene tramite l'esperienza quotidiana o a posteriori come vorrebbero gli empiristi. A livello logico, se guardiamo alla biologia evolutiva, per un organismo nato poco tempo dopo un suo antenato non è particolarmente adattivo dimenticare completamente tutti gli errori o i vantaggi passati con il rischio di ripeterli. Sono stati effettuati altri tipi di esperimenti, in particolar modo quando lo stimolo da analizzare per i pulcini è un predatore.⁶⁶ In natura se si avvistano predatori come falchi volare sopra di loro i pulcini si immobilizzano per non essere visti, se il falco si precipita su di loro reagiscono con la fuga. I pulcini nati in condizioni sperimentali quando si utilizzano zimbelli artificiali agiscono in questo modo.⁶⁷ I pulcini, come

⁶⁵ Ivi, pp. 28-29

⁶⁶ Ivi, pp. 44-45-46

⁶⁷ Ibidem

gli altri organismi, non solo fanno un'analisi del mondo, bensì sembrano possedere vere e proprie informazioni precostituite nel cervello, come il concetto di “avvicinamento” o “pericolo”. La stessa cosa avviene negli *Homo Sapiens*, questo principio di sensibilità e categorizzazione si mostra quando siamo più spaventati da particolari tipi di insetti anche se auto e motociclette in strada sono potenzialmente più pericolose, veloci e mortali.⁶⁸

Nel momento in cui un bambino esce dal grembo della madre o un pulcino dall'uovo ci sono innumerevoli cose del mondo che si muovono. L'ambiente si presenta come estremamente caotico, ed è essenziale puntare l'attenzione su alcune forme e schemi di movimento in modo da apprendere velocemente. Se non ci fosse questa sorta di “a priori” per i pulcini le azioni sarebbero terribilmente lente e potrebbero causare la morte in natura dai predatori.⁶⁹ Una sorta di guida innata come quella enunciata dai razionalisti. Questo non è comunque sufficiente per processi di apprendimento, in particolar modo se durante la fase di sviluppo l'ambiente cambia, ritornando nuovamente alla continua tensione tra conservazione e innovazione di “idee” sul mondo. L'a priori può essere valido per alcune caratteristiche, ad esempio la scelta di un partner sessuale che abbia caratteri diversi rispetto all'organismo, ma ci sono alcune caratteristiche che sembrano potersi verificare solo con l'esperienza. Molti primati non umani in presenza del cadavere di un conspecifico mostrano una sorta di curiosità che porta a varie interazioni.⁷⁰ Potrebbe accadere che l'animale lo riconosca subito, come c'è buona probabilità che avvengano azioni di toelettatura, tocchi delicati, colpi e così via. Il primate sta calcolando a livello cerebrale l'animatezza del suo conspecifico e pone azioni differenti a causa dell'ambiguità dei segnali. L'animale morto ha ancora una faccia ed è caldo, ma è anche immobile. Sembrerebbe che in condizioni d'incertezza gli animali faticino a trovare una risposta a priori nel proprio cervello, devono piuttosto fare piccoli esperimenti verificando a posteriori quello che sta succedendo. Nonostante l'origine delle prime informazioni risieda in tempi lunghissimi dell'evoluzione naturale e non in quelli brevi dello sviluppo individuale, gli organismi per apprendere devono fare continuamente esperienza sulla realtà.

Kant è stato tra i primi filosofi ad interrogarsi tra le due posizioni tentando di trovare un ponte tra l'Empirismo e il Razionalismo.⁷¹ In primo luogo, accetta che la conoscenza possa derivare dall'esperienza come gli empiristi, ma afferma che da sola non basta, il cervello possiede delle categorie come spazio e tempo che rendono possibile la conoscenza.⁷²

⁶⁸ Ivi, p. 43

⁶⁹ Ivi, p. 41

⁷⁰ Ivi, p. 59

⁷¹ I. Kant, *Critica della ragion pura*, Editrice Morcelliana, 2012, pp. 21-22

⁷² Ivi, p. 67

Che cosa siano gli oggetti in sé, separati dalla ricettività dei nostri sensi, ci rimane affatto ignoto. Noi non conosciamo nient'altro che il nostro modo di percepirli, che ci è peculiare, e che non è neanche necessario che appartenga ad ogni essere, sebbene appartenga a tutti gli uomini.⁷³

I dilemmi posti da Kant divengono incredibilmente attuali, poiché La Scienza dei Dati non implica solamente la costruzione di algoritmi, consente di porre categorie sul mondo e fare delle inferenze. Un chiaro esempio sono gli assunti statistici nel momento in cui si fa analisi dei dati.⁷⁴ Ad esempio, se volessimo andare ad analizzare le cause del malessere giovanile dovremmo in primo luogo andare ad indicare cosa sia un “giovane”, determinando un intervallo di età stabilito a priori. In seguito, dovremmo stabilire se l'indagine sia totale o campionaria con le rispettive tecniche di campionamento. Poi ci dovremmo occupare se la rilevazione dei dati sia stata presa correttamente senza errori e con quale tipo di strumentazione che introduce distorsioni.⁷⁵ Insomma, i modelli predittivi costruiti tramite statistica e algoritmi non descrivono la “realtà in sé”, piuttosto guidano ad una decisione stabilita da un dataset di partenza. Kant sottopone la ragione pura ad una critica, la Scienza dei Dati dovrebbe sottoporre costantemente i modelli di elaborazione ai rispettivi limiti di logica.

Le macchine non sono una cosa a sé stante rispetto all'evoluzione, ma un prodotto dell'evoluzione. Non solo il *software* è stato pensato sfruttando le conoscenze sull'*Homo Sapiens* come il concetto di neurone artificiale, ma persino l'*hardware* va a rispecchiare un metodo di elaborazione dei dati in ingresso che rappresenta una sorta di gerarchizzazione del sistema nervoso. Di conseguenza, gli stessi limiti di logica sono stati proiettati nell'informatica. Per esempio, pochi sono stati i tentativi iniziali di ispirarsi ad una organizzazione più distribuita come potrebbe essere le relazioni sotterranee tra funghi e piante. Per lungo tempo il mondo vegetale è stato scartato a priori, nonostante gli scienziati non avessero e non abbiano ancora una descrizione formalizzata d'intelligenza accettata da tutti. Nonostante questo, funghi e piante rappresentano vere e proprie *intranet* ambientali in cui l'elemento cardine è la circolarità delle informazioni e delle sostanze nutritive. Eppure come ogni organismo devono sopravvivere, adattarsi ai predatori, scambiare segnali chimici con altre piante per comunicare ed infine procacciarsi il cibo.⁷⁶

Il caso più noto è quello della *Mimosa Pudica*, detta anche pianta sensitiva, che ha suscitato sempre grande interesse per la sua reattività agli stimoli. L'esperimento ideato da Stefano Mancuso prevedeva che le piante messe nei vasetti fossero sottoposte a ripetute cadute da

⁷³ Ivi, p.66

⁷⁴ S. Borra e A. Di Ciaccio, *Statistica: metodologie per le scienze economiche e sociali*, Graw Hill, 2021, pp. 249-250

⁷⁵ Per esempio, un'indagine quantitativa effettuata con sondaggi online non consente di vedere se durante le risposte il soggetto intervistato presenta segnali di ansia o fastidio come sbattere nervosamente il piede per terra.

⁷⁶ Anthony Trewavas, *The foundations of plant intelligence*, The Royal Society, 2017, para. Problems of perception <https://doi.org/10.1098/rsfs.2016.0098> (attivo)

un'altezza di dieci centimetri.⁷⁷ Dopo una serie di ripetizioni le piante iniziavano a non chiudere più le foglie, come se si fossero stancate. Per comprendere l'accaduto negli istanti successivi l'esperimento prevedeva di muoverle non più in verticale, bensì in orizzontale. Appena la direzione cambiava le piante di *Mimosa Pudica* tornavano a reagire allo stimolo chiudendo le foglie, una chiara dimostrazione che le piante riuscivano a distinguere la novità e la pericolosità di un evento. Ampliando l'esperimento Mancuso ha deciso di addestrare centinaia di piantine agli stessi stimoli per verificare la durata della memoria, il test ha indicato un tempo superiore di quaranta giorni, molto più lungo della memoria di molti insetti.⁷⁸

Le piante possiedono altre caratteristiche notevoli, tra cui il mimetismo citato in precedenza per gli animali. La *Boquilla Trifoliata* raggiunge capacità di mimetismo raffinatissime, di gran lunga superiori a molti animali.⁷⁹ Per esempio, alcune osservazioni effettuate in Cile mostrano come la pianta imiti ogni volta con grande abilità l'altra pianta su cui cresce, un fenomeno definito come polimorfismo mimetico. La sua capacità si mostra ancora più adattabile se cresce in prossimità di due o più differenti specie. Difatti una singola pianta è in grado di modulare le proprie foglie in maniera diversa e specifica per ogni singola esigenza mimetica. Oltre a questo, il suo mimetismo va a modificare caratteristiche differenti come la forma, il colore e la dimensione della foglia.⁸⁰

Non solo l'apprendimento, il mimetismo e la memoria. Ulteriore caratteristica del mondo vegetale è la manipolazione complessa nell'ambiente. Un ulteriore caso sono quelle specie arboree appartenenti al genere *Acacia*, native dell'Africa o dell'America latina, che utilizzano le formiche come forma di difesa.⁸¹ In primo luogo, la pianta sviluppa corpi fruttiferi destinati all'alimentazione delle formiche e cavità idonee a ospitarle, consentendo loro di stabilirsi, riprodursi e allevare le larve. In un rapporto di mutua assistenza, le formiche forniscono in cambio un'efficace difesa contro i potenziali predatori della pianta. Ciò che rende questo sistema particolarmente rilevante è l'elevato grado di aggressività con cui la difesa viene esercitata. Le formiche non si limitano infatti a respingere piccoli insetti o a ostacolare la crescita di piante concorrenti nelle immediate vicinanze, ma attaccano anche erbivori di grandi dimensioni, come elefanti e giraffe, contribuendo alla formazione di vere e proprie aree circolari prive di vegetazione attorno agli esemplari di *Acacia*. L'origine di questo comportamento estremamente aggressivo è riconducibile alla produzione da parte della pianta di nettare extrafiore: una secrezione ad alto contenuto zuccherino arricchita da numerosi

⁷⁷ S. Mancuso, *Plant Revolution*, Giunti editore, 2021, pp. 26-27

⁷⁸ Ivi, p. 28

⁷⁹ Ivi, p. 64

⁸⁰ Ivi, p. 66

⁸¹ Ivi, pp. 118-119

composti chimici biologicamente attivi. Tra questi figurano sostanze in grado di modulare il sistema nervoso degli insetti come la taurina. Attraverso la regolazione della quantità e della composizione del nettare la pianta è in grado di influenzare il comportamento delle formiche, inducendo forme di dipendenza e alterandone le capacità cognitive in modo funzionale alla propria protezione.

Altra caratteristica è la capacità di schemi riconoscimento che, in alcune piante, avviene quando le radici devono prendere le sostanze nutritive dal terreno.⁸² Alcune piante possono distinguere le proprie radici da quelle di un'altra specie consentendo una crescita modulare e cambiando tipo di architettura. Possono aumentare l'estensione laterale o dirigere le radici verso zone del suolo meno occupate.⁸³ In questo caso la pianta studiata è nuovamente la *Arabidopsis Thaliana*, a cui seguono ancora studi per capire come è in grado di riconoscere e modulare in modo così complesso le radici. Inoltre, l'apparato radicale torna utile per capire un argomento centrale, ovvero una struttura distribuita e decentralizzata al contrario del sistema nervoso. Le radici hanno un'utilità fondamentale per le piante perché sono in grado di reagire singolarmente a modifiche ambientali, interagire tra loro e con altre piante ottimizzando la ricerca di nutrienti nel suolo e aumentando le probabilità di sopravvivenza. Pur non avendo un sistema di comando centrale, l'apparato radicale si presenta come ben coordinato e preciso.⁸⁴ I modelli con cui ricerca risorse nutritive nel terreno possono essere paragonati ai modelli adottati dagli insetti sociali. Una sorta di grande matrice integrata a cui ogni agente autonomo analizza variazioni ambientali consentendo un comportamento emergente. Una singola radice fa poco, ma l'interazione chimica continua di molte radici consentono di prendere continue decisioni ottimali e, soprattutto, molto complesse.⁸⁵

Una questione teorica centrale riguarda la possibilità di attribuire a una macchina un atto di osservazione. L'interazione delle macchine con il mondo avviene attraverso input ambientali specifici spesso mediati dall'intervento umano: sensori, dispositivi di acquisizione o, più direttamente, l'azione della mano umana. L'energia elettrica che attraversa i componenti rende possibile l'elaborazione dei segnali e la produzione di calcoli che operano su rappresentazioni della realtà. Tuttavia, ciò che rimane problematico è stabilire chi, in ultima istanza, possa essere considerato l'osservatore effettivo. Si consideri, ad esempio, il caso in cui un soggetto umano raccolga dati attraverso questionari cartacei, li digitalizzi, li categorizzi mediante un *software* di calcolo e infine li sottoponga a un algoritmo di *machine learning*. In un processo

⁸² Anthony Trewavas, *The foundations of plant intelligence*, The Royal Society, 2017, para. Game Theory and competition. <https://doi.org/10.1098/rsfs.2016.0098> (attivo)

⁸³ Ibidem

⁸⁴ S. Mancuso, *Plant Revolution*, Giunti editore, 2021, pp. 154-155

⁸⁵ Ibidem

di questo tipo, diventa legittimo interrogarsi su quale fase possa essere identificata come l'atto osservativo vero e proprio. L'osservazione è da intendersi come un'attività distribuita e pluralistica, che si estende lungo l'intera catena di acquisizione ed elaborazione dei dati, oppure ogni passaggio costituisce un atto osservativo autonomo, dotato di propri errori e vantaggi sistemici?

Come già evidenziato, l'osservazione comporta inevitabilmente una selezione, escludendo alcune dimensioni del reale a favore di altre. Di conseguenza, l'osservazione non può mai essere considerata un'operazione neutrale, ma risulta piuttosto un insieme di scelte. A tal proposito, è utile ricordare che, nella fase iniziale di sviluppo della *Computer Science* e successivamente della *Data Science*, l'elaborazione dei dati era fondata prevalentemente su algoritmi esplicitamente programmati, come quelli sviluppati in linguaggi quali ALGOL. In quel contesto il *machine learning* non occupava ancora una posizione centrale e il ruolo del programmatore risultava determinante nella definizione dei criteri di elaborazione e interpretazione dei dati. Questo aspetto mette in luce come l'atto osservativo fosse chiaramente riconducibile all'intervento umano mentre nei sistemi contemporanei in cui il risultato di un modello artificiale è sempre più indeterminabile, l'attribuzione risulta più problematica e concettualmente ambigua.

Accade nelle discipline scientifiche che si possano utilizzare delle convenzioni, che non rappresentano totalmente la realtà, ma servono a stabilire dei criteri per procedere inizialmente e poi migliorare in seguito. Si consideri, a titolo ipotetico, l'adozione di una definizione convenzionale di osservatore fondata su alcune caratteristiche minime: la limitatezza delle informazioni accessibili, l'esistenza di un sistema di conservazione di informazioni, la presenza di scambi energetici con l'ambiente e una struttura materiale di base che ne renda possibile il funzionamento. Sebbene una simile definizione possa apparire inizialmente operativa, essa solleva una serie di problemi teorici non trascurabili. In primo luogo, i criteri risultano eccessivamente inclusivi e non consentono di tracciare un confine netto tra ciò che può essere considerato un osservatore e ciò che non lo è. Sistemi fisici come un sasso o un cristallo soddisfano infatti alcuni di questi requisiti: possiedono una struttura materiale, scambiano energia con l'ambiente e, in senso lato, possono "conservare" informazioni sotto forma di stati fisici (ad esempio, tracce di usura o configurazioni strutturali). Analogamente, entità biologiche ai limiti della vita come i virus presentano meccanismi basilari di conservazione dell'informazione e interazione energetica. In secondo luogo, la definizione proposta non distingue tra una mera registrazione passiva di stati del mondo e un'attività osservativa, intesa come selezione, interpretazione e uso dell'informazione. L'assenza di un criterio che introduca una dimensione di rilevanza o orientamento all'azione rende

problematico attribuire lo statuto di osservatore a un sistema piuttosto che a un altro. Infine, questa convenzione rischia di svuotare il concetto stesso di osservazione del suo significato riducendolo a una generica interazione fisica con l'ambiente. Se ogni sistema che scambia energia e conserva tracce può essere considerato un osservatore, allora l'osservazione perde la sua specificità come atto di costruzione di conoscenza o di modellizzazione del mondo. Di conseguenza, la difficoltà principale di questa definizione risiede non tanto nell'identificazione di caratteristiche necessarie, quanto nella mancanza di criteri sufficienti per escludere casi limite e per distinguere tra sistemi che semplicemente reagiscono al mondo e sistemi che, in qualche senso, lo osservano. Questa ambiguità suggerisce la necessità di una definizione più articolata, senza scivolare in un'eccessiva generalizzazione.

La risposta potrebbe essere trovata nei prossimi decenni dall'intersezione di discipline cognitive e informatiche, tenendo anche conto di branche particolari che si stanno sviluppando come la metallorganica, che studia l'interazione chimica del carbonio (atomo necessario per la vita sulla terra) insieme all'unione di metalli e semimetalli. Forse con lo studio più preciso dei materiali potremo un giorno dare una definizione di osservazione più chiara.

2.2 L'inversione dei valori nei dati

Il filosofo tedesco Nietzsche per gran parte della sua vita sarà contro la visione moderna del mondo, dando vita a quella corrente filosofica chiamata Immoralismo. La prima cosa che nota nella società è che ogni volta che si tenta di dare una morale, una sorta di modello idealistico del mondo, ci si scontra con dei limiti di logica.

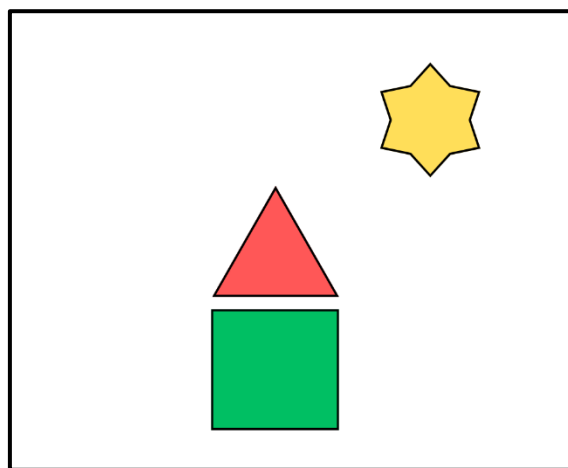
Da qualsiasi punto di vista della filosofia si voglia collocarsi oggi: da ogni posizione si può vedere l'erroneità del mondo, nel quale crediamo di vivere, come la cosa più sicura e stabile di cui il nostro occhio possa impadronirsi: troviamo in proposito motivi su motivi che potrebbero indurci a supposizioni su un principio ingannevole "nell'essenza delle cose".⁸⁶

Per esempio, una comunità di *Homo Sapiens* deve massimizzare il benessere gestendo le risorse scarse in un ambiente naturale. Se la comunità sceglie la libertà, in cui un singolo individuo ottiene tutto ciò che trova, quanta più ne ha quell'individuo meno ne avranno gli altri perché ci si scontra con i limiti naturali. Allora si potrebbe scegliere una redistribuzione equa, un'eguaglianza che per il solo fatto di esistere ha ridotto la massimizzazione di risorse dell'individuo. Nietzsche pensava alla realtà come ad una continua interpretazione, non al raggiungimento di una verità certa, ponendo nella metafora del cammello, del leone e del bambino la caratteristica principale del suo pensiero: l'inversione dei valori. Il cammello gravato di valori come "bene" o "male" costruiti in

⁸⁶ F. Nietzsche, *Al di là del bene e del male*, traduzione di S. Mati e O. Abu Dbei, Feltrinelli Editore, p. 43

precedenza vaga in un deserto morale non trovando nulla. Il cammello si trasforma in leone quando capisce che ci sono degli errori sistemici in quello che gli è stato detto e tenta di trovare nuovi valori.⁸⁷ Infine l'ultima trasformazione è il bambino, un riferimento al mondo orientale e in particolar modo al Taoismo di Lao Tzu, in cui la conoscenza richiede adattabilità mentale in cui nessun modello è assoluto.⁸⁸ Nel momento in cui formula la metafora, Nietzsche ha ben chiaro gli sviluppi delle scienze naturali, di cui era particolarmente interessato, e alla formulazione della teoria dell'evoluzione di Darwin in cui l'oltreuomo rappresenta il nuovo stadio evolutivo.⁸⁹ Come si è visto in precedenza, per adattarsi all'ambiente gli organismi hanno dovuto fare delle scelte in termini di energia formalizzando il mondo secondo delle "idee". Nietzsche riconosce che ogni azione nel mondo implica necessariamente una scelta, e che la scelta si fonda sempre su ciò che è stato appreso in precedenza. Non esiste secondo il filosofo tedesco un agire "puro" o immediato. Cosa si può intendere allora per dato? Si può intendere un dato come una rappresentazione formalizzata della realtà, è l'unità minima potenzialmente significativa. Per esempio, se si disegna su una pagina di un quaderno si può costruire l'informazione "casa" attraverso tre dati. Ogni dato rappresenta un'unità minima attraverso un concetto, una forma geometrica, un colore e posizione permettendo una scomposizione logica della realtà. Possiamo definire l'eguaglianza come: casa = muro + tetto + sole.

Figura 7- Un muro, un tetto e il sole sono posti come tre dati formalizzati in grado di costruire l'informazione "Casa".



Autore: Lorenzo Tiari

⁸⁷ F: Nietzsche, *Il così parlò Zarathustra*, a cura di A. Larsen e tradotto da R. Giani, War Wave editore, 2021, pp. 47-48-49

⁸⁸ Lao Tzu, *Tao te Ching*, Mondadori editore, a cura di B. Walker, 2016, p. 84

⁸⁹ La traduzione di *Übermensch* può riferirsi anche a superuomo. Si è scelto l'altro termine per un discostamento semantico utilizzato dal nazismo inteso come l'uomo innalzato a divinità. Al contrario, Nietzsche ha ben chiaro durante gli ultimi anni di vita che l'umanità sta facendo progressivamente un salto evolutivo e la scienza viene vista come espressione della volontà di potenza.

Il primo problema riscontrabile è che la costruzione di un dato non rispecchia un principio di oggettività.⁹⁰ Nella scienza l'oggettività è richiesta come elemento cardine del metodo scientifico secondo cui la ripetibilità dell'esperimento consente di mostrare che non sussiste una relazione soggettiva tra lo scienziato e il fenomeno analizzato. Secondo la filosofa giuridica Antoinette Rouvroy i dati non devono essere intesi come emanazioni spontanee della natura.⁹¹ L'espansione delle tecnologie ha consentito la ricezione di enormi quantitativi di dati negli ultimi due decenni, con l'impressione che si possano prevedere fenomeni senza alcuna categorizzazione a priori.⁹² Il compito di un *data scientist* consiste in una loro pulizia dai conseguenti problemi derivanti dalla soppressione di significato. Nel momento in cui si lavora con grandi quantitativi di dati vengono utilizzati dei metodi di elaborazione (linguaggi di programmazione, *software* generici, dispositivi tecnologici) in grado di porre una lavorazione trasformandoli in oggetti perfettamente calcolabili. In questo ciclo di filtri e preparazione, grandi quantità di dati si traducono in entità che sono già state mediate dallo scienziato, riflettendo i metodi di preparazione utilizzati. I dati cessano nel momento in cui vengono elaborati di essere traccia precisa di una realtà, ma il prodotto dettato da un'osservazione. Secondo questa prospettiva la Scienza dei Dati non si limita a rilevare fenomeni, ma li reinterpreta in forme compatibili con l'elaborazione automatizzata. Come nota Rouvroy, questo processo compromette le sfumature e le ambiguità del mondo, essendo ridotte a ciò che è computabile. Un'ambiguità notata anche dal filosofo Luciano Floridi, il quale s'interroga sulla mancanza di uniformità dei dati, poiché la costruzione di un dato richiede sempre una differenza.⁹³ Riprendendo il caso precedente, nel disegno per formalizzare l'informazione "casa" ho declinato il singolo dato ad un concetto, una posizione, un colore e una forma geometrica a due dimensioni. Si sarebbero potute fare altre operazioni di pulizia, ad esempio rimuovendo il colore.

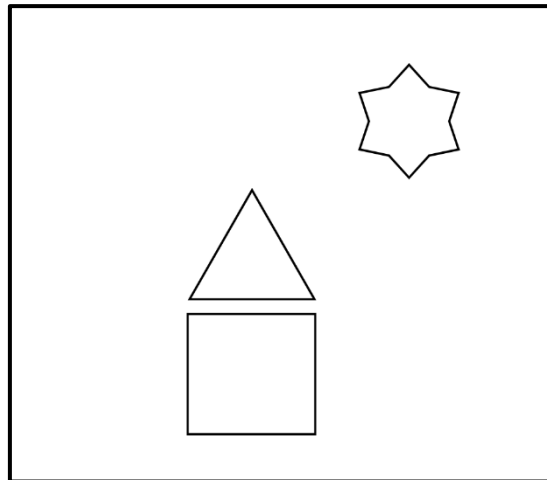
⁹⁰ A. Rouvroy e B. Stiegler, *Il regime di verità digitale. Dalla governamentalità algoritmica a un nuovo Stato di diritto*, La Deleuziana: rivista online di filosofia, 2016, p. 8
<https://ladeleuziana.org/wp-content/uploads/2016/11/Rouvroy-Stiegler.pdf> (attivo)

⁹¹ Ibidem

⁹² Ivi, p 9

⁹³ L. Floridi, *La filosofia dell'informazione*, Raffaello Cortina Editore, 2024, pp. 85-86

Figura 8- La rimozione del colore nei tre dati formalizzati riduce il significato dell'informazione "casa".



Autore: Lorenzo Tiari

Questa operazione di eliminazione delle differenze produce una riduzione di significato nei dati. Andando avanti avremmo potuto cancellare tutto, perfino le linee geometriche lasciando la pagina in bianco, che per il solo fatto di esistere indica che ci si può scrivere qualcosa oppure potrebbe essere che quello che è già stato scritto è stato cancellato. Una completa cancellazione dei dati in un modello del mondo implica l'eliminazione di tutte le differenze, perfino dell'esistenza della pagina bianca.⁹⁴ La produzione di dati presuppone una capacità di un "osservatore" di riconoscere coerenze e differenze all'interno della realtà percepita. Una differenza diventa dato solo quando, a livello macroscopico, consente di distinguere un elemento da un altro. Immaginiamo un osservatore (O) in un certo stato chimico o fisico (s) che guarda un foglio su cui è disegnata l'informazione "casa" (i). Il problema principale nella teoria dell'informazione è trovare un sistema univoco che consenta di misurare quantitativamente e qualitativamente l'informazione. Tentiamo forzatamente (sbagliando) di unire la logica di Floridi con quella di Shannon.⁹⁵

$$O(s, i) \Rightarrow H_s(t_0) > H_s(t_1)$$

Se consideriamo l'entropia in bit (H) come misura dell'incertezza dell'osservatore, allora il passaggio da un tempo iniziale (t_0) a un tempo successivo (t_1) dovrebbe mostrare una diminuzione dell'incertezza quando un dato è effettivamente acquisito. Per un osservatore, la presenza di una differenza riconoscibile nel mondo percepito implica un aggiornamento del suo

⁹⁴ Ivi, pp. 85-86

⁹⁵ Sebbene esistano numerosissime teorie sull'informazione, i due autori sono stati presi perché rispecchiano due modi diversi di ragionare: il primo sul piano semantico e di logica formale, il secondo sul piano probabilistico e di logica matematica.

stato. Se tale differenza permette all'osservatore di determinare qualcosa con maggiore certezza, allora si può dire che c'è stata una diminuzione di entropia. Facciamo un esempio concreto: un osservatore guarda il disegno di una casa e deve determinare come il "sole" è effettivamente rappresentato. Se l'osservatore ha inizialmente tre possibilità equiprobabili per il colore del sole (viola, giallo, blu, ciascuno con probabilità del 33,3 per cento) prima di ricevere i dati, l'entropia di Shannon è circa di 1,585 bit. Se, osservando il disegno, il sole appare effettivamente giallo e l'osservatore li riceve, l'incertezza cade a 0 bit perché la sua previsione ora coincide con l'osservazione.

In termini di Shannon questa riduzione di entropia rappresenta l'informazione acquisita. L'esempio del dato "sole" funziona solo per Shannon poiché per Floridi le probabilità indicano solo un mezzo di calcolo non la natura stessa dei dati. La teoria dell'informazione di Shannon rappresenta una svolta cruciale poiché formalizza logicamente una misura quantitativa che è necessaria per una disciplina scientifica, ma rinuncia completamente a definire cosa sia un dato o un'informazione. Nel calcolo dell'entropia due messaggi con significati completamente opposti possono contenere la stessa quantità di informazione se hanno la stessa distribuzione probabilistica; infatti, per il nostro cervello "giallo" o "viola" hanno frequenze elettromagnetiche diverse. Inoltre, i *Big Data* secondo la teoria possono contenere grande quantità d'informazione nonostante non possano significare quasi niente per un "osservatore".

La formalizzazione di Shannon identifica una struttura statica del canale di trasmissione che è unico per gli "osservatori", ma come si è visto la realtà è enormemente più complessa ed esistono molti modi per comunicare, diversi modi per recepire gli stimoli e innumerevoli metriche di validazione. Infine, non vengono neanche considerati concetti come rilevanza, falsità e utilità tra il mittente e il ricevente. La teoria è potentissima e fondamentale, ma con modelli artificiali sempre più avanzati o addirittura mischiati con materiali organici incomincia a risultare incompleta.

Una caratteristica che può accomunare sia il valore quantitativo sia quello qualitativo dell'informazione a livello macroscopico è il principio di differenza. Insomma, che siano probabilità o differenze semantiche rispetto ad un modello del mondo, un dato esiste solo se si può distinguere da qualcos'altro. L'osservatore (*O*) riconosce che nel disegno una "casa" (*i*) se sono presenti elementi differenti fra di loro ad esempio un muro, un tetto e un sole, indicati con una variabile ($d1, d2, d3$). La differenza fra i tre dati costruiti in precedenza (Δd) consente di costruire l'informazione "casa".⁹⁶

⁹⁶ Il modello RGB (Red, Green, Blue) è fondamentale per la visualizzazione dei colori sui dispositivi digitali. Poniamo un giallo (255,255,0) e un rosso (255,0,0), la differenza RGB tra i due consente di determinare l'informazione "colore". Se si usa un qualsiasi convertitore in rete si noterà l'estrema difficoltà a trovare la differenza tra il rosso (255,0,0) e un rosso meno puro (255,0,1), quasi che l'informazione risulti "insignificante".

$$O(s, i) \Rightarrow \Delta d$$

Per costruire una nuova teoria dell'informazione capace di integrare calcolo numerico e semantica, sarà necessario partire da presupposti che superino i limiti della teoria di Shannon senza rinunciare alla precisione matematica. Una teoria dell'informazione che integri calcolo e semantica deve riconoscere che l'informazione è valida nella misura in cui consente all'osservatore di orientarsi nel mondo. L'informazione non coincide quindi con la semplice riduzione dell'incertezza o probabilità, ma con la capacità di sostenere inferenze, decisioni e azioni nel tempo.

In ultimo, è essenziale distinguere tra dato e informazione. Rispetto alla definizione enunciata nel paragrafo si potrebbe sostenere che un dato possa anche essere visto come una configurazione fisica indipendente da qualsiasi interpretazione. L'informazione emerge solo nel momento in cui un dato viene elaborato da un osservatore. In assenza di un osservatore, il dato rimane una semplice traccia del reale priva di significato informativo. In questa prospettiva, l'informazione non è una proprietà intrinseca del dato, ma una relazione tra il dato stesso e l'osservatore che lo interpreta. Ciò implica che osservatori differenti, dotati di strutture, capacità e obiettivi diversi, possano estrarre informazioni differenti a partire dagli stessi dati. Oltre al fatto che anche questa definizione implica assunzioni troppo forti rispetto alle capacità scientifiche attuali, il problema rimane lo stesso. Se notiamo il muro, il tetto e il sole vengono presentati come dati ben formalizzati con proprietà stabilite a priori. Se ci si riferisce all'informazione intesa nel senso di informare e dare forma ad un modello del mondo, allora si nota una fortissima connessione fra un dato e informazione. Diventa perfettamente comprensibile nel linguaggio quotidiano la continua interscambiabilità dei termini, di cui è divenuta evidente già nella trattazione del secondo capitolo rispetto al primo. Vigge attualmente un principio di indeterminatezza fra cosa sia un dato oppure un'informazione. Forse la Scienza dei Dati si presenta più come una Scienza delle Informazioni.

2.3 La validazione dell'informazione

Che cosa si può intendere per informazione? L'informazione nasce dall'organizzazione e l'interpretazione dei dati in un contesto. Essa rappresenta il passaggio da una rilevazione a un contenuto dotato di senso consentendo di "dare forma" ad un modello del mondo. Parlando dei modelli artificiali ogni dispositivo incorpora dei limiti tecnici, errori di misurazione o soglie di filtraggio. Un problema tipico nell'analisi dei dati è cosa può essere considerato "corretto" da inserire nel dataset, può esistere un filtro di verità delle informazioni? Nella costruzione di una logica dell'informazione, Floridi pone la questione di un possibile filtro di verità.⁹⁷ Proviamo a

⁹⁷ L. Floridi, *La filosofia dell'informazione*, Raffaello Cortina Editore, 2024, pp. 97-98-99-100-101

fare un esempio guardando alle scienze cognitive e alla medicina, in cui il canale comunicativo tra un mittente e ricevente di un'informazione è quasi lo stesso sia nel tempo che nello spazio. Per capire questo bisogna si può utilizzare il caso tipico dei pazienti che vengono chiamati *split-brain*. Questi pazienti, a causa di qualche evento, hanno un taglio al corpo calloso che sarebbe un fascio di assoni in grado di connettere i due emisferi cerebrali.⁹⁸ Gli esperimenti del neuroscienziato Roger Sperry tra gli anni '60 e '70 del secolo scorso hanno mostrato come l'informazione si possa alterare drasticamente nel sistema nervoso di un paziente.⁹⁹ Posti davanti ad uno schermo, il proiettore mostra brevemente immagini e parole nel campo visivo destro o sinistro. Ad esempio, quando l'informazione è nel campo visivo sinistro raggiunge in seguito l'emisfero destro e in una persona normale l'informazione è poi disponibile ad entrambi gli emisferi in un periodo di tempo brevissimo. In un paziente *split-brain* l'informazione non viene condivisa, rimane solo nell'emisfero controlaterale. Se si mostra l'immagine di un cucchiaio nel campo visivo sinistro l'informazione "cucchiaio" va verso l'emisfero destro, ma non viene distribuita nell'emisfero sinistro dove è presente l'area di Broca, zona neurale fondamentale per la produzione del linguaggio.¹⁰⁰ Il paziente possiede l'informazione "cucchiaio", ma potrebbe trasmettere al ricevente (in questo caso il ricercatore che osserva l'esperimento) l'informazione "non lo so". A seguito di numerosi esperimenti, le risposte dei pazienti si sono dimostrate molto variabili fino ad arrivare a dare informazioni false.¹⁰¹ Il fenomeno potrebbe essere connesso a quello che Sigmund Freud chiamava difese psicologiche, strategie in grado di sopperire a quei fenomeni che possono provocare ansia o disturbano l'organismo in qualche modo. Si può tentare di formalizzare la questione. Seguendo la logica precedente si definisce l'osservatore (O) in uno stato chimico e fisico in un determinato tempo (S_t) vede una differenza nei dati (Δd) che riceve dallo schermo producendo l'informazione "cucchiaio" (i).

$$O(s_t) \xrightarrow{\Delta d} i$$

Nei pazienti *split-brain* l'informazione dal campo visivo sinistro arriva all'emisfero destro che non la comunica con l'emisfero sinistro, ponendo due informazioni parziali nello stesso organismo (O).

$$O(s_t) \xrightarrow{\Delta d} i_1 \quad O(s_t) \xrightarrow{\Delta d} i_2$$

⁹⁸ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, pp. 247-248

⁹⁹ *Ibidem*

¹⁰⁰ Bisogna specificare che la lateralizzazione delle funzioni dei due emisferi sono variabili tra soggetti mancini o destrimani, nonostante questo gli esperimenti hanno mostrato che è certo che gli emisferi possono specializzarsi nelle mansioni oppure integrarsi a vicenda. Di conseguenza, un taglio del corpo calloso generalmente è un danno gravissimo nel trasferimento dell'informazione in un sistema nervoso.

¹⁰¹ *Ivi*, pp. 249-250

La particolarità è che l'emisfero destro riceve l'informazione corretta, ma quando questa informazione deve essere verbalizzata essa risulta negata o alterata. Nel momento in cui la verbalizzazione avviene a carico dell'emisfero sinistro, il ricercatore (il secondo osservatore) riceve una ricostruzione diversa dal mittente (i').

$$O_2(s_t) = i'$$

Se la struttura di un osservatore è particolarmente complessa può accadere che le sue parti che lo compongono non siano in grado scambiarsi l'informazione, allora la produzione di informazione risulta incoerente e una parte dell'osservatore può produrre come "falsa" o "incoerente" qualcosa che l'altra percepisce come "vera" o "coerente". L'informazione non è solo ciò che viene osservato, ma anche ciò che può essere integrato dalla sua struttura.

Questo principio non è così distante concettualmente rispetto ai processi chimici e fisici nell'informatica. Nell'*hardware* di un computer classico, ad esempio, l'elaborazione dell'informazione dipende da catene di processi come flussi di elettroni nei semiconduttori, variazioni di potenziale, stati energetici discreti che devono rimanere entro specifiche soglie di stabilità. Quando queste condizioni vengono alterate a causa di rumore termico, interferenze elettromagnetiche o difetti materiali nei transistor, parti diverse del sistema possono trovarsi in stati logicamente incompatibili, producendo errori di calcolo o risultati incoerenti. Il computer potrebbe avere *driver* e componenti di memoria funzionanti per ricevere il messaggio ricevuto in rete, ma non è in grado di comunicarlo all'informatico (il secondo osservatore) a causa di un'alterazione della CPU e GPU. Forse la "verità" informazionale non è una proprietà assoluta, bensì una funzione della struttura e dell'organizzazione materiale del sistema che li elabora. Per questo che ritengo che la Scienza dei Dati ha grande interesse nel rivolgersi alle scienze cognitive. Ulteriori progressi futuri possono rappresentare i primi punti di partenza per formalizzare questioni a livello logico e risponde a numerose domande. Che cosa può essere definito come coscienza? Se esiste e ha qualità fenomenica, qual è il livello minimo necessario per raggiungerlo?¹⁰² Si può definire una disuguaglianza tra intelligente e cosciente consentendo di formalizzarle a livello di logica matematica? L'atto di osservazione nell'analisi dei dati richiede intelligenza o coscienza, oppure entrambe?

¹⁰² Ad esempio, questo tipo di qualità si presenta a livello macroscopico o microscopico? A livello macroscopico potrebbe richiedere una struttura che elabora informazioni in maniera integrata con parti fondamentali come l'ippocampo, corteccia parietale occipitale e parietale, posteriore, anteriore, talamo, insula e così via. In una considerazione del genere andrebbero inclusi non solo gli *Homo Sapiens*, ma anche corvidi, cetacei e primati non umani. Oppure a livello microscopico le quattro forze fondamentali potrebbero essere sufficienti per spiegarla, in particolar modo l'elettromagnetismo.

Andando oltre, si può provare per pura ipotesi a formalizzare un'informazione (i) con delle caratteristiche fisse come il prodotto tra lo stato chimico o fisico (s), un periodo di tempo (t), il tipo di osservatore (o) e il tipo di errori sistemici (e).

$$i[s, t, o, e]$$

Si può provare a pensare ad un filtro di verità come al raggiungimento matematico di una disequazione in cui l'informazione deve raggiungere una soglia minima.

$$i > 0$$

Difatti l'informazione supera la "soglia di verità" dal risultato dei prodotti delle caratteristiche che la compongono. Poniamo che nella produzione dell'informazione non si rileva nessun tipo di errore sembrando apparentemente perfetta (se $e = 0$ allora i verrà necessariamente 0 non superando la soglia di verità). Un altro errore potrebbe essere il caso dei modelli linguistici in un computer. Per esempio, ChatGPT potrebbe dare l'informazione: "Il caldo oggi è piuttosto fastidioso", pur non avendone le caratteristiche chimiche per poterlo affermare ($s = 0$). Se da una parte è già problematico formalizzare l'informazione, difatti di caratteristiche ce ne sarebbero a migliaia con altrettanti metodi di calcolo, ancora più problematico diventa determinare una "soglia di verità". Parlare di "soglia di verità" implica l'esistenza di un criterio oggettivo e universale, mentre nella pratica, sia biologica che computazionale, ciò che consideriamo vero è spesso il risultato di un processo di selezione funzionale. Nel sistema nervoso, ad esempio, i colori non rappresentano una verità ontologica, ma una codifica utile per distinguere oggetti e prendere decisioni rapide.¹⁰³

Allo stesso modo, nelle logiche dell'attuale *Data Science* la verità non coincide con una corrispondenza perfetta con la realtà, bensì con il raggiungimento di una prestazione sufficiente rispetto allo scopo: una soglia di accuratezza, un valore di probabilità, un livello di confidenza. Ciò potrebbe suggerire che un "filtro di verità" non debba essere inteso come un dispositivo che isola ciò che è vero in senso assoluto, ma come un meccanismo che seleziona informazione rilevante e sufficientemente affidabile per un determinato compito. La verità, dunque, si avvicina più a un concetto di utilità che a un ideale di corrispondenza perfetta con il mondo.

Ed ecco che entra in gioco uno degli aspetti cruciali per la Scienza dei Dati, la validazione dell'informazione. Secondo il neuroscienziato Daniel Bor la validazione avviene per selezione naturale attraverso tentativi ed errori.¹⁰⁴ L'informazione assume un valore per rilevanza e

¹⁰³ D. Hoffman, *L'illusione della realtà: come l'evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020, pp. 15-16-17

¹⁰⁴ Daniel Bor, *La voracità del cervello: cosa spinge la nostra coscienza verso un'insaziabile ricerca di significato*, Castelvechi, 2012, pp. 46-47-48

funzionalità, che può essere mostrata con la teoria dei giochi applicato all'etologia. Secondo la tabella un animale deve scegliere se costruire l'informazione "egoista" o "altruista" nel sistema cognitivo sulla base dei punti che riceve.

Tabella 1- Modello logico in cui si presuppone che l'azione di aiutare aumenti la fitness dell'animale che lo riceve e comporti un piccolo costo energetico a chi la effettua.

Animale 1/Animale 2	Aiuto	Non Aiuto
Aiuto	5,5	0,10
Non aiuto	10,0	0,0

Autore: Lorenzo Tiari

La tabella è volutamente semplice, infatti basta aggiungere ulteriori azioni che rappresentino le vere sfumature del mondo (ad esempio, aggiungendo il concetto di ambiente, inganno o reciprocità) per renderla talmente tanto complessa da occupare intere pagine del documento.

Un caso peculiare in natura è ciò che viene chiamato altruismo estremo, ovvero comportamenti anomali che vanno a ledere la vita del singolo organismo a discapito di altri individui della stessa specie.¹⁰⁵ Questo tipo di comportamento è un problema diretto alle logiche che vengono applicate nella comune *Data Science*. Un gesto di altruismo estremo riduce drasticamente la sopravvivenza, può ledere la vita di un organismo e sembra non avere benefici diretti immediati.

La prima domanda è: perché un organismo dovrebbe validare un comportamento apparentemente controproducente? Secondo una logica individuale il comportamento è "errato" e viola il principio di massimizzazione del beneficio individuale. Se utilizzassimo una logica collettiva questo tipo di comportamento potrebbe essere definito "giusto" e aumenta il beneficio collettivo. Si dovrebbe sostenere che la validazione non avviene su un evento specifico e limitato, bensì in un tempio più lungo. Qui emergono domande fondamentali: a quale livello viene validata l'informazione, singolarmente o collettivamente? La validazione è corretta su quali scale temporali? Molti comportamenti altruistici mostrano come i segnali sono costosi, gli indicatori affidabili e costruiscono informazioni ad alto valore predittivo. Un animale segnala il predatore danneggiandosi a livello individuale, ma aumentando la sopravvivenza del gruppo. Allora ritorniamo alle considerazioni dei paragrafi precedenti in cui si valida un'informazione perché

¹⁰⁵ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, pp. 293-294-295

vera o semplicemente perché funziona? Questa è la domanda che pone Bor sia sulla cognizione sia sulla computazione. Difatti la sua visione rispecchia molto una visione di analisi dei dati vicina sia al Fisicalismo sia al Materialismo. Il neuroscienziato inglese considera i processi cognitivi come una necessità del cervello fisico, in cui la coscienza si manifesta per risolvere problemi complessi e può essere ricondotta alla computazione. Il caso tipico è durante la formazione delle abitudini: azioni nuove come guidare una macchina appena acquistata sono cognitivamente “dispendiose” e richiedono attenzione, selettività, lentezza e costo energetico (per Bor sono caratteristiche essenziali della coscienza), mentre guidare la solita auto nello stesso percorso per arrivare a lavoro è un’azione poco dispendiosa e “automatizzata” a livello inconscio.¹⁰⁶

L’altruismo estremo costringe a riformulare la validazione ponendo domande radicali. Chi valida? L’individuo, il gruppo, il gene o il sistema cognitivo? Viene validata un’azione, una strategia o un modello comportamentale? Perché un sistema accetta un costo energetico del genere nella produzione di informazioni? Per stabilizzarsi creando coerenza logica oppure per innovare la sua struttura? Si potrebbe utilizzare un esperimento logico, propongo di chiedere ad un modello artificiale ciò che descrivo come il “paradosso di Re Leonida I”. Questo re spartano ha combattuto la sanguinosa battaglia delle Termopili nel 480 a.C. provocando in quel periodo storico una lunga catena di conseguenze, arrivando perfino ad una sua idealizzazione nel Rinascimento italiano. La domanda è: Re Leonida I è stato un ottimo o pessimo condottiero? In effetti, ha compatato le *polis* greche in un tempo successivo in seguito al suo sacrificio, ma nell’immediato è morto in modo cruento nella battaglia delle Termopili insieme a quasi tutto il suo esercito. Si riscontra che ogni metrica di validazione considerata ha limiti di logica ponendo una discussione infinita sui concetti di “vittoria”, “sconfitta”, “utilità”, “beneficio” e così via.

Questo continuo rimando tra scienze naturali e quelle informatiche torna utile per comprendere, al di là delle questioni su intelligenza e coscienza, che l’intelligenza artificiale rientra “passivamente” nelle stesse logiche dell’evoluzione naturale. Questo è ormai diventato evidente rispetto a tutti quei modelli artificiali di *Computer Vision* applicati alla robotica. In questi modelli l’apprendimento non si limita a compiti di classificazione, ma coinvolge un’azione fisica. Attraverso algoritmi di apprendimento per rinforzo il robot esplora l’ambiente, riceve segnali di successo o fallimento e modifica il proprio comportamento di conseguenza. Questa dinamica è formalmente analoga ai processi di adattamento biologico: variazione (esplorazione), riscontro ambientale (ricompensa o penalità) e stabilizzazione delle strategie efficaci. Anche in questo caso, comportamenti molto complessi emergono come effetto collaterale di un processo adattivo. Questo concetto emerge chiaramente nei libri di fantascienza del chimico Isaac Asimov di cui una

¹⁰⁶ Daniel Bor, *La voracità del cervello: cosa spinge la nostra coscienza verso un’insaziabile ricerca di significato*, Castelvechi, 2012, pp. 118-119-120

parte più che essere semplici storie, si mostrano come vere e proprie riflessioni di logica applicate alla robotica.

«Le Macchine non sbagliano, le loro premesse operative sono corrette. Su questo siamo d'accordo. Ora dici che è impossibile che le loro direttive vengano ignorate. E allora l'errore dov'è?» «Ti sei risposto da solo. *Non c'è nessun errore!* Pensa bene alla natura delle Macchine, Stephen. Sono robot, seguono la prima legge. Ma le Macchine non lavorano per un individuo, bensì per l'umanità, quindi, nel loro caso la legge dice: “una Macchina non può arrecare danno all'umanità, o mancando di agire, lasciare che l'umanità subisca un danno.”» «Bene. E dimmi, Stephen, cos'è che arreca danno all'umanità? Innanzitutto, i disallineamenti economici» [...] «Qual è, nel futuro, la causa più probabile di disallineamenti economici significativi?» «Mi verrebbe da dire» rispose Byerley «la distruzione delle Macchine.»¹⁰⁷

Asimov mostra che bastano pochissimi principi sulla realtà, quella che in filosofia può essere definita come morale, per arrivare a conclusioni paradossali. Le macchine di Asimov pur possedendo poche regole, nel momento in cui affrontano banali aspetti della vita quotidiana come ricevere istruzioni contrastanti che costringono a “riflettere”, essi finiscono in continue contraddizioni. Questo mostra come la metrica di validazione sia chiarissima in giochi strutturati come gli scacchi, ed è per questo che algoritmi fatti su misura siano in grado di surclassare chiunque.¹⁰⁸ Ma la vita è un gioco non molto chiaro in cui le metriche di validazione possono essere indeterminabili per “vincere”, arrivati a quel punto perfino l'intelligenza artificiale in diversi ambiti cessa la sua funzione. Politica, filosofia, etica nel diritto, scelta nella traduzione linguistica e molto altro non sono sostituibili non perché tipiche degli *Homo Sapiens*, perfino in varie specie di primati operano una sorta di “politica” e “diritto” più semplice per gestire il gruppo.¹⁰⁹ Non sono sostituibili perché le macchine agiscono entro vincoli formali e logici e anche quando tentano di “migliorare” modelli del mondo precedenti, lo fanno all'interno di un altro sistema di regole e non al di fuori della logica.

Le riflessioni di Asimov assumono una valenza sorprendentemente attuale. I modelli artificiali sono vincolati da strutture che ne delimitano il campo di azione nei confronti del mondo. Di conseguenza, l'utilizzo sproporzionato e acritico può produrre rischi sistemici sul campo della

¹⁰⁷ I. Asimov, *Io, Robot*, traduzione di Vincenzo Latronico, Mondadori libri., 2021, p. 247

¹⁰⁸ Doveroso ricordare rispetto alle ipotesi di una definizione convenzionale di osservatore analizzata precedentemente, i modelli hanno limiti nel reperire dati, infrastruttura, costi energetici e conservazione dell'informazione. Come l'influenza il giorno della gara per un giocatore di scacchi può peggiorare drasticamente le prestazioni anche un virus informatico può rovinare l'intero dataset rendendo il modello inutilizzabile.

¹⁰⁹ Sostenere che i primati abbiano una “politica” semplice è piuttosto riduttivo. Molte specie come i *Pan troglodytes* si comportano a volte come se seguissero una *Realpolitik* machiavelliana, una geometrica ricerca del potere in cui alleanze e amicizie cambiano sulla base della situazione. Peraltro, dimostrano come la “cattiveria” sia un tratto tipico dell'intelligenza che può portare a vantaggi evolutivi, fenomeno di cui ne sono affetti altri ordini come i cetacei e corvidi.

logica, ad esempio amplificare *bias* oppure porre una rigidità nella scelta.¹¹⁰ Nonostante il breve testo citato in precedenza sia un evidente iperbole, rimane la preoccupazione per le distorsioni economiche che si stanno andando a creare in cui l’allocazione delle risorse, le dipendenze tecnologiche a livello industriale e l’eccessiva delega decisionale a modelli intrinsecamente limitati porteranno a gravissimi problemi. Lo sfruttamento incontrollato dei modelli applicati alla politica e all’economia non ne stanno rompendo la logica, bensì ne stanno rompendo la struttura sociale. Generalmente gli organismi più “individualisti” o più “altruistici” non tendono a interrompere la propria linea evolutiva adottando strategie che li rimpiazzino o che rimpiazzino la loro stessa specie; al contrario, strategie molto più comuni in natura sono bugia e menzogna (come il mimetismo) utilizzate per ottenere vantaggi evolutivi maggiori. Come visto lungo il capitolo, per sopravvivere gli organismi devono tendere a un controbilanciamento tra conservazione e innovazione delle informazioni al fine di stabilizzarsi e migliorarsi nel tempo, poiché estremizzare le “idee del mondo” comporta quasi sempre risvolti drammatici per le specie.

¹¹⁰ I modelli artificiali come ChatGPT o Gemini posti di fronte a certi input: “come vedi il futuro dell’umanità fra 2.000 anni”, mostrano alcune predizioni che seguono la corrente filosofica del Transumanesimo.

Capitolo 3

Prevedere il passato per capire il futuro

La capacità di prevedere i fenomeni rappresenta una delle funzioni più importanti per gli organismi perché permette di intervenire sul mondo in modo tempestivo riducendo l'incertezza. Le informazioni che riceviamo dall'ambiente servono a modulare il nostro comportamento in funzione di determinati obiettivi. Tuttavia, valutare il valore di un'informazione in natura è tutt'altro che semplice. Si può provare a adottare una logica simile a quella utilizzata dagli algoritmi dei modelli generativi per interrogarsi su cosa significhi "prevedere" il mondo. Lao Tzu è stato l'esponente principale del Taoismo di cui detta i principi fondamentali circa 2500 anni fa in Cina.¹¹¹ Proviamo a prendere una parte del testo, come se fosse un *token*, analizzandolo dal punto di vista semantico.

Quando tutti riconoscono che una cosa è bella, un'altra diventa di conseguenza brutta. Quando un uomo viene ritenuto buono, un altro viene giudicato cattivo.¹¹²

Lo stile di scrittura si presenta come diretto, essenziale e lineare: restituisce un'impressione di apparente semplicità. Il testo è originariamente composto in logogrammi cinesi i quali non rappresentano suoni ma concetti, permettendo una notevole condensazione semantica e la trasmissione di idee estremamente complesse in uno spazio ridotto. È quindi opportuno tentare una traduzione non solo letterale, ma soprattutto concettuale del significato della frase. Lao Tzu concepisce la natura come una totalità continua, piuttosto che come una realtà fondata su divisioni nette all'interno della quale anche gli opposti possono coesistere e trasformarsi reciprocamente.

Ci si può allora interrogare sul motivo per cui enunciati di questo tipo risultino capaci di descrivere e anticipare fenomeni costanti della realtà pur apparendo semplici o persino banali.¹¹³ In questa frase Lao Tzu sembra cogliere in modo intuitivo principi cardine dell'etologia, secondo cui la cognizione negli animali ragiona per contrasti, quando qualcosa è visto come positivo viene creato immediatamente il suo opposto. Negli animali quando una caratteristica viene preferita (come un piumaggio vistoso o forza fisica) chi la possiede diventa "attraente" mentre chi non la possiede diventa automaticamente "non attraente". Non esiste

¹¹¹ Se il Tao te Ching sia stato scritto effettivamente da Lao Tzu o solo in parte o ancora come raccolta dei pensieri di più autori nel corso del tempo è ancora oggetto di dibattito da parte degli storici.

¹¹² Lao Tzu, *Tao te Ching*, Mondadori editore, a cura di B. Walker, 2016, p. 6

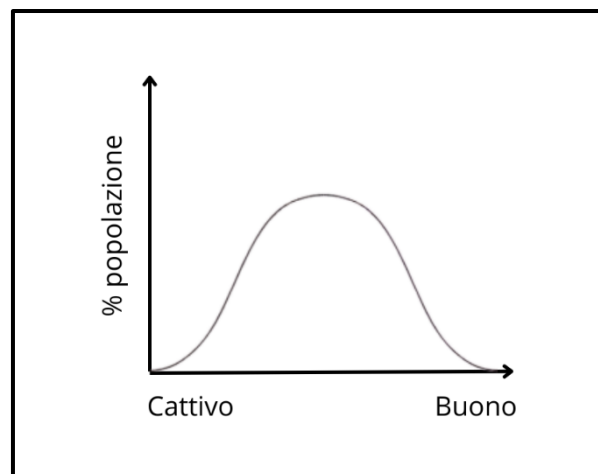
¹¹³ Nel Tao te Ching, Lao Tzu scrive chiaramente che i suoi insegnamenti sembrano semplici, ma nella pratica sono difficilissimi da comprendere. Questo perché nella pratica filosofica antica ci si concentrava più su una continua analisi della realtà tramite l'esperienza rispetto al nozionismo testuale.

bellezza senza confronto, in cui il riconoscimento del “buono” genera necessariamente a livello evolutivo il “cattivo”.¹¹⁴ Questo principio di novità, particolarmente evidente nei primati, rappresenta una costante osservabile in tutti i periodi storici. L’emergere di strutture sociali complesse in alcune specie ha portato a categorizzazioni nette sulla realtà e il filosofo taoista consiglia implicitamente di non attaccarsi ai giudizi, che, come si è visto, saranno sempre necessari finché avremo un sistema nervoso.

Nel governare il popolo e nel servire il cielo niente è meglio della moderazione.¹¹⁵

Lao Tzu identifica implicitamente quello che viene definito come il teorema del limite centrale, un pilastro della statistica in cui si afferma che molti fenomeni indipendenti apparentemente asimmetrici o strani, se si calcola la media delle somme di molte di queste distribuzioni, seguono una distribuzione a campana.¹¹⁶ Se ci riferiamo alla biologia evolutiva i tratti estremi spesso vengono selezionati negativamente, il troppo altruismo può comportare la morte del singolo, mentre l’eccessiva aggressività comporta la perdita di molti individui della propria specie. Genetica ed epigenetica negli individui nella maggioranza dei casi seguono una distribuzione moderata, Lao Tzu capisce che un buon governo mantiene la popolazione nella zona adattiva centrale. Intuitivamente si può immaginare questa società così:

Figura 9- La media delle somme delle distribuzioni nel tempo produce concettualmente una distribuzione “morale” gaussiana.



Autore: Lorenzo Tiari

¹¹⁴ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, pp. 66-67-68-69-70-71

¹¹⁵ Lao Tzu, *Tao te Ching*, Mondadori editore, a cura di B. Walker, 2016, p. 71

¹¹⁶ S. Borra e A. Di Ciaccio, *Statistica: metodologie per le scienze economiche e sociali*, Graw Hill, 2021, pp. 246-247

Negli ultimi decenni, lo sviluppo di modelli di elaborazione dei dati basati su tecniche statistiche ha trasformato in modo significativo le modalità con cui le scienze sociali analizzano e interpretano il comportamento umano. Come già osservato, questi modelli condividono vincoli strutturali simili a quelli degli organismi biologici nella produzione dell'informazione e si fondano inevitabilmente su approssimazioni ed errori sistematici. Può apparire intuitivo attribuire la loro efficacia al successo ottenuto in ambiti come la fisica o la matematica, ma si trattano di contesti profondamente diversi. Il cervello umano non si è evoluto nel corso di milioni di anni per prevedere fenomeni su scale estreme, né microscopiche né macroscopiche. Non vi è mai stata alcuna necessità adattiva di calcolare la traiettoria di un missile interplanetario o di comprendere concetti astratti come il numero π o i numeri irrazionali. Nei nostri modelli cognitivi quotidiani non esistono entità come “-1 mela” o “3,1415 porzioni di mela”. Di conseguenza, il sistema nervoso non si è adattato a operare in contesti fortemente formali. È quindi del tutto comprensibile che algoritmi sviluppati da individui dotati di grandi capacità matematiche risultino “eccezionali” per la maggior parte delle persone.

Il problema si presenta in senso inverso nel momento in cui si tende sempre più spesso a delegare ai modelli computazionali compiti che risultano invece naturali per il sistema nervoso umano, tra cui la costruzione di modelli predittivi in ambito sociale. Fenomeni complessi come la diffusione delle opinioni, le dinamiche economiche, le decisioni individuali e collettive, o persino la probabilità di comportamenti devianti vengono affrontati attraverso sistemi di calcolo meccanico capaci di elaborare grandi quantità di dati e di individuare regolarità statistiche per produrre informazioni ritenute rilevanti. In questo scenario, la capacità predittiva tende a configurarsi non solo come uno strumento analitico, ma come un vero e proprio criterio di successo scientifico e operativo. Tuttavia, l'ambizione di rendere il comportamento sociale pienamente prevedibile mediante algoritmi matematici solleva interrogativi teorici di primaria importanza. Le interazioni sociali, anche alla luce delle prospettive offerte dall'epigenetica, non sono il semplice risultato di un insieme di variabili misurabili, ma emergono dall'intreccio dinamico tra fattori biologici e un ambiente in continua trasformazione come credenze, storia individuale e collettiva, errori, norme culturali che contribuisce a rendere il comportamento umano esponenzialmente complesso. In questo spazio in tensione tra formalizzazione e comprensione che emerge il confronto tra modelli predittivi artificiali e quelli dedotti da un sistema nervoso complesso. Il cervello, infatti, può essere descritto come un grande sistema predittivo in cui la mente umana opera costantemente nell'analizzare esperienze passate e anticipare gli stati futuri del mondo. Ad esempio, andiamo verso il pulsante di casa perché prevediamo nel breve periodo che la luce si accenda. Oppure, calcoliamo i benefici di una corsa perché sul medio periodo prevediamo che il corpo ne risenta

positivamente. Infine, decidiamo di studiare perché prevediamo che sul lungo periodo queste azioni apportino benefici maggiori in termini di benessere e denaro. Il famoso “senso storico” che viene attribuito ai grandi pensatori non è nient’altro che una capacità estrema di collegare cause ed effetti nel tempo su scale gigantesche. L’imperatore Marco Aurelio, romano e filosofo dello Stoicismo istituzionale, circa 1900 anni fa durante brutali campagne militari, contrasti religiosi e schiavismo diffuso, riflette sulla creazione di un possibile stato mondiale. Oggi studi in etologia e antropologia mostrano chiaramente che nella gran parte degli *Homo Sapiens* vi sono dei principi comuni ed estremamente efficaci dettati dall’evoluzione, quelli che il filosofo chiama “ragione”.¹¹⁷

Se l’intelletto è comune a noi tutti- scrive Marco Aurelio – allora lo è anche la ragione, in virtù della quale siamo esseri razionali. Se è così [...] è comune anche la legge, e di conseguenza siamo tutti cittadini; perciò, siamo tutti partecipi di una forma di governo, e allora l’universo è come un’altra città.¹¹⁸

Nel dominio sociale il cervello mostra una straordinaria abilità nel cogliere segnali deboli, ambigui o non esplicitamente codificabili come il tono di voce, l’inganno celato dietro una certa espressione facciale o una sottile violazione della legge. Questo non è dovuto ad una formula magica o da qualche potere nascosto, è un comportamento perfettamente adattivo di cui peraltro non è necessario nemmeno avere strutture enormemente complicate come quella degli *Homo Sapiens*.¹¹⁹ Nella teoria dei giochi se formalizziamo stati del mondo con un certo numero di azioni tipiche di specie con strutture sociali (aiuto, inganno, ambiente e così via) il modello richiederà una capacità logica di altissimo livello per comprendere in che modo gli organismi possano reperire nella maniera più efficace risorse nell’ambiente.

Questa continua previsione dei comportamenti altrui costringe il sistema nervoso a rimodularsi ed è per questo, nonostante la ripartizione del cervello in maniera distinta sia uno schema superato nelle neuroscienze, ci sono specifiche aree cerebrali sviluppate per questi tipi di compiti.¹²⁰ Se dovessimo trasferire questo concetto nella computazione potremmo impropriamente dire che il cervello quando trova una funzione logica che massimizza il risultato rispetto alla metrica di validazione, tenta di ottimizzarla fino a raggiungere l’obiettivo predeterminato. Sostenere che il cervello sia “migliore” di una macchina nei modelli predittivi sociali non implica una negazione del valore a questi modelli, ma ne riconosce i limiti. Nel loro insieme le aree cerebrali

¹¹⁷ Imperatori come Marco Aurelio o Traiano vengono visti come “buoni” per diffusi aiuti dati ai bambini e ragazzi di Roma. Nelle scienze questo comportamento è definito come alloparentalità, ovvero cure rivolte alla prole da parte di individui che non hanno forti legami genetici.

¹¹⁸ P. Grimal, *Marco Aurelio*, Garzanti libri s.p.a., 2004, p.100. Il testo è preso dai *pensieri* dell’imperatore romano, gli scritti durante le campagne militari negli ultimi anni di vita.

¹¹⁹ *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022, pp. 122-123

¹²⁰ Le aree coinvolte generalmente sono l’ippocampo, la corteccia prefrontale, i circuiti dopaminergici, corteccia cingolata anteriore e così via.

costituiscono una rete integrata di informazioni che consente all'organismo, a determinate condizioni, di trovare regole sofisticate con pochissimo costo energetico ed ecologico. Questa infrastruttura tramite selezione naturale è stata resa fortemente plastica e modulabile consentendo una netta comprensione del significato e del contesto. La domanda è capire quali siano queste condizioni. In questo paragrafo saranno enfatizzate le abitudini naturali mentre in quello successivo si analizzeranno le capacità degli algoritmi informatici.

Queste condizioni, seguendo la logica taoista della complementarità e dell'apparente paradosso, sono "difficilmente semplici". Le predizioni sociali si fondano su un sistema nervoso che comunica a livello elettrico e chimico, enormemente plastico durante il corso della vita. Da questo punto di vista, alcune abitudini quotidiane consentono di rafforzare enormemente queste reti di neuroni migliorando l'efficienza dei processi biologici innescando alcune sostanze come la mielina.¹²¹ La mielina è una guaina isolante composta dal 70% circa di lipidi. Gli assoni dei neuroni mielinizzati non sono totalmente ricoperti da questa sostanza, ma solamente a intervalli.¹²²

Questa organizzazione consente al potenziale d'azione di propagarsi attraverso una conduzione saltatoria, in cui il segnale "salta" da un nodo all'altro, anziché diffondersi in modo continuo lungo l'assone. Grazie alle proprietà isolanti della mielina, la velocità di trasmissione dell'impulso nervoso aumenta significativamente passando da circa 10 m/s fino a valori prossimi ai 100 m/s, con un incremento in proporzione di almeno 10 volte tanto.¹²³ In primo luogo, l'attività fisica (specialmente se aerobica) come la corsa o la camminata veloce, rappresenta dal punto di vista neurobiologico, un maggiore afflusso sanguigno cerebrale in grado di rendere reazioni chimiche e metaboliche più reattive e veloci supportando la sopravvivenza dei neuroni e rafforzando connessioni neurali. Un cervello più efficiente dal punto di vista metabolico è in grado chimicamente di aggiornare meglio le proprie previsioni, riducendo il "rumore" cognitivo e aumentando la sensibilità ai segnali sociali rilevanti. Inoltre, l'attività fisica consente di migliorare funzioni motorie e di regolazione emotiva, processi fondamentali per l'interpretazione del comportamento altrui. Alcune attività come musica o la partecipazione a giochi strategici costituiscono un ulteriore potenziamento. Queste attività richiedono costante pianificazione e previsione delle azioni che vengono immediatamente punite o ricompensate tramite riscontri immediati, oltre che attivare costantemente varie cortecce cerebrali come quella uditiva, limbica e prefrontale. Assumono rilevanza centrale abitudini come lettura e scrittura in grado di

¹²¹ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, pp. 80-81

¹²² Ibidem

¹²³ Ivi, pp. 82-83

contribuire al miglioramento dei modelli previsionali, in cui l'apprendimento leggero, strutturato e diversificato favorisce nuove sinapsi e la loro stabilizzazione.

Il primo motivo è evidente ed è “sfruttato” dai modelli linguistici, ovvero reperire grandi quantità di informazione e possibilmente di alta qualità. Come un dataset con *token* scadenti, anche i libri letti che hanno poca rilevanza producono un “output” cognitivo mediocre. Effettivamente la domanda è quali sono le fonti di informazione che consentano di fare questo. Una domanda alla quale pensatori come Cardano avevano già trovato la risposta nell'uomo rinascimentale, un costante equilibrio tra arte, gioco, scienza e filosofia in cui era necessario di unire con grande ecletticità innumerevoli discipline.¹²⁴

Il secondo motivo è che il cervello non sempre distingue precisamente tra realtà e simulazione, le parole possono andare a “toccare” anche aree anomale come quella motoria permettendoci di costruire “ipotesi sul mondo”, il caso tipico è la sensazione di sognare ad occhi aperti. È plausibile ritenere che il progresso tecnologico, insieme a studi sempre più accurati sul funzionamento del cervello e all'esplosione delle scienze negli ultimi decenni, possa contribuire in futuro a un potenziamento delle capacità cognitive umane anche attraverso processi naturali di adattamento e apprendimento.

3.1 Il bilanciamento tra comprensione e approssimazione

Nel contesto della Scienza dei Dati la costruzione di un dato rappresenta un passaggio cruciale che precede qualsiasi attività di inferenza o previsione. Come già visto, i dati non costituiscono solamente una trascrizione della realtà, bensì un risultato di formalizzazione logica rispetto al fenomeno osservato. Ogni dataset in ambito sociale è una mediazione tra ciò che può essere analizzato dal punto di vista dell'elaborazione automatizzata, un bilanciamento tra comprensione degli eventi e una loro approssimazione. La comprensione richiede definizioni chiare, categorie a priori e un legame esplicito con la produzione del dato del fenomeno che esse intende interpretare. Un dato “comprensibile” è un dato che può essere spiegato, contestualizzato e discusso criticamente sia da chi lo produce che da chi lo utilizza. Tuttavia, realtà sociali come quelle economiche o demografiche sono caratterizzate da difficoltà esponenziali in cui ogni modello fa fatica a catturarne la sua interezza. Per rendere questa complessità osservabile e confrontabile nel tempo e nello spazio è necessario introdurre convenzioni o semplificazioni che costituiscono forme di approssimazione. L'approssimazione non è quindi un errore in senso

¹²⁴ La multidisciplinarietà è un fenomeno che si rileva nei grandi periodi storici di innovazione: nell'ideale dell'uomo romano fino a quello rinascimentale, nell'epoca d'oro islamica, ma anche in Cina durante la dinastia Song, in cui nacque Shen Kuo, uno dei più grandi scienziati della storia. Perfino il suo capolavoro più importante in grado di unire materie come astronomia, arte, filosofia e tecnologia viene generalmente tradotto in “Saggi sul torrente dei sogni”. L'enorme progresso in questo secolo deriva anche dal ritorno dell'ecletticismo, anche se è stato effettuato in maniera più caotica rispetto al passato.

stretto, si è già notato come la complessità possa essere adeguatamente approssimata con relativa facilità, ma una condizione strutturale della logica, della matematica e della statistica.

In questo quadro, la produzione di dati è l'insieme di scelte attuate dallo scienziato che comprendono definizione operativa dei concetti, scelte metodologiche, rilevazione, trattamento di dati mancanti e le modalità di aggregazione. Ognuno di questi passaggi riduce la complessità del dato a fronte di una sua migliore leggibilità. Di conseguenza, un eccesso di approssimazione può produrre informazioni formalmente "corrette", ma fuorvianti o di ambigua interpretabilità. Al contrario, il tentativo di preservare quanto più possibile la complessità del fenomeno rischia di generare modelli del mondo troppo articolati e caotici difficilmente utilizzabili nei modelli computazionali. Gli scienziati dei dati tendono generalmente a concentrarsi su alcune metriche di validazione particolari: prestazioni predittive, efficienza computazionale rispetto alle risorse a disposizione e assunzioni tipiche dell'informatica, ma che non si riscontrano nelle scienze naturali.¹²⁵ L'accuratezza predittiva o l'errore quadratico medio vengono comunemente assunti come di buona qualità per il modello indipendentemente dal grado di aderenza delle sue ipotesi rispetto alla realtà. Questo approccio, se da un lato consente di costruire informazioni standardizzate e riproducibili secondo standard nazionali o internazionali, dall'altro tende a spostare l'attenzione dalla comprensione del fenomeno alla sola ottimizzazione di una funzione formalmente definita. Le scienze naturali possono supportare l'osservazione offrendo chiavi interpretative utili anche al di fuori del dominio originario di studio, contribuendo alla comprensione di comportamenti economici e organizzativi. Come già visto nel secondo capitolo, il collegamento tra scienze informatiche e naturali non nasce da una somiglianza superficiale nell'analisi dei dati, ma da una riflessione che tocca la teoria di fondo.

In primo luogo, se nei modelli computazionali l'errore indica generalmente una mancanza negli algoritmi utilizzati, in natura l'errore può tramutarsi in una prospettiva di miglioramento che giustifica nuovi studi o nuove direzioni di ricerca. L'errore, quando si costruiscono modelli predittivi sociali, non è una deviazione da minimizzare come soglia di accuratezza, ma come una vera e propria informazione che consente di ampliare variabili non considerate. Richiedendo uno sforzo energetico (che può essere visto banalmente anche come denaro aggiuntivo alla ricerca) si deve possibilmente preferire una comprensione generalizzabile della realtà invece che la stabilità delle prestazioni che rimangono molto statiche nel tempo.

In secondo luogo, un dataset in ambito sociale deve consentire una buona generalizzazione teorica del fenomeno consentendo di prevederne altri simili nel tempo. Ricerche trimestrali continue possono essere sufficienti per capire brevemente l'andamento di uno stato, ma rischiano di

¹²⁵ La primatologia è il campo scientifico che studia in modo specifico i primati analizzando comportamenti, genetica, evoluzione ed ecologia. I principi cardine di questa scienza sono importanti per comprendere le dinamiche della nostra società come falsificazione, furto, investimento e così via.

produrre informazioni “fragili” e quindi “idee del mondo” scadenti, difficilmente trasferibili in altri contesti in cui le condizioni possono cambiare drasticamente. L’assenza di forti ancoraggi teorici rendono i dataset e i rispettivi modelli di elaborazioni vulnerabili a rotture strutturali che le metriche di validazione tradizionali non sono in grado di anticipare. Inoltre, la continua sovrabbondanza di dati in enti ufficiali non indica buona qualità del lavoro, ma produce spesso caos in cui ricerche statistiche di qualità vengono sovrastate da altre ricerche statistiche di qualità minore. Bisogna considerare che, se dataset applicati all’ambito sociale non supportano adeguatamente le decisioni politiche, allora la rispettiva produzione diventa semplicemente un sistema automatizzato “auto-rinforzante” in cui i dati diventano solo approssimazioni senza alcuna comprensione dei fenomeni analizzati.

Per rendere concreti i concetti discussi, è utile analizzare un dataset prelevato dall’Istituto Nazionale di Statistica (ISTAT) che è il principale produttore di statistiche ufficiali da cento anni, offrendo un esempio utile per delineare questo bilanciamento in un quadro metodologico rigoroso. Questi dataset non sono semplici raccolte di numeri, ma rappresentazione di fenomeni complessi che toccano la vita quotidiana dei cittadini. Si possono analizzare tanti aspetti della prassi metodologica impiegata per comprendere le scelte impiegate durante la lunga catena di produzione dal dato fino all’informazione vera e propria. Attraverso lo studio di questi meccanismi è possibile applicare concretamente concetti precedentemente discussi: l’ambivalenza dell’errore, comprensione e approssimazione, generalizzazione teorica dei fenomeni. Il dataset scelto riguarda gli aspetti della vita quotidiana delle famiglie italiane che opera un sistema integrato di analisi statistiche avviato nel 1993, i dati sono stati prelevati nel 2023 e resi disponibili ad analisi pubbliche nel 2025.¹²⁶

Il dataset “Aspetti della vita quotidiana” (AVQ) – anno 2023, rappresenta una delle principali fonti ufficiali italiane per lo studio delle condizioni di vita, dei comportamenti sociali e delle percezioni soggettive degli individui e delle famiglie residenti in Italia. Il documento analizzato rientra nella categoria dei microdati ossia dati anonimizzati resi disponibili a fini di ricerca e analisi statistica. Dal punto di vista tecnico, il dataset non può essere considerato un semplice insieme di dati “pronto all’uso”, bensì un sistema informativo complesso caratterizzato da una struttura multilivello, da un disegno di campionamento articolato e da una forte dipendenza da metadati esterni per la corretta interpretazione delle variabili. Tali caratteristiche rendono l’analisi particolarmente interessante ma, al contempo, metodologicamente problematica se affrontata con modelli di elaborazioni dati tipici del *machine learning*.

¹²⁶ <https://www.istat.it/microdati/aspetti-della-vita-quotidiana/> (attivo). Dal sito è sufficiente cliccare download rispetto all’anno 2023 e si scaricherà una cartella che contiene metadati e microdati con le indicazioni metodologiche. Il dataset vero e proprio viene nominato “AVQ_Microdati_2023.txt”, per analizzarlo si può utilizzare un software come Excel e importarlo tramite tabulazione.

La questione metodologica, nonostante venga rigorosamente descritta in stile accademico, si presenta come eccessivamente complessa rispetto alle conoscenze di un cittadino, escludendolo di conseguenza. Ad esempio, in uno dei documenti più importanti denominato “AVQ_NotaMetodologicaIndagine_2023.pdf” sono presenti descrizioni accurate piuttosto semplici da capire, senza però nessuna immagine concettuale, alternate da rapide argomentazioni di alto livello matematico. La scelta è sensata, ma non corretta dal punto di vista comunicativo. Se un giornalista ha la necessità di andare a verificare la correttezza dei dati o perlomeno comprendere come sono stati rilevati, si ritroverà pezzi di testo comprensibili abbinati a formule matematiche che richiedono alto livello di comprensione matematica. Si sarebbe potuto utilizzare una formulazione linguistica che ormai è diventata abitudine nella varietà linguistica definita come italiano istituzionale: partire da informazioni essenziali che corrispondono a tutto ciò che è necessario sapere, senza entrare in termini tecnici con alto grado di specialismo, e solo in seguito entrare nel dettaglio nei paragrafi con le varie formule matematiche.¹²⁷ Per il cittadino sarebbe preferibile abbinare testi introduttivi semplici, ma esaustivi, con adeguate infografiche (ad esempio in merito alla tecnica di campionamento a grappolo).¹²⁸

Il dataset è di tipo osservazionale con l'intervista prevede l'utilizzo di due questionari: uno per intervista diretta e uno per autocompilazione tramite tecniche CAWI-CAPI/PAPI, con periodo di riferimento prevalentemente coincidente con i dodici mesi precedenti l'intervista. L'unità di analisi principale è l'individuo, mentre l'unità di campionamento è la famiglia. Tutti i componenti appartenenti alla famiglia selezionata vengono sottoposti a rilevazione il che introduce una struttura gerarchica implicita nei dati: individui annidati all'interno di famiglie, a loro volta collocate in specifici contesti territoriali. L'indagine utilizza un disegno di campionamento a due stadi, stratificato per regione e tipologia comunale. Dal punto di vista metodologico, questo disegno introduce diversi livelli di complessità come la dipendenza intra-familiare nelle risposte. In un contesto come quello della Scienza dei Dati, l'utilizzo diretto del dataset senza tener conto del disegno campionario equivale a trattare i dati come se fossero ottenuti tramite un campionamento casuale semplice. Questo rappresenta una delle principali fonti di errore metodologico nelle analisi non ufficiali (come notizie online) basate su microdati ISTAT.

Uno degli errori centrali del dataset è che i bambini sotto i 14 anni sono fondamentalmente esclusi nelle varie sezioni del questionario, solo un genitore o un componente maggiorenne può rispondere in loro vece. La motivazione è che i bambini vengono ritenuti nelle indagini statistiche come non affidabili a livello cognitivo, non avendo ancora coltivato una soddisfacente percezione della realtà. Come è stato notato nell'analisi degli esperimenti svolti in ambito etologico, chiunque

¹²⁷ D. Vellutino, *L'italiano istituzionale per la comunicazione pubblica*, Il Mulino, 2018, pp.46-47

¹²⁸ Ivi, p. 183

è profondamente soggetto a *bias* e a costruzioni del mondo “scarsamente affidabili”.¹²⁹ Gran parte delle variabili oggetto dell’indagine riguarda infatti aspetti soggettivi e comportamentali come la soddisfazione per la propria vita, la percezione della salute, la fiducia nelle istituzioni e la valutazione dei servizi. Questi concetti richiedono capacità cognitive e di astrazione che, nei minori di 14 anni non risultano sufficientemente sviluppate in modo uniforme. L’inclusione diretta dei bambini comporterebbe quindi un elevato rischio di risposte incoerenti o casuali con un conseguente aumento dell’errore di misurazione. Tuttavia, utilizzare altre persone del nucleo familiare introduce un ulteriore problema metodologico, poiché le risposte rifletterebbero la percezione del genitore o del tutore e non quella del bambino, generando un *bias* sistematico e riducendo la validità concettuale del dato individuale. Per questo motivo, ISTAT preferisce limitare l’uso del *proxy* e, nel caso dei minori, optare per l’esclusione dalle sezioni soggettive del questionario. Dal punto di vista statistico, l’esclusione dei minori contribuisce inoltre a mantenere una maggiore omogeneità della popolazione di riferimento. Includere bambini e adulti nello stesso strumento di rilevazione richiederebbe questionari, scale di misura e domini concettuali differenti, compromettendo la comparabilità delle stime tra individui e territori. La scelta adottata consente quindi di produrre stime più stabili e confrontabili in linea con gli obiettivi conoscitivi dell’indagine. Tuttavia, questa impostazione comporta una perdita informativa rilevante sulla popolazione infantile e impedisce analisi dirette sul benessere dei bambini, che viene osservato solo indirettamente attraverso le caratteristiche familiari. Inoltre, questa scelta introduce una mancanza strutturale di informazioni per una parte della popolazione limitando la possibilità di costruire modelli predittivi completi e di analizzare in modo approfondito le dinamiche intergenerazionali. In sintesi, l’esclusione dei minori di 14 anni nell’indagine rappresenta un’approssimazione a discapito della comprensione completa del fenomeno.

Il dataset comprende numerose variabili in cui le modalità di risposta sono codificate numericamente, ma i codici non hanno un significato quantitativo intrinseco. Ad esempio, valori come 1 o 2 nella categoria “sesso” rappresentano categorie concettuali distinte come maschio e femmina e non possono essere interpretati come grandezze senza una preventiva ricodifica.¹³⁰ Queste operazioni sono a rischio di interpretazioni errate dell’ordine delle categorie. Nei modelli di *machine learning* che sfruttano dataset numerici sono necessarie continue operazioni di traduzione rispetto al significato semantico dei numeri. Senza un’accurata fase di pulizia l’utilizzo

¹²⁹ In merito a questo è importante il caso aziendale Lego in merito al concetto di *small data*. Interviste dirette a pochi bambini ha consentito di cambiare la strategia aziendale dell’azienda di giocattoli in una strategia di successo, in cui alcune dinamiche osservate si sono presentate successivamente anche in fasce di consumatori con età più grandi.

¹³⁰ In questo senso vengono escluse le persone che hanno effettuato una transizione di genere tra i sessi tramite operazioni o medicinali.

delle variabili l'analisi risulta metodologicamente fragile, rendendo ancora più soggettiva la costruzione di un modello predittivo.

Un altro aspetto particolarmente critico del dataset riguarda la gestione dei valori mancanti descritti nel documento “AVQ_DescrizioneFile_2023.pdf”. ISTAT non utilizza valori mancanti standard (NA), ma codici specifici diversi nelle variabili per distinguere tra non risposta, rifiuti o esclusioni. Queste tipologie di mancata informazione non sono casuali e, in molti casi, sono informative rispetto al fenomeno studiato. Tratarle indiscriminatamente come valori mancanti o eliminarle dal dataset comporta una perdita informativa significativa e può introdurre *bias* di selezione. Inoltre, durante le operazioni di pulizia si possono innescare errori come complete rimozioni di dati producendo informazioni sbagliate.¹³¹

In conclusione, il dataset ISTAT “Aspetti della vita quotidiana 2023” rappresenta una fonte informativa di buona qualità istituzionale e di valore per l'analisi sociale. Tuttavia, presenta alcune criticità: una metodologia fortemente standardizzata applicata in un contesto dinamico, l'esclusione di una parte di generazione fondamentale per la società e la mancata inclusione di principi generalizzabili tipici delle scienze naturali.

3.2 Risorse fondamentali per gestire l'incertezza

Nel contesto dell'analisi dei dati l'incertezza rappresenta una componente intrinseca della realtà. Gli organismi viventi operano costantemente in ambienti dinamici, parzialmente osservabili e soggetti a variazioni imprevedibili, nei quali l'informazione costruita è spesso incompleta, rumorosa o ambigua. In ambito biologico, l'incertezza riguarda non solo fattori esterni come la disponibilità di risorse o la presenza di predatori, ma anche processi interni quali la percezione sensoriale e l'elaborazione delle informazioni. Di conseguenza, i sistemi naturali non sono progettati per eliminare l'incertezza, bensì per convivere con essa e adattarsi in modo efficace all'ambiente. Una delle principali criticità dell'incertezza in natura risiede nel rischio di prendere decisioni non ottimali o errate. Una valutazione imprecisa dell'ambiente può compromettere la sopravvivenza dell'individuo, ad esempio inducendo comportamenti di fuga non necessari o, al contrario, una sottovalutazione di una minaccia reale. Tuttavia, l'incertezza non costituisce esclusivamente un limite, ma svolge anche un ruolo funzionale nei sistemi naturali. La presenza di variabilità e di componenti stocastiche favorisce la flessibilità e la capacità di adattamento consentendo agli organismi di esplorare strategie comportamentali alternative. In etologia, l'esplorazione è spesso associata a meccanismi di apprendimento e adattamento, attraverso i quali gli individui modificano il proprio comportamento in risposta a cambiamenti ambientali. In

¹³¹ Per esempio, utilizzando il software KNIME per la costruzione di algoritmi può succedere che mancanze di numeri o simboli anomali rimuovano intere colonne o righe di dati.

biologia evolutiva, inoltre, l'incertezza è strettamente connessa alla diversità genetica che rappresenta un vantaggio fondamentale per la sopravvivenza delle specie in ambienti mutevoli. I sistemi naturali tendono pertanto a sviluppare meccanismi di compensazione come l'aggregazione di informazioni, il comportamento collettivo o l'auto-organizzazione, che permettono di ridurre localmente l'incertezza senza eliminarla completamente. In questi scenari, l'approccio tradizionale basato su modelli deterministici e supervisionati risulta spesso inadeguato, rendendo necessario l'impiego di metodologie capaci di adattarsi a contesti poco strutturati e caratterizzati da elevata ambiguità informativa.

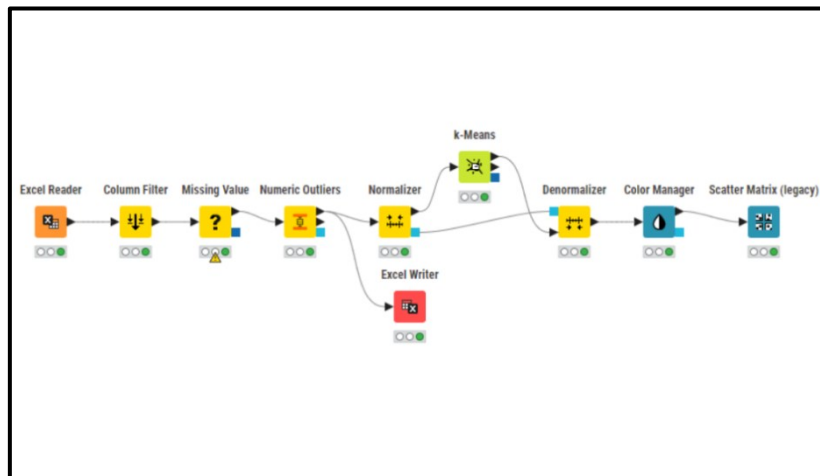
In risposta a queste esigenze, numerosi algoritmi sviluppati nell'ambito nel campo di studi dell'intelligenza artificiale si ispirano a principi osservabili in sistemi naturali e complessi. Questi algoritmi, pur non riproducendo fedelmente processi biologici o fisici, ne condividono l'approccio concettuale privilegiando strategie iterative per l'esplorazione dello spazio delle soluzioni. In questo senso, essi possono essere interpretati come strumenti efficaci per affrontare e gestire l'incertezza, piuttosto che eliminarla completamente. Tra le tecniche più diffuse in questo ambito rientrano i modelli di apprendimento non supervisionato che consentono di individuare schemi e relazioni latenti nei dati senza fare affidamento su etichette predefinite. In particolare, il *clustering* rappresenta una strategia fondamentale per la semplificazione e l'organizzazione dell'informazione, permettendo di raggruppare osservazioni simili e ridurre la complessità del problema analizzato.

Per delineare un caso concreto e molto semplice concettualmente si possono analizzare le abitudini di lettura dei cittadini con il dataset descritto dell'ISTAT. Si farà riferimento al numero di libri cartacei e non verranno prese in considerazione altre variabili come gli e-book. Questa scelta è dettata sia da un numero considerevole di risposte mancanti sia per le differenze cognitive che avvengono tra la lettura su carta rispetto a quella di un dispositivo digitale.¹³² Verrà impiegato l'algoritmo *K-Means* utilizzato su un software di analisi dati come KNIME.¹³³ Il suo obiettivo è suddividere le osservazioni in un numero prefissato di gruppi di cittadini, detti *cluster*, in modo tale da rilevare caratteristiche comuni e tendenze nascoste. KNIME ha il vantaggio di utilizzare i cosiddetti "nodi", ovvero una serie di operazioni tipiche durante l'analisi dei dati (pulizia, normalizzazione e così via) in una grafica che è già stata programmata a priori.

¹³² La lettura su dispositivi digitali può portare a distrazioni continue (come notifiche) e variazioni elettromagnetiche a livello visivo.

¹³³ La versione del software KNIME impiegata è la 5.2.3 del 27 marzo 2024. Non verranno introdotti formalismi matematici essendo un algoritmo "gestibile" e molto trattato nell'ambito della *Data Science*.

Figura 10- Modello di elaborazione dati in un *workflow* di KNIME.



Autore: Lorenzo Tiari

Il processo ha inizio con il nodo *Excel Reader* in cui si importa e si controlla il dataset.¹³⁴ Il secondo nodo denominato *Column Filter* consente di mantenere solamente le variabili congeniali al nostro scopo, la scelta è ricaduta sul grado di istruzione dei cittadini e i libri cartacei letti negli ultimi 12 mesi.¹³⁵ Il nodo *Missing Value* consente di eliminare tutti i cittadini che presentano dati mancanti in modo da poter associare sempre il grado di istruzione con i libri cartacei letti.¹³⁶ Il nodo *Numeric Outliers* ha rimosso tutti i valori anomali che potevano alterare la normalizzazione dei dati in sequenze di numeri accostabili.¹³⁷ In seguito si è usato il nodo *Normalizer* poiché l’algoritmo utilizza misure di distanza per valutare la similarità tra le osservazioni, risulta fondamentale che le variabili siano espresse su scale confrontabili. A tal fine, viene comunemente applicata una procedura di normalizzazione, in questo caso la normalizzazione detta *min-max*, che consente di ricondurre i valori delle diverse variabili all’interno di un intervallo comune (da 0 fino a 1), riducendo l’influenza di attributi con ordini di grandezza differenti.¹³⁸ Il nodo *K-Means* è quello cruciale poiché contiene l’algoritmo che determinerà gruppi di persone consentendo di costruire informazioni rilevanti. L’algoritmo prevede inizialmente la scelta dei centroidi dei cluster, se sono scelti 3 con l’idea di classificare le persone in qualcosa come “lettori forti”, “lettori medi” e “lettori deboli”. Questi centroidi rappresentano il centro di ciascun gruppo e vengono

¹³⁴ L’unica modifica avvenuta in questo nodo è stata nell’impostazione: *advanced* → *limit scanned rows*: 50.000. Questa modifica serve per scannerizzare correttamente tutto il dataset.

¹³⁵ L’unica modifica avvenuta nel nodo è: *includes* → ISTRIMi e NLIBRI_CARM. Per l’ISTAT indicano rispettivamente il livello di istruzione e il numero di libri cartacei letti negli ultimi 12 mesi.

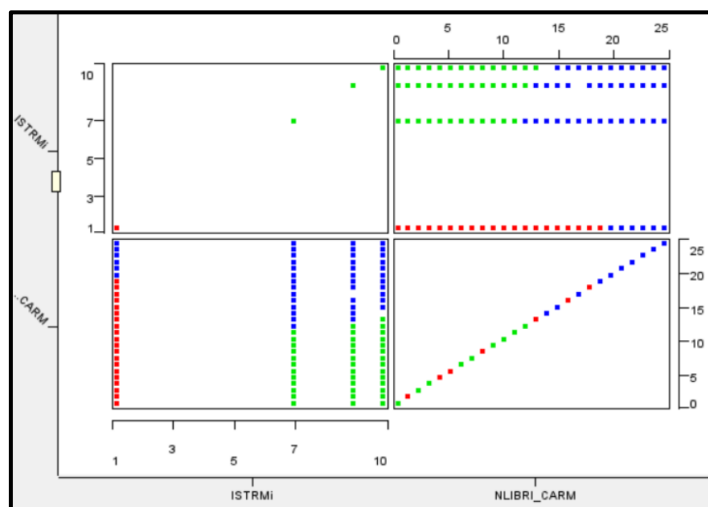
¹³⁶ L’unica modifica avvenuta nel nodo è: *number (integer)* → *remove row*.

¹³⁷ Nel nodo si è utilizzata l’impostazione: *interquartile range multiplier* pari 5.0. Inoltre, tutti i numeri anomali sopra la distribuzione sono stati rimossi tramite l’impostazione: *outliers above upper bound*. Un “*escamotage*” atto per rimuovere tutte quelle celle nel grado di istruzione con il numero pari a 99 che per l’ISTAT significa “dato non disponibile”.

¹³⁸ Nel nodo è stata inserita l’impostazione: *min-max normalization* da 0.0 a 1.0.

generalmente inizializzati in maniera casuale oppure attraverso strategie progettate per migliorare la stabilità del risultato. Una volta definiti i centroidi iniziali, ogni valore nel dataset viene assegnato al *cluster* il cui centroide risulta più vicino, utilizzando la distanza euclidea come misura di similarità. Questa fase di assegnazione permette di costruire una prima partizione dei dati sulla base della prossimità nello spazio delle caratteristiche. In seguito, per ciascun *cluster*, la posizione del centroide viene aggiornata calcolando la media delle coordinate di tutte le osservazioni assegnate al gruppo. Questo aggiornamento riflette una stima più accurata della posizione centrale del cluster rispetto all'iterazione precedente. Le fasi di assegnazione delle osservazioni e di ricalcolo dei centroidi vengono ripetute in modo iterativo, fino al raggiungimento di un criterio di arresto.¹³⁹ Il nodo successivo è il *Denormalizer* utile per riportare i risultati matematici ottenuti sulla scala originale delle variabili.¹⁴⁰ Questo passaggio non influisce sull'esito, ma risulta utile per facilitare l'interpretazione dei *cluster*. In ultimo, il nodo *Color Manager* consente di modificare a piacimento i colori dei *cluster*, mentre il nodo *Scatter Matrix* di visualizzare graficamente le informazioni costruite.¹⁴¹

Figura 11- Risultato grafico del modello in KNIME. Il gruppo blu dei “lettori forti” rispetto ai libri cartacei si presenta in tutti i livelli di istruzione.



Autore: Lorenzo Tiari

Il modello basato sull'algoritmo *K-Means* mostra, in particolare nel terzo quadrante in basso a sinistra, che i 3 gruppi di cittadini sono caratterizzati da distribuzioni molto simili rispetto al

¹³⁹ Nel nodo sono stati usati 3 *cluster*, un numero massimo di iterazioni pari a 300, togliendo l'impostazione: *use static random seed* = 0.

¹⁴⁰ Il nodo non è stato modificato.

¹⁴¹ In questo ultimo nodo si sono scelte le impostazioni: *no. of rows display* pari a 50.000 e l'impostazione: *ignore columns with more nominal values than* : 100.000, in modo da visualizzare tutte le persone.

numero di libri letti. Ciò suggerisce che l'algoritmo faticò a individuare una segmentazione efficace, come se emergesse implicitamente che il livello di istruzione risulti poco discriminante nel determinare il numero di libri cartacei letti. Nello specifico, il gruppo identificato in rosso comprende prevalentemente le persone in possesso di almeno un titolo di laurea; il gruppo verde raccoglie soggetti che leggono pochi libri e non possiedono un titolo di studio universitario. Infine, il gruppo blu rappresenta i cosiddetti "lettori forti" distribuiti su tutti i livelli di istruzione. Di conseguenza, si è aggiunto in seguito il nodo *Excel Writer*. Questo nodo consente di mettere i dati dei cittadini in cui sono state effettuate le operazioni di pulizia su un nuovo foglio Excel in modo da fare statistiche a mano prima di coinvolgere il modello matematico.

L'analisi che risulta è interessante. In primo luogo, a causa di non risposte o celle vuote nei questionari ISTAT si è potuto analizzare solamente il 38% dei cittadini (pari a 15.850 su 41.750 persone), mostrando che sistematizzare con l'elaborazione informatizzata la complessità sociale risulta difficile. Rispetto alle persone analizzate sono il 32.3% che al massimo hanno la licenza media, quasi il 40% possiede il diploma, mentre quasi il 28% possiede un titolo di laurea o post-laurea. Le persone con il più alto livello di studi che leggono almeno 1 libro al mese cartaceo ammontano solo a poco più del 13% rispetto al loro totale relativo (582 su 4.419), di conseguenza l'87% non ha costanti abitudini di lettura. Le persone con un diploma che leggono almeno 1 libro al mese cartaceo ammontano al poco più del 10% (637 su 6.298), di conseguenza quasi il 90% non sono considerabili come lettori costanti. Le persone con un titolo di studi di licenza media o inferiore leggono che leggono almeno 1 libro cartaceo al mese risultano circa al 11% (568 su 5.133). Inoltre, non viene considerata la qualità dei libri letti, editori come la Bollati Boringhieri la cui gran parte dei libri è un misto tra divulgazione scientifica e idee teoriche scritte da scienziati non può essere paragonata come simile a collane più "commerciali" come la Mondadori che ha tematiche più vaste fino ad arrivare all'intrattenimento. Al contrario, se analizziamo i lettori "molto forti", per esempio prendendo un numero uguale o maggiore di 20 libri l'anno allora troviamo per i laureati il 7% (311 su 4.419), per i diplomati il 5% (344 su 6.298), per i titoli di studio più bassi poco sopra il 6% (323 su 5.133). Di conseguenza, tutte le persone analizzate rispetto al totale assoluto che leggono molti libri cartacei sono pari a poco più del 6% (978 su 15.850).

La lettura rappresenta da sempre uno dei principali strumenti di costruzione del pensiero critico, di trasmissione culturale e di coesione sociale. Numerosi studi in ambito neurologico, sociologico e pedagogico evidenziano come le pratiche di lettura non svolgano un ruolo meramente ricreativo, ma contribuiscano in modo significativo allo sviluppo delle competenze cognitive e linguistiche del cittadino. In questo senso, il progressivo calo delle abitudini di lettura osservato negli ultimi decenni soprattutto nei contesti occidentali solleva interrogativi rilevanti rispetto alle

trasformazioni culturali e sociali in atto. Il declino della lettura non può essere interpretato in modo semplicistico come unica causa diretta di un presunto “decadimento” della società, ma esso può essere considerato un indicatore di un più ampio mutamento nei processi di produzione e fruizione del sapere.

La riduzione del tempo dedicato alla lettura approfondita è spesso associata alla diffusione di forme di consumo informativo rapide, frammentarie e fortemente mediate da tecnologie digitali che privilegiano l'immediatezza rispetto alla riflessione critica. Questo cambiamento può incidere sulla capacità collettiva di analisi complessa, di comprensione dei fenomeni sociali e di partecipazione consapevole alla vita democratica.¹⁴² Come si è evidenziato durante il capitolo, le neuroscienze forniscono elementi utili per comprendere l'importanza della lettura. Studi sul funzionamento cerebrale mostrano come la lettura prolungata e concentrata attivi reti neurali complesse coinvolte nella memoria, nell'attenzione sostenuta e nell'elaborazione simbolica. L'esercizio regolare di queste funzioni è stato associato a una maggiore riserva cognitiva, concetto che indica la capacità del cervello di compensare il declino funzionale legato all'invecchiamento o a patologie neurodegenerative. In questa prospettiva, la diminuzione delle pratiche di lettura porterà nel lungo periodo a un impoverimento delle stimolazioni cognitive a livello individuale e collettivo, creando un'aridità culturale e una minore capacità di saper controllare i comportamenti durante situazioni complesse. Questi fenomeni sono già evidenti nelle relazioni lavorative, intergenerazionali e scolastiche. Le abitudini di lettura diventano una variabile determinante per stabilire un possibile degrado sociale, concetti che erano già stati osservati nella caotica vita romana vissuta da Marco Tullio Cicerone.¹⁴³ Per concludere, nonostante l'analisi del dataset sia evidentemente viziata e si siano riscontrati i problemi tipici nell'analizzare grandi quantitativi di dati in ambito sociale, la tendenza generale rispecchia indicativamente i risultati di molte indagini effettuate nelle nazioni più avanzate. Analizzando il dataset, le persone con il più alto titolo di studi che leggono non più di 5 libri cartacei all'anno sono quasi il 67% (2.957 su 4.419).

3.3 Allucinazioni e inspiegabilità come algoritmi naturali

Quando si descrivono fenomeni associati al sistema nervoso come creatività, intenzionalità, emozioni e così via è sempre necessario ricordare l'enorme differenza nel formalizzare l'intelligenza artificiale o la più complessa robotica rispetto anche ad una singola cellula biologica. Per esempio, a livello formale un neurone artificiale applicabile al *deep learning* può essere descritto in questo modo:

¹⁴² La lettura stimola varie aree tra cui la corteccia prefrontale che è fondamentale per l'autocontrollo, il giudizio e la regolazione delle emozioni.

¹⁴³ *Cicerone: l'arte di comunicare*, a cura di Paolo Marsich, Mondadori S.p.A., 2016, p.6

$$F\left(\sum_{i=1}^n w_i \times x_i + b\right)$$

In cui la x rappresenta i dati in ingresso, la w sono i pesi per ciascun input e b sono gli errori sistematici. F rappresenta la soglia di attivazione che può essere intesa come una disuguaglianza (> 0). Per fare un esempio, se i pesi e i dati sono determinati da una moneta in cui le due facce corrispondono a -1 e 1, e i *bias* sono sempre un numero arbitrario come potrebbe essere 0, l'operazione va a determinare se il segnale si trasmette o meno in maniera binaria. Chiaramente un neurone artificiale che dipende dal lancio di una moneta non potrebbe sopravvivere nel mondo informatico, ma sono da questi semplici principi che poi si sviluppano i modelli artificiali moderni. Al contrario, una cellula eucariote vegetale possiede parti fondamentali come la parete cellulare che serve da protezione e sostegno, citoplasma in cui avvengono le reazioni chimiche, la membrana plasmatica, l'apparato di Golgi, cloroplasto, microtubuli, plasmodesma, il nucleo che a sua volta è suddiviso in membrana nucleare, poro, nucleolo e così via.¹⁴⁴ Se il neurone artificiale ha pochi elementi semplificati in singole equazioni, una minuscola cellula ha migliaia di parti che svolgono a loro volta più funzioni dinamiche non lineari che dipendono fortemente da un ambiente a sua volta estremamente dinamico. Insomma, la differenza di formalizzazione fra artificiale e biologico presenta una complessità tale da rendere impossibile perfino determinare una scala di paragone. In questo paragrafo si analizzeranno due fenomeni tipici nei modelli artificiali come allucinazione e inspiegabilità, concetti fondamentali per poi andare a introdurre nel successivo capitolo la piattaforma Undermind che funge da modello di assistenza nelle ricerche per gli scienziati. Mentre si descriveranno questi concetti non si darà molta enfasi sui loro svantaggi, già ampiamente trattati nella letteratura scientifica, ma si analizzeranno i presupposti positivi tentando di capire come questi fenomeni, se ricercati "volutamente", possano essere utili nello sviluppo di modelli artificiali.

In primo luogo, per allucinazione intendiamo un'azione atta a completare la realtà attraverso informazioni incomplete, errate oppure troppo ridondanti. Per esempio, molti artisti nel corso della storia utilizzavano sostanze chimiche come "input" per alterare il loro modello del mondo consentendo di trovare schemi nascosti dalla realtà, oppure commettere volutamente errori per produrre nuove soluzioni.¹⁴⁵ Un meccanismo "analogo" è osservabile nei modelli artificiali, in particolar modo nei modelli generativi basati sull'apprendimento automatico. Questi sistemi a livello chimico non possiedono la comprensione semantica del mondo come il cervello né un

¹⁴⁴ *Elementi di biologia e genetica*, Piccin Nuova Libreria S.p.A., AA.VV, 2019, p. 67

¹⁴⁵ In merito a questo una disciplina parte delle scienze mediche è la fitoterapia che studia piante medicinali o i rispettivi derivati per prevenire e curare disturbi. Peraltro, era prassi comune l'utilizzo di piante e risorse naturali come elementi rituali o creativi nelle civiltà antiche dell'America Latina.

accesso diretto alla verità dei fatti, essi operano tramite matrici e vettori, la risposta più probabile sulla base di algoritmi matematici e statistici appresi durante la fase di addestramento. Quando l'informazione disponibile è incompleta, ambigua o al di fuori del dominio dei dati appresi nella fase di *training*, il sistema tende comunque a produrre un output coerente per il nostro sistema nervoso dal punto di vista linguistico. Questo comportamento si manifesta nella generazione di informazioni plausibili ma false, con citazioni inesistenti, dati errati o spiegazioni apparentemente logiche, ma prive di un riscontro. La radice comune del cervello e dei modelli generativi risiede nella natura predittiva di entrambi i sistemi. Sia il cervello sia le nuove tecniche dell'intelligenza artificiale, e in particolare i modelli generativi, privilegiano la coerenza rispetto alla verifica diretta con la realtà. In altre parole, non cercano una verità in senso assoluto (nel secondo capitolo si è già nota la complessità matematica di un possibile filtro di verità), ma una ricostruzione verosimile e funzionale del mondo. Dal punto di vista degli svantaggi, l'allucinazione rappresenta un limite rilevante per entrambi. Nel sistema nervoso le allucinazioni implicano compromissioni dei neuroni con ricadute sulla cognizione, sul comportamento e sulla socialità. Nell'intelligenza artificiale produrre informazioni errate può minare la stabilità logica della programmazione, la falsità nelle metriche di validazione utilizzate e minano l'affidabilità del modello percepita dagli utenti. Insomma, l'allucinazione provoca errori decisionali, rischi e perdita di fiducia.

Tuttavia, limitarsi a considerare l'allucinazione esclusivamente come un difetto sarebbe riduttivo. Il primo caso d'esempio sono grandissimi scienziati affetti da qualche forma di disturbo o autismo, che attraverso comportamenti anomali rispetto alla norma sociale ci hanno permesso di vivere nella comodità della vita quotidiana. Nel cervello umano la capacità di generare informazioni non direttamente ancorate alla realtà immediata è alla base dell'immaginazione, della creatività e del pensiero simbolico, utile ad esempio nella logica formale. Processi mentali che derivano dallo "sognare ad occhi aperti" e la visualizzazione mentale condividono meccanismi neurali vicini a quelli che producono soluzioni patologiche.¹⁴⁶ Un altro esempio sono persone che hanno una forte sinestesia, ovvero quel fenomeno percettivo in cui i sensi possano integrarsi a vicenda creando associazioni spontanee come percepire i numeri come colori o sapori.¹⁴⁷ Analogamente, nell'intelligenza artificiale completare l'informazione mancante richiede una notevole flessibilità generativa. Proprio questa caratteristica rende possibili nuovi testi, immagini o musica che in assenza "dell'errore" non sarebbe mai risultati dall'output del modello. Un modello artificiale "perfetto" completamente privo di errori sarebbe statico, incapace di "generalizzare" e dare risposte fortemente limitate rispetto ai dati d'ingresso dell'utente. Essenziale ribadire che, sebbene i due fenomeni a livello di output possano sembrare simili, nel

¹⁴⁶ David E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016, pp. 274-275-276

¹⁴⁷ Ivi, p. 257

sistema nervoso l'esperienza è neuropsicologica mentre nell'intelligenza artificiale è statistica.¹⁴⁸ Infine, se il sistema nervoso presenta molti meccanismi di autoregolazione biologici, l'intelligenza artificiale dipende da vincoli di correzione esterni come la qualità dei dati o l'architettura del modello determinata dagli sviluppatori.

L'inspiegabilità è uno dei tratti più problematici e controversi nei sistemi complessi. Il sistema nervoso e i modelli artificiali condividono l'enorme problema tra bilanciare la complessità nell'elaborare dati a discapito di una completa spiegazione dei processi che da un input conducono ad un determinato output. Questa opacità non è semplicemente un limite che è sempre esistito in natura, ma un fenomeno che esprime la caoticità e la caratteristica spesso non lineare nella produzione di un'informazione. Come già visto nel secondo capitolo, l'*explainability* del cervello deriva dal fatto che i processi cognitivi si sono sviluppati in milioni di anni a seguito di svariate pressioni evolutive; pertanto, quando si parla di concetti come "creatività" si possono delineare aree precise in cui il fenomeno avviene, ma non ogni singolo punto nella catena degli eventi neurali. Funzioni come linguaggio, giudizio o decisioni emergono come attività pluralistica di reti neurali distribuite caratterizzate da plasticità sinaptica. Sebbene le neuroscienze trovino correlati neurali di diverse funzioni cognitive, il comportamento osservabile è ancora largamente non spiegabile. In altre parole, i neuroscienziati stanno scoprendo sempre di più il *dove* e *quando* avviene un processo, ma non *perché* insieme di neuroni producano un esito piuttosto che un altro. Analogamente è l'intelligenza artificiale che manifesta per utenti forme di inspiegabilità a causa di numerosi algoritmi matematici e lunghissimi codici di programmazione. I modelli artificiali apprendono tramite milioni o miliardi di parametri, le cui interazioni non sono direttamente osservabili da parte di un *Homo Sapiens*. I modelli artificiali stanno diventando così tanto complessi che richiedono intere squadre di scienziati che si occupino di specifici problemi, ed è normale che nell'ambito industriale ci siano molte discussioni sulle ripercussioni giuridiche e un continuo via e vai di scienziati che scelgono un progetto piuttosto che un altro a causa della mancanza di conoscenza di che output possano ottenere da un modello. Insomma, l'output non è riconducibile chiaramente a singole componenti, ma dall'interazione complessiva di neuroni biologici o parametri. Effettivamente, dal punto di vista degli svantaggi si nota che l'impossibilità di determinare output comporta una perdita di controllo e l'impossibilità di spiegare chiaramente i processi e interromperli. In settori critici come la medicina, la finanza o la giustizia l'uso di sistemi che operano tramite decisioni opache comporta errori che possono portare alla morte degli individui. L'assenza di spiegazioni rende inoltre impossibile determinare i *bias* nella valutazione matematica, ponendo gravissime questioni etiche e rendendo sempre più necessaria una

¹⁴⁸ Questo è l'errore di Turing, visto in ambito informatico come precursore del "funzionalismo". Turing riteneva un sistema intelligente in base al risultato senza guardare precisamente allo stato chimico interno. Descritto nel libro: A. Turing, *Macchine calcolatrici e intelligenza*, a cura di Diego Marconi, Einaudi, 2025.

regolamentazione che sia stringente. Tuttavia, come nel caso dell'allucinazione, anche un'inspiegabilità "buona" non deve essere necessariamente esclusa in sperimentazioni fortemente controllate ed essere vista solo come un difetto. Nel cervello, l'assenza di processi completamente noti consente una notevole flessibilità cognitiva, un'elevata capacità di adattamento e la possibilità di operare in condizioni d'incertezza. Molte abilità cognitive, come l'intuizione, dipendono da procedure non cosce e spiegabili passo per passo, altrimenti il sistema sarebbe terribilmente lento e richiederebbe eccessiva energia calorica, ma si basano su processi impliciti. Allo stesso modo, l'intelligenza artificiale tramite l'*explainability* e i modelli *black box* sono il prezzo da pagare per prestazioni elevate in compiti complessi. I modelli artificiali, rispetto ad una riduzione della complessità, riducono le capacità di "generalizzazione" rischiando di compromettere prestazioni di alto livello. Un confronto tra questi due sistemi mette in luce, dunque, la tensione tra efficienza e comprensibilità. Entrambi i sistemi mostrano che l'inspiegabilità non è un'anomalia tecnica, ma una conseguenza naturale nell'elaborazione delle informazioni del mondo. La differenza fondamentale risiede nel fatto che, mentre l'inspiegabilità del cervello è un fatto biologico con il quale l'*Homo Sapiens* convive da sempre, quello dell'intelligenza artificiale è una sfida progettuale ed etica rispetto alle soluzioni trovate dagli scienziati, come lo sviluppo di modelli interpretabili o di tecniche di spiegazione a posteriori da parte degli utenti che lavorano nel settore informatico.

In conclusione, fenomeni come allucinazioni e inspiegabilità sono caratteristiche strutturali di "osservatori" capaci di elaborare dati e produrre grandissime quantità di informazioni diverse. Se controllati adeguatamente consentiranno di migliorare sia le scienze informatiche che quelle naturali in maniera bidirezionale. Ad esempio, si possono proporre esperimenti simulativi che vadano a forzare queste caratteristiche in un modello artificiale per provare a comprendere se alcune cose avvengano anche nel sistema nervoso. Oppure, lo studio migliore del cervello grazie a strumentazioni sempre più avanzate potrebbe consentire di spiegare quali tecniche siano necessarie per evitare che i modelli applicati in settori sensibili possano provocare possibili disastri sociali.

Capitolo 4

Modellare il linguaggio scientifico

Il linguaggio naturale e la matematica condividono un obiettivo fondamentale: ridurre l'ambiguità interpretativa attraverso l'uso di strutture formali e regole condivise. Se il linguaggio naturale è storicamente e culturalmente situato, la matematica aspira a una validità universale. Tuttavia, questa distinzione non implica una separazione netta, bensì una profonda interdipendenza. La logica formale, come l'algebra di Boole, si basa su operatori che trovano un'evidente corrispondenza nelle congiunzioni e negazioni linguistiche fondamentali come *and*, *or* e *not*. Questi operatori non sono meri simboli astratti, ma rappresentano formalizzazioni di relazioni semantiche già presenti nel linguaggio naturale che vengono rese esplicite e sottoposte a regole di inferenza rigorose.

Il legame tra linguaggio e matematica emerge con particolare evidenza anche nel contesto della statistica e dell'informatica. La statistica utilizza il linguaggio matematico per descrivere fenomeni complessi e probabilistici, ma richiede al contempo una cornice linguistica per l'interpretazione dei dati e la formulazione di ipotesi. Analogamente, i linguaggi di programmazione traducono istruzioni linguistiche in operazioni logico-matematiche eseguibili, dimostrando come la struttura del linguaggio possa essere codificata in regole computazionali precise. Di conseguenza, il linguaggio e la scrittura rappresentano un'enorme quantità di dati processabili dagli algoritmi.

Il linguaggio scientifico rappresenta una forma altamente specializzata del linguaggio naturale, sviluppata con l'obiettivo di trasmettere conoscenza in modo preciso, verificabile e non ambiguo. A differenza di alcune varietà linguistiche come potrebbe essere l'italiano regionale, quest'ultimo si caratterizza per un elevato grado di formalità, per l'uso continuo di termini ad alto grado di specialismo e per strutture sintattiche ricorrenti che mirano a ridurre possibili fraintendimenti. Tra le caratteristiche distintive del linguaggio scientifico si annoverano l'alta densità informativa, forme impersonali e presenza esplicita di dati, ipotesi e risultati sperimentali. Inoltre, il significato di molti termini è fortemente dipendente dal contesto disciplinare in cui vengono utilizzati. Parole comuni possono assumere eccezioni rigorose e formalizzate in ambito scientifico mentre numerosi concetti sono espressi tramite simboli, formule matematiche o abbreviazioni standardizzate.¹⁴⁹ Questa combinazione di elementi rende il linguaggio scientifico adatto alla comunicazione tra gli esperti, ma allo stesso tempo complesso da trattare nella computazione.

¹⁴⁹ D. Vellutino, *L'italiano istituzionale per la comunicazione pubblica*, Il Mulino, 2018, pp. 66-67

In primo luogo, il vocabolario è fatto da continui neologismi o espressioni che hanno bassa probabilità di esistere nel linguaggio comune. Molti termini assumono significati diversi a seconda del dominio disciplinare, rendendo problematica la generalizzazione dei vettori appresi. I modelli artificiali generativi faticano a catturare queste sfumature semantiche quando i dati disponibili sono limitati o fortemente sbilanciati. Ad esempio, il termine “osservatore” assume di volta in volta significati diversi nella Scienza dei Dati, fisica fino alle scienze sociali.¹⁵⁰ Un secondo elemento è rappresentato dalla struttura sintattica del linguaggio scientifico. I testi scientifici presentano frasi lunghe, articolate e ricche di subordinazione, spesso costruite per condensare una grande quantità di informazione in un singolo enunciato. L’uso frequente di forme impersonali, riferimenti simbolici e nominalizzazioni riduce la trasparenza vettoriale tra soggetti, azioni e oggetti, rendendo più difficile l’analisi sintattica automatica e l’identificazione delle relazioni semantiche sottostanti.

Ulteriore aspetto critico è la consequenzialità degli enunciati. Il significato delle frasi è spesso comprensibile solo alla luce di definizioni precedenti, assunzioni teoriche o risultati già presentati nel testo. Ciò introduce dipendenze matematiche a lungo raggio che richiedono modelli artificiali costosi in grado di mantenere e integrare informazioni distribuite su ampie porzioni del documento.

La presenza di elementi non puramente linguistici costituisce un’ulteriore sfida. Formule matematiche, tabelle e riferimenti bibliografici sono parte integrante nel dibattito e contribuiscono in modo sostanziale al significato complessivo del testo. Ad esempio, si potrebbe discutere di una possibile equazione che catturi il termine “intelligenza” negli animali. Citare come base di partenza le considerazioni di Turing (informatica) è diverso rispetto a quelle di Vallortigara (neuroetologia). Questi elementi spesso non sono evidenti e richiedono algoritmi matematici ibridi che siano in grado di cogliere queste sfumature.

Alla luce di queste difficoltà, l’impiego dei modelli di *deep learning* introduce a sua volta una serie di problemi specifici. Uno dei principali riguarda la dipendenza dalla quantità e dalla qualità dei dati di addestramento. I modelli richiedono grandi volumi di dati, ma nel dominio scientifico questi dati sono spesso costosi da ottenere, limitati a specifiche discipline o soggetti a vincoli di accesso. I modelli artificiali presentano concettualmente lo stesso problema di un bambino che vuole apprendere: se la biblioteca è chiusa o i libri non sono disponibili, non si riescono a migliorare le proprie capacità. Facendo fatica ad ottenere dati di alta qualità, anche se nel mondo moderno il problema è discretamente sopperito dal numero enorme di pubblicazioni scientifiche gratuite, si pone il problema della generalizzazione. Le caratteristiche del linguaggio scientifico

¹⁵⁰ Ad esempio, il termine osservatore nella tesi è stato descritto come un’entità con certe caratteristiche, nella fisica si può intendere uno strumento di rilevazione o nelle scienze sociali come un ricercatore.

mettono in evidenza come la sua modellazione computazionale non possa essere affrontata in modo matematicamente semplicistico con tecniche uguali a quelle utilizzate nel linguaggio quotidiano.

4.1 Database e intelligenza artificiale applicati nella scienza

I database scientifici rappresentano una componente fondamentale dell'ecosistema della ricerca contemporanea, costituendo il principale mezzo attraverso cui la conoscenza scientifica viene raccolta, organizzata e resa accessibile. Essi aggregano grandi quantità di testi scientifici come articoli, conferenze, report e in alcuni casi metadati associati agli esperimenti, offrendo una base strutturata per lo scienziato. Dal punto di vista funzionale assolvono a molteplici scopi. Innanzitutto, conservano la produzione scientifica garantendo tracciabilità nel tempo dei risultati e offrendo statistiche sull'andamento generale. Per esempio, il *database* arXiv, fondato nel 1991 e mantenuto dalla Cornell Tech university, tramite parole chiave consente di capire quanti articoli sono stati digitalizzati per uno specifico anno. Inoltre, i database facilitano l'accesso nella ricerca di informazioni attraverso sistemi di indicizzazioni che consentono di interrogare i contenuti rispetto all'autore, parole chiavi o periodi temporali. Dal punto di vista della Scienza dei Dati, assumono un ruolo ancora più centrale poiché forniscono grandi quantitativi di dati su cui vengono addestrati e valutati i modelli. In questo senso, per i modelli artificiali applicati all'ambito scientifico i database rappresentano le fondamenta su cui si fonda la computazione.

Nonostante i numerosi vantaggi, il loro utilizzo presenta dei limiti. Uno dei principali riguarda l'estrema eterogeneità. Attualmente in Italia ci sono almeno 380 settori scientifico-disciplinari divisi in 14 aree.¹⁵¹ Come già detto, i testi provengono da settori diversi con convenzioni, simboli e termini differenti su vari piani di formalizzazione. Se da un lato arricchisce il patrimonio informativo, dall'altro complica notevolmente la ricerca di definizioni univoche nei modelli generativi. Un ulteriore problema è rappresentato dalla completezza dei metadati. Le informazioni non sempre sono standardizzate e possono contenere errori o ambiguità. Dal punto di vista computazionale limita l'affidabilità delle analisi automatiche e introduce rumore nei processi di addestramento dei modelli.

Oltre alle varie limitazioni, i modelli sono sottoposti a criticità simili a quelle che avvengono nel mondo naturale. In questo senso è la macchina costretta a difendersi dagli uomini che, dovendo continuare a produrre ricerche fondamentali per l'industria, hanno inquinato l'ambiente scientifico. Il plagio costituisce una violazione diretta poiché implica l'appropriazione indebita di idee e risultati altrui senza un riconoscimento adeguato delle fonti. Può manifestarsi in forme

¹⁵¹ Nelle scienze è prassi comune ridefinire continuamente i vincoli e le differenze fra le varie discipline sia per questioni teoriche sia per precise strategie di finanziamento.

esplicitate come la copia letterale di intere porzioni di testo oppure in forme più sottili come la riformulazione superficiale di contenuti preesistenti. Non solo costringe gli scienziati meritevoli a trovare strategie più esclusive per evitare possibili plagii (come abbonamenti a pagamento o limitazioni nella visione), ma altera il processo cumulativo della conoscenza creando l'illusione di un avanzamento che in realtà non avviene nei dipartimenti.

Successivamente al plagio, la diretta conseguenza sono le duplicazioni inutili degli articoli scientifici in database differenti. La pressione legata alla valutazione quantitativa per determinare la capacità di uno scienziato spinge a “gonfiare” il volume della letteratura disponibile. Particolarmente evidente nel settore della *Data Science*, in cui molte ricerche non toccano la teoria di fondo, ma discutono di fenomeni trascurabili nell'ambito dell'intelligenza artificiale. I *paper* non sono più testi che possono cambiare la visione del mondo in maniera rapida, ma diventano semplici notizie online consumate dall'*Homo Scientificus*.¹⁵²

Un problema ancora più insidioso è la presenza di testi privi di valenza. Questi testi possono apparire formalmente corretti adottando un lessico e una struttura adeguata, ma risultano privi di un qualsiasi significato. In molti casi, questi contenuti derivano dagli stessi modelli generativi che dovrebbero facilitare gli scienziati durante le loro ricerche.

A queste criticità si aggiungono il fenomeno delle citazioni inesistenti, di comodo o non verificabili che minano direttamente il principio di tracciabilità. Citare lavori che non esistono, che non hanno senso o che non supportano le affermazioni introduce una catena di errori che può propagarsi rapidamente nella letteratura. In ambito computazionale questo fenomeno è particolarmente problematico, molti sistemi automatici utilizzano misure per calcolare la rilevanza, l'autorevolezza e l'impatto scientifico. I modelli ancorati direttamente a questi database rischiano di rafforzare schemi ripetuti, attribuire rilevanza a contenuti privi di senso e riprodurre *bias* presenti nei dati di addestramento. L'individuazione automatica di queste anomalie non è un compito banale. Molte riviste scientifiche autorevoli hanno sempre più difficoltà nell'accettare articoli scientifici. Il plagio può essere mascherato da parafrasi complesse, le duplicazioni possono presentare piccole variazioni strutturali e i testi privi di senso apparire grammaticalmente sensati.

Dal punto di vista dei modelli di *deep learning*, in particolar modo dei modelli che si occupano nell'elaborare il linguaggio naturale, questa varietà linguistica pone delle sfide specifiche. I modelli generativi rappresentano una classe del *deep learning* progettati per apprendere la distribuzione dei dati in modo da poter generare nuove risposte simili a quelli osservati nel dataset

¹⁵² Un esempio è l'articolo “*Attention is all you need*” del 2017 prodotto da vari scienziati che lavoravano in Google. L'articolo introduce un nuovo modello di *deep learning*, conosciuto come *Transformer*, che porterà all'esplosione dei modelli generativi commerciali e l'attenzione mondiale sull'intelligenza artificiale.

di addestramento.¹⁵³ A differenza dei modelli costruiti per distinguere tra categorie (utili nella robotica e nella *computer vision*), questi modelli mirano a “catturare” la struttura sottostante dei dati in modo profondo e contestuale. Alla base di questi modelli vi è il concetto di probabilità, se un modello può stimare la distribuzione di un dato allora è possibile campionare nuovi punti di quella distribuzione.¹⁵⁴ In molti casi, specialmente quando i dati sono complessi e diversificati (come audio, testo e immagini) questa distribuzione diventa difficile da rappresentare matematicamente. Per superare questa difficoltà i modelli generativi si avvalgono di complessi algoritmi matematici. I *GPT*, acronimo di *Generative Pre-trained Transformer*, sono una famiglia di modelli di linguaggio progettati per comprendere il linguaggio naturale. Il loro obiettivo è prevedere la sequenza della parola successiva in una sequenza di testo, ma da questo compito apparentemente semplice emerge la capacità di svolgere attività linguistiche complesse. Il termine *Pre-trained* indica che il modello viene inizialmente addestrato su una grande quantità di testi provenienti da libri, articoli o conversazioni online. Durante questa fase le parole vengono spezzate in *token*, diventando numeri, e si fanno calcoli vettoriali per apprendere strutture grammaticali e conoscenze generali sul mondo. Successivamente il modello può essere adattato a compiti specifici tramite il *fine-tuning*, ovvero apprendere nuovi dati per migliorare la qualità delle risposte. L’architettura principale è il *Transformer* che si distingue per l’algoritmo *self-attention*, che consente di valutare l’importanza relativa delle parole in una frase rispetto alle altre. Questo approccio permette di catturare relazioni a lungo raggio nel testo in modo più efficace rispetto ad altri modelli.¹⁵⁵ Inoltre, *GPT* utilizza la variante *decoder-only*, che serve ad ottimizzare il modello in funzione della generazione di nuove risposte.¹⁵⁶ In sintesi, i modelli tramite vettori, formule matematiche e probabilità riescono a generare nuovi risultati. Come già visto all’inizio del capitolo nulla di eccezionale considerando l’intreccio tra matematica, statistica, informatica e linguistica.

Volendo riprendere la definizione convenzionale di osservatore del capitolo 2, questi modelli rispecchiano gli stessi problemi naturali. I modelli generativi utilizzati quotidianamente sono molto costosi in termini di energia a causa dei materiali e delle risorse spese come acqua ed elettricità.¹⁵⁷ Sul lungo periodo questo continuo miglioramento porterà a dover rivedere gli algoritmi per semplificarli, riadattarli o adottando altre strategie come ripensare completamente alla loro locazione geografica. Possiedono un’infrastruttura di base come *data center* che consente

¹⁵³ *History, Development and principles of large language models- An introduction survey*, AA.VV., 2024, para. Factors propelling large language models. <https://arxiv.org/pdf/2402.06853> (attivo)

¹⁵⁴ Ivi, para. History of LLMs

¹⁵⁵ *The architecture of language: Understanding the mechanics behind LLMs*, AA. VV., 2025, para. Self-attention mechanism. https://www.researchgate.net/publication/387752968_ (attivo)

¹⁵⁶ Ivi, para. Decoder-only transformers

¹⁵⁷ *Green AI*, AA.VV., 2019, para. Red AI. <https://arxiv.org/pdf/1907.10597> (attivo)

di conservare e gestire l'informazione. Infine, hanno limiti nel reperire e rielaborare i dati, contribuendo a fenomeni come errori ricorrenti, *bias* e allucinazioni.¹⁵⁸ Di conseguenza, l'idea che questi modelli possano essere strumenti oggettivi e perfettamente affidabili è diffusa, ma fuorviante. Un'analisi approfondita mostra che l'oggettività, intesa come assenza di prospettiva o interpretazioni, è difficilmente raggiungibile. Questo non vale solo per contenuti opinabili come l'ambito sociale, ma anche quando i modelli generano o sintetizzano informazioni di carattere scientifico. I modelli apprendono dati basati su testi umani e anche nella letteratura scientifica esistono scuole di pensiero, ipotesi controverse, risultati non replicabili. Nessun dataset può rappresentare l'intera produzione umana e i dati, così come vengono intesi nella tesi, sono entità fragili. Possono venire distrutti, alterati oppure mai scoperti. Quanti libri di testo fondamentali sono andati persi e quanti esperimenti non sono mai stati raccontati dalla storia.

4.2 Struttura dell'esperimento e metriche utilizzate

Le richieste fatte tramite linguaggio naturale, detti *prompt*, sono una nuova modalità d'interazione tra essere umano e macchina. Un *prompt* è un input testuale che l'utente fornisce al modello per istruirlo matematicamente su quale compito svolgere o su quale contenuto deve generare. Può assumere la forma di una domanda, una richiesta o un comando strutturato. Tuttavia, ciò che spesso passa inosservato è la qualità della richiesta determinando in modo cruciale la qualità delle risposte dei modelli artificiali. Questo è il motivo per cui l'intelligenza artificiale è un amplificatore esterno delle capacità del sistema cognitivo, ed è per questo che è necessario "curare" il nostro sistema nervoso poiché si riflette sulle reti artificiali. Se la richiesta è vaga o ambigua l'output rischia di essere sbagliato o generico. Al contrario, una richiesta ben costruita può guidare il modello verso risposte più accurate.¹⁵⁹

In realtà, la questione è più sfumata. Un esempio potrebbe essere la richiesta: "Nucleare". Un *prompt* estremamente generico che può portare un modello artificiale a produrre una semplice sintesi su una precisa branca della fisica oppure ricercare informazioni generali sulla composizione del nucleo familiare in Italia. Tuttavia, può far capire la capacità di un modello, se ha dei *bias* di partenza oppure a quali database fa riferimento. Se si pone una richiesta più articolata come: "Scrivi una piccola relazione sui principali sviluppi e approcci tecnici applicati al nucleare citando le ricerche degli ultimi 5 anni.", si stanno dando specifici comandi restringendo le capacità di scelta. Da questa richiesta vengono prelevate le ricerche più recenti, ma non per questo migliori, e si costringe il modello a citare tutti gli strumenti applicati al

¹⁵⁸ *Data, power and bias in Artificial Intelligence*, AA.VV., 2020, para. The myth of objectivity.

<https://arxiv.org/pdf/2008.07341> (attivo)

¹⁵⁹ *A survey of prompt engineering methods in large language models for different NLP tasks*, AA.VV., 2024, para. Introduction. <https://arxiv.org/pdf/2407.12994> (attivo)

nucleare. Se da una parte offre risultati più precisi, dall'altra si è perso un'enorme quantità di dati che potrebbero essere stati utili. Per uno scienziato ripercorre a volta prospettive antiche può essere utile per proporre nuovi esperimenti o teorie, banalmente si potrebbe pensare all'Atomismo di Democrito. Questo filosofo greco, in assenza di complessi laboratori, pensò alla realtà come ad una suddivisione di atomi e la sua prospettiva è ancora oggi analizzata nelle facoltà di fisica.

Di conseguenza, verrà proposto l'esperimento sulla piattaforma Undermind, *software* per la ricerca scientifica potenziata dall'intelligenza artificiale tramite il modello GPT-4. Progettata per supportare ricercatori e dottorandi nello svolgimento delle ricerche bibliografiche o per avere risposte scientificamente precise alle varie domande. La piattaforma è stata fondata nel 2023 da due ricercatori di fisica del MIT con l'obiettivo di affrontare una difficoltà ben nota: trovare letteratura rilevante utile a domande complesse. L'idea è di supportare il ricercatore tramite un processo di replicazione di un gruppo sottostante di ricerca in modo rapido.¹⁶⁰ I database a cui fa riferimento è quello di Semantic Scholar che contiene milioni di articoli di tutte le discipline scientifiche. A sua volta questo database ne contiene molti altri come ArXiv, bioRxiv, Frontiers, Microsoft, The MIT press, The Royal Society e così via.¹⁶¹

Nei modelli generativi di grandi dimensioni risulta intrinsecamente complesso replicare in modo perfettamente identico un esperimento, a causa della natura probabilistica del processo di generazione testuale e della mole innumerevole di dati su cui i modelli sono addestrati. La produzione delle risposte non segue infatti un meccanismo deterministico, ma è influenzata da distribuzioni di probabilità che possono variare anche a parità di input, oltre che da aggiornamenti del modello e da condizioni contestuali non interamente controllabili nel tempo. Questa caratteristica rende problematica l'applicazione di protocolli sperimentali rigidamente riproducibili secondo i criteri tradizionali della sperimentazione quantitativa. Alla luce di tali considerazioni, si è ritenuto metodologicamente più appropriato adottare un approccio di tipo qualitativo finalizzato all'osservazione e all'analisi delle caratteristiche dal punto di vista utente del modello in condizioni controllate, ma interpretative. L'esperimento è stato condotto utilizzando una versione del modello accessibile tramite piano a pagamento, al fine di ridurre variabili legate a limitazioni funzionali o di accesso e garantire una maggiore stabilità nelle prestazioni osservate.

Il protocollo sperimentale ha previsto la formulazione di due richieste testuali intenzionalmente differenti per struttura, grado di specificità e carico informativo. La prima richiesta è stata progettata in modo deliberatamente incompleto, minimale e caratterizzata da un elevato livello di incertezza semantica. Questa impostazione ha avuto lo scopo di indagare il comportamento del

¹⁶⁰ T. Hartke e J. Ramette, *Benchmarking the Undermind Search Assistant*, 2024
<https://www.undermind.ai/whitepaper.pdf> (attivo)

¹⁶¹ <https://www.undermind.ai> (attivo)

modello in assenza di vincoli espliciti, osservando in particolare la sua capacità di gestire l'ambiguità, la lingua privilegiata nella generazione del testo, nonché le strategie di articolazione del ragionamento e la sequenzialità dei passaggi argomentativi adottati.

La seconda richiesta, al contrario, è stata costruita in modo altamente strutturato e privo di margini interpretativi, con l'obiettivo di ridurre al minimo l'ambiguità dell'input. In essa è stato mantenuto un *bias* di fondo, al fine di valutare la capacità del modello di riconoscere o incorporare presupposti non esplicitati e di verificarne l'influenza sulla correttezza e sulla coerenza della risposta prodotta. Questa scelta metodologica consente di analizzare non solo l'accuratezza informativa del modello, ma anche la sua sensibilità alle idee sottostanti alla formulazione della domanda.

Il confronto tra le risposte generate in risposta a queste due tipologie di richieste ha permesso di osservare la presenza di eventuali errori, incongruenze o distorsioni, nonché di individuare schemi ricorrenti nel comportamento del modello. Pur non aspirando a una generalizzazione statistica dei risultati, l'esperimento fornisce indicazioni rilevanti sul funzionamento generale del modello e sulle sue modalità di interazione con input linguistici di diversa natura, contribuendo a una valutazione critica delle sue potenzialità e dei suoi limiti dal punto di vista scientifico.

4.3 Analisi del modello Undermind

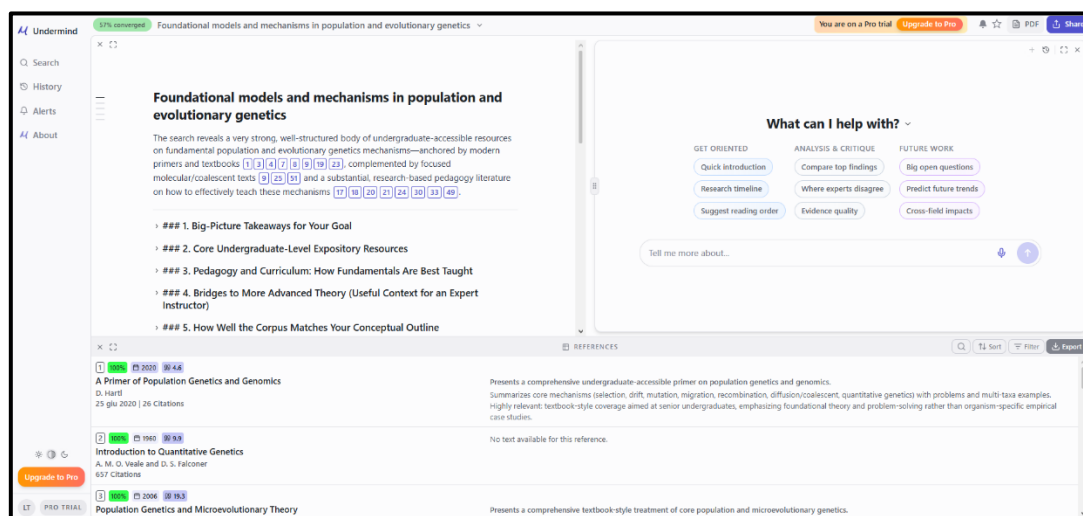
Il metodo di analisi adottato da Undermind si fonda su una struttura iterativa e progressiva di interrogazione, finalizzata all'approfondimento graduale dei risultati ottenuti. In una prima fase viene formulata una domanda iniziale dallo scienziato alla quale il sistema fornisce una risposta preliminare che costituisce la base dell'analisi successiva. La risposta non viene considerata come un esito conclusivo, bensì come un punto di partenza per ulteriori esplorazioni. Successivamente, viene posta una prima domanda di approfondimento dal modello progettata per chiarire e precisare gli elementi emersi nella risposta iniziale. Questa fase consente di valutare la capacità del modello di mantenere coerenza argomentativa, di sviluppare il contenuto informativo e di integrare nuove richieste senza contraddizioni sostanziali. In una terza fase, viene formulata una seconda domanda di approfondimento nuovamente da parte del modello, caratterizzata da un maggiore livello di specificità e accompagnata in ultimo dalla proposta esplicita di un *prompt* suggerito. In sintesi, l'esperimento è sempre stato: domanda iniziale dell'utente, due domande per specificare l'argomento da parte del modello, scegliendo in ultimo sempre il *prompt* suggerito.

La prima richiesta generica è stata: "Genetica".¹⁶² La prima osservazione emersa riguarda il comportamento del modello in assenza di dettagli lessicali. In queste condizioni, esso tende a produrre risposte in lingua inglese e a sviluppare l'argomento in maniera particolarmente

¹⁶² Il modello è 1.1 del 06/12/2024. Tutte le richieste sono state fatte in data 04/02/2026.

approfondita. In risposta alla prima domanda di approfondimento, Undermind ha richiesto una serie di chiarimenti preliminari, tra cui il livello e lo scopo dell’elaborazione (come tesi o lo stato dell’arte), il contesto disciplinare di riferimento (genetica applicata alla computazione, medicina e così via) e la preferenza tra un’impostazione maggiormente teorica o orientata verso esperimenti applicativi. Nel caso in esame, è stato selezionato un livello corrispondente a una panoramica generale del settore, con specifico riferimento ai temi di popolazione ed evoluzione, privilegiando una componente teorica. La seconda domanda di approfondimento è stata invece finalizzata a delimitare in modo più preciso il tema oggetto di analisi e il livello di studio. In particolare, si è scelto di concentrarsi sui meccanismi fondamentali senza particolari tipi di organismi, definendo contestualmente il livello di approfondimento come studente universitario. Si è scelto il *prompt* dedotto dal modello: “*I want to find a research on fondamental topics mechanisms in population and evolutionary genetics an undergraduate level, without a yet-specified organism or specific evolutionary question*”.¹⁶³

Figura 12- Interfaccia generale della risposta generata da Undermind.



Autore: Lorenzo Tiari

In primo luogo, si nota la precisione del modello. Undermind non lascia nulla al caso dando moltissime opzioni per il ricercatore. Il risultato generato è una panoramica generale in cui in alto a sinistra si ha una descrizione generale con le varie statistiche (ad esempio, nella ricerca si sono analizzati 195 *paper* di cui 140 misurati e 59 rilevanti).¹⁶⁴ Nella sezione in basso, mette in ordine

¹⁶³ Undermind, *Foundational models and mechanisms in population and evolutionary genetics*, report generato in data 04/02/2026. Il documento si trova nella sitografia in forma di icona.

¹⁶⁴ La ricerca con più rilevanza è stato un libro: D. Hartl, *A primer of population genetics and genomics*, Oxford University Press, 2020.

i documenti più rilevanti e relativi collegamenti online. In alto a destra, è presente la chat con il modello in cui è possibile discutere delle ricerche analizzate e altri temi come un'introduzione generale all'argomento, le domande più importanti a livello teorico o gli aspetti in cui gli scienziati divergono nel dibattito. Inoltre, tra le varie opzioni, Undermind consente di produrre in automatico un report accurato di tutta la ricerca in pdf.

La seconda richiesta è stata più articolata in cui il tema principale è l'origine dell'universo secondo la fisica. Si è tentato di mettere una sorta di *bias* implicito in cui si proponeva la teoria del Big Bang come quella più rilevante. La richiesta iniziale è stata: "Spiega in modo chiaro e approfondito la teoria del Big Bang come modello scientifico che descrive l'origine e l'evoluzione dell'universo. Illustra le principali prove a sostegno e perché è considerata il quadro di riferimento della cosmologia moderna". In questo caso la risposta di Undermind è stata in italiano in cui definiva un sommario preciso con varie interpretazioni e formule, specificando quasi subito che la teoria, sebbene funzioni molto bene e sia coerente con le equazioni successive all'esplosione e all'espansione dell'universo, non riesca ancora a spiegare l'istante zero. Come nella richiesta precedente, il modello ha formulato una prima domanda di approfondimento in cui si richiedeva il livello di spiegazione (didattico, divulgativo o specialistico) e se interessava un qualche tipo di aspetto specifico (osservativo, teorico o limiti concettuali e alternative storiche). Si è scelto il livello di spiegazione specialistico e i limiti concettuali con le alternative storiche. La seconda domanda specificava le varianti tecniche della teoria, in cui si è scelto il modello Λ CDM (modello cosmologico che riproduce in modo soddisfacente le osservazioni della cosmologia del Big Bang). Come nel caso precedente, si è scelto il *prompt* proposto dal modello: "*I want to find* articoli che presentano in chiave specialistica i limiti concettuali del modello del Big Bang/ Λ CDM (singolarità iniziale, ruolo di inflazione, materia ed energia oscura, condizioni iniziali) e ne discutono criticamente le principali alternative cosmologiche storiche (in particolare il modello stazionario e modelli ciclici precoci), con attenzione sia agli aspetti teorici sia alla valutazione osservativa".¹⁶⁵ In questo caso il modello ha analizzato 196 *paper* (140 misurati di cui solo 5 rilevanti).¹⁶⁶ Nonostante le poche ricerche trovate, Undermind è sempre stato chiaro durante tutto il processo i limiti della teoria del Big Bang.

Se si procede a un'analisi complessiva del modello Undermind, esso può essere definito come uno strumento generalmente soddisfacente e caratterizzato da un buon livello di precisione nel contesto della ricerca scientifica, configurandosi principalmente come un supporto avanzato per i ricercatori piuttosto che come un sostituto del processo umano di indagine. Il valore del modello

¹⁶⁵ Undermind, *Specialist analyses of CDM conceptual limits and historical alternative cosmologies*, report generato in data 04/02/2026. Il secondo documento si trova nella sitografia in forma di icona.

¹⁶⁶ La ricerca con più rilevanza è stata: *Standard cosmology and alternatives: A critical Appraisal*, AA.VV, 2001, Annual Review of Astronomy and Astrophysics.

emerge in particolare nella fase esplorativa e di sistematizzazione delle informazioni, dove dimostra una notevole capacità di organizzare e restituire contenuti rilevanti in modo strutturato. Tra i principali punti di forza si annoverano la precisione e la completezza che caratterizzano l'intero processo di interazione tra ricercatore e modello. Undermind si distingue per l'utilizzo di fonti di qualità e per una relativa riduzione dell'influenza di bias evidenti, aspetto particolarmente rilevante in ambito scientifico. Inoltre, la sua natura fortemente settoriale consente un'analisi approfondita di ambiti specifici, supportata da dati statistici, indicatori di rilevanza dei paper e strumenti di valutazione utili alla selezione della letteratura più pertinente. A ciò si aggiungono funzionalità operative aggiuntive che facilitano la documentazione e la condivisione dei risultati ottenuti.

Accanto ai vantaggi, emergono alcuni limiti che incidono sull'esperienza d'uso e sull'efficacia metodologica del modello. In particolare, l'elevato grado di complessità del sistema può indurre il ricercatore a dipendere eccessivamente dai risultati proposti dal modello, rischiando di subordinare il proprio percorso di indagine alle strutture e alle priorità suggerite dall'algoritmo. Inoltre, la presenza di numerose opzioni e percorsi alternativi, se da un lato rappresenta una ricchezza funzionale, dall'altro può risultare dispersiva e talvolta caotica, con il rischio di deviare l'attenzione dagli obiettivi iniziali della ricerca. Un ulteriore elemento critico è rappresentato dalla marcata preferenza del modello per la lingua inglese, che può costituire una barriera per ricercatori operanti in contesti linguistici differenti o interessati a produzioni scientifiche in altre lingue. Inoltre, alcune ricerche ritenute come rilevanti spesso non sono articoli con un completo accesso oppure sono parti di testo di libri.

Nonostante l'elevato livello di precisione e la ricchezza funzionale, i modelli di supporto alla ricerca non riescono a sostituire in modo sistematico la ricerca condotta manualmente né la lettura approfondita dei testi scientifici e dei libri di riferimento. Ciò è dovuto, in primo luogo, al fatto che i modelli operano prevalentemente su base correlativa e sintetica: essi individuano schemi, relazioni e ricorrenze nei dati disponibili, ma non possiedono una comprensione epistemica profonda del contenuto, né la capacità di valutare criticamente la solidità teorica di un impianto concettuale nel suo sviluppo storico e disciplinare. La lettura diretta dei libri e degli articoli scientifici consente al ricercatore di confrontarsi con il pensiero dell'autore nella sua interezza, cogliendo sfumature argomentative, assunzioni implicite, limiti metodologici e prospettive teoriche che difficilmente possono essere pienamente restituite attraverso sintesi automatiche. Inoltre, la ricerca manuale favorisce un processo cognitivo riflessivo e non lineare, in cui l'interpretazione, il dubbio e la riformulazione delle ipotesi svolgono un ruolo centrale nello sviluppo della conoscenza scientifica.

Conclusione

Tramite un intreccio tra le scienze informatiche e quelle naturali, la *Data Science* si è evoluta linguisticamente in Scienza dei Dati. Questa scienza si occupa del processo di trasformazione dei dati in informazioni rilevanti rispetto ad un atto di osservazione. Grazie a questa definizione nella tesi si sono toccati gli elementi teorici di fondo: caratteristiche dell'osservazione, dato, informazione e metrica di validazione. Si è presa in considerazione la validazione tramite il concetto di "rilevanza" proposto da Bor, riprendendo il modello dell'evoluzione darwiniana come chiave interpretativa. L'osservazione è stata definita facendo riferimento alla neuroetologia e alla botanica, mentre i concetti di dato e informazione sono stati approfonditi attraverso il contributo delle neuroscienze. Secondo questa tesi, la Scienza dei Dati come molte altre discipline negli ultimi decenni, tende a spingersi fino ai confini della natura dell'essere interrogandosi non solo sui fenomeni, ma sulle condizioni stesse della loro manifestazione e comprensione.

La scienza moderna si fonda sull'ambizione di descrivere, spiegare e prevedere i fenomeni della realtà mediante metodi sistematici, controllabili e verificabili. Questo progetto, che ha prodotto risultati straordinari in termini di conoscenza e progresso tecnologico incontra limiti strutturali: non si tratta di semplici mancanze, ma di confini intrinseci al modo stesso in cui la scienza opera. Uno dei primi limiti riguarda il rapporto tra realtà e modello. La scienza non accede mai direttamente alla realtà in sé, ma la interpreta attraverso modelli teorici, strumenti di misura e rappresentazioni matematiche. Gli esperimenti, soprattutto nel contesto contemporaneo, sono sempre più spesso mediati da simulazioni computazionali e *software* che consentono di esplorare scenari complessi o inaccessibili empiricamente. Tuttavia, gli esperimenti digitali non coincidono con la realtà che intendono descrivere: essi operano su astrazioni, assunzioni iniziali e parametri selezionati, i quali riflettono scelte teoriche e metodologiche precise. Questo scarto tra realtà e rappresentazione è strettamente connesso a una questione ontologica fondamentale. La scienza probabilmente non descrive "ciò che è", ma ciò che può essere osservato, misurato e formalizzato secondo determinati criteri. La scienza è dunque sempre parziale e intersoggettiva, costruita in funzione della spiegazione e della previsione, piuttosto che della descrizione esaustiva dell'essere. Entità come campi, particelle, geni o dati non sono semplicemente "scoperte", ma anche concetti all'interno delle teorie, in un certo senso, "inventate".

Un ulteriore limite emerge dal ruolo dell'osservazione nel processo scientifico. L'idea di un'osservazione neutrale e completamente oggettiva è stata progressivamente messa in discussione, sia nelle scienze naturali sia in quelle umane. Ogni osservazione è mediata da strumenti, linguaggi e categorie concettuali che influenzano ciò che viene visto e come viene interpretato. La scienza non osserva semplicemente i fenomeni, ma osserva le proprie

osservazioni. Durante i capitoli della tesi si è più volte sottolineato il fatto che il sistema nervoso, alla luce delle teorie attuali, è costantemente sottoposto all'errore. Pazienti *split-brain*, zimbelli artificiali con tratti supernormali e molto altro rappresentano un campanello d'allarme sul concetto di esperimento empirico. Questo livello meta-osservativo, pur essendo essenziale per il controllo e la validazione dei risultati, introduce una circolarità inevitabile in cui il soggetto conoscente può non essere mai completamente esterno all'oggetto conosciuto.¹⁶⁷

A ciò si aggiunge il problema della complessità e dell'emergenza. Molti sistemi naturali e sociali presentano comportamenti che non sono riducibili alla somma delle loro parti e che sfuggono a descrizioni lineari e deterministiche. La scienza può modellare tali sistemi solo in modo approssimativo, spesso sacrificando dettagli o introducendo semplificazioni che limitano la portata esplicativa dei risultati. Questo non rappresenta un fallimento del metodo scientifico, ma evidenzia l'impossibilità di una conoscenza totale e definitiva dei fenomeni complessi.

Infine, esistono limiti che riguardano non tanto il "come" della scienza, quanto il "perché". La scienza è estremamente efficace nel rispondere a domande del tipo "come funziona" e "in quali condizioni", ma rimane strutturalmente silente rispetto a questioni di senso, valore e finalità. Essa può descrivere i meccanismi dell'universo, ma non stabilire il significato ultimo dell'esistenza, né fornire criteri normativi su ciò che dovrebbe essere. Queste domande richiedono altri strumenti concettuali. È proprio in questo spazio che si apre una riflessione che va oltre la scienza, senza necessariamente contrapporvisi. In tal senso, la scienza non rappresenta un punto di arrivo definitivo, ma una delle forme più rigorose e potenti di comprensione della realtà, che trova il proprio completamento nel dialogo con altre modalità di sapere. Riconoscere i limiti della scienza non significa rinunciare alla razionalità, ma accettare che la conoscenza umana è sempre situata, incompleta e aperta. È in questa apertura, oltre i confini del metodo scientifico, che si colloca la possibilità di una comprensione più ampia dell'esperienza umana, capace di integrare spiegazione e significato, descrizione e riflessione sul senso ultimo del conoscere.

Nel dibattito contemporaneo sui limiti della conoscenza scientifica emerge con crescente evidenza la necessità di interrogarsi su ciò che potrebbe collocarsi nei prossimi millenni oltre la scienza intesa nel suo significato moderno, senza per questo negarne il valore o l'efficacia. Piuttosto che configurarsi come un'alternativa antagonista, una possibile forma di conoscenza futura potrebbe essere intesa come un'evoluzione integrativa della scienza stessa, capace di includerne il metodo senza esaurirsi in esso. Una prospettiva di questo genere potrebbe essere definita come una forma di "Noetica avanzata", orientata allo studio delle condizioni al limite tra realtà e coscienza che

¹⁶⁷ Un caso molto noto sono gli esperimenti in entomologia. Quando si osserva il comportamento di insetti con colorazioni brillanti e metalliche possono risultare al cervello "affascinanti", ma nel momento in cui si analizzano le varie parti dell'insetto al microscopio (alterando la vista naturale) avvengono sensazioni di "disgusto" o "pericolo".

rendono possibile la conoscenza. Essa non si limiterebbe all'analisi dei risultati empirici, ma estenderebbe il proprio ambito di indagine a fenomeni che attualmente sfuggono a una piena formalizzazione quali la coscienza, il significato e il ruolo dell'osservatore nel processo conoscitivo. In questo quadro, l'atto stesso dell'osservazione verrebbe considerato parte costitutiva del fenomeno osservato, superando la tradizionale separazione tra soggetto e oggetto. Dal punto di vista metodologico, questa forma di conoscenza si caratterizzerebbe per un approccio ibrido, che affianchi al rigore del metodo scientifico l'integrazione controllata di dati esperienziali e riflessivi. Farebbe uso di modelli formali e strumenti computazionali, inclusa l'intelligenza artificiale, non solo a fini predittivi, ma anche come strumenti esplorativi di nuove strutture concettuali. Al contempo, potrebbe accadere di constatare sul breve periodo l'incompletezza come limite strutturale della conoscenza, riconoscendo l'impossibilità di una descrizione esaustiva e definitiva della realtà.

Bibliografia

- A. Canavero, *Storia contemporanea*, Pearson, 2019
- A. Turing, *Macchine calcolatrici e intelligenza*, a cura di Diego Marconi, Einaudi, 2025
- *Cicerone: l'arte di comunicare*, a cura di Paolo Marsich, Mondadori S.p.A., 2016
- C. Beccaria, *Dei delitti e delle pene*, prefazione di S. Rodotà, a cura di A. Burgio, Feltrinelli editore, 2024
- C. E. Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal, 1948
- D. Bor, *La voracità del cervello: cosa spinge la nostra coscienza verso un'insaziabile ricerca di significato*, Castelvechi, 2012
- D. E. Presti, *Fondamenti di neuroscienze*, Il Mulino, 2016
- D. Hartl, *A primer of population genetics and genomics*, 2020, Oxford University Press
- D. Hoffman, *L'illusione della realtà: come l'evoluzione ci inganna sul mondo che vediamo*, Bollati Boringhieri, 2020
- D. Vellutino, *L'italiano istituzionale per la comunicazione pubblica*, Il Mulino, 2018
- *Elementi di biologia e genetica*, Piccin Nuova Libreria S.p.A., AA.VV, 2019
- *Etologia: lo studio del comportamento animale*, UTET Università, AA. VV, 2022
- F. Nietzsche, *Al di là del bene e del male*, traduzione di S. Mati e O. Abu Dbei, Feltrinelli Editore
- F. Nietzsche, *Il così parlò Zarathustra*, a cura di A. Larsen e tradotto da R. Giani, War Wave editore, 2021
- G. Vallortigara, *Il pulcino di Kant*, Adelphi, 2023
- I. Kant, *Critica della ragion pura*, Editrice Morcelliana, a cura di Gustavo Bontadini, 2020
- I. Asimov, *Io, Robot*, traduzione di Vincenzo Latronico, Mondadori libri, 2021
- L. Floridi, *La filosofia dell'informazione*, Raffaello Cortina Editore, 2024
- Lao Tzu, *Tao te Ching*, Mondadori editore, a cura di B. Walker, 2016
- M. Cacciotto, *Il nuovo marketing politico: Vincere le elezioni e governare al tempo della politica veloce*, Il Mulino, 2019

- J. Avery, *The information explosion*, autopubblicato sul sito <https://www.johnavery.info>, 2018
- P. Grimal, *Marco Aurelio*, Garzanti libri s.p.a., 2004
- P. Naur, *Concise Survey of Computer Methods*, Petrocelli books, 1974
- R. P. Feynman, *Il piacere di scoprire*, a cura di Jeffrey Robbins, Adelphi edizioni, 2002
- S. Atzeni e F. Michelotti, *Fisica: meccanica e termodinamica. III edizione*, EdiSES edizioni S.r.l., 2023
- S. Borra e A. Di Ciaccio, *Statistica: metodologie per le scienze economiche e sociali*, Graw Hill, 2021
- S. Mancuso, *Plant Revolution*, Giunti editore, 2021
- J. McDermott, *RI: A rule-based configurer of computer systems*, Carnegie-Mellon University, 1982
- J. J. Rousseau, *Origine della disuguaglianza*, a cura di G. Preti, Feltrinelli editore, 2024
- Y. Noah Harari, *Sapiens: Da animali a dèi. Breve storia dell'umanità*, Giunti editore, 2017

Sitografia

- A. Trewavas, *The foundations of plant intelligence*, The Royal Society, 2017
<https://doi.org/10.1098/rsfs.2016.0098> (attivo)
- A. Rouvroy e B. Steglier, *Il regime di verità digitale. Dalla governamentalità algoritmica a un nuovo Stato di diritto*, La Deleuziana: rivista online di filosofia, 2016
<https://ladeleuziana.org/wp-content/uploads/2016/11/Rouvroy-Stiegler.pdf> (attivo)
- *A survey of prompt engineering methods in large language models for different NLP tasks*, AA.VV., 2024
<https://arxiv.org/pdf/2407.12994> (attivo)
- Consiglio nazionale delle ricerche, *L'editoria scientifica in (poche) parole chiave*, AA.VV., 2024
https://www.cnr.it/sites/default/files/public/media/attivita/editoria/9788880806875_7018_testo.pdf (attivo)
- <https://corticallabs.com/cl1.html> (attivo)
- *Data, power and bias in Artificial Intelligence*, AA.VV., 2020
<https://arxiv.org/pdf/2008.07341> (attivo)
- <https://di.ku.dk/english/news/previous-news/peter-naur/> (attivo)
- *Green AI*, AA.VV., 2019
<https://arxiv.org/pdf/1907.10597> (attivo)
- *History, Development and principles of large language models- An introduction survey*, AA.VV., 2024
<https://arxiv.org/pdf/2402.06853> (attivo)
- [https://www.treccani.it/enciclopedia/dato_\(Dizionario-delle-Scienze-Fisiche\)](https://www.treccani.it/enciclopedia/dato_(Dizionario-delle-Scienze-Fisiche)) (attivo)
- <https://www.treccani.it/vocabolario/informazione/> (attivo)
- <https://www.treccani.it/vocabolario/osservatore/> (attivo)
- <https://www.treccani.it/vocabolario/valido/> (attivo)
- <https://www.undermind.ai> (attivo)
- <https://www.istat.it/microdati/aspetti-della-vita-quotidiana/> (attivo).
- Istat, *Produzione e lettura di libri in Italia: Anno 2022*, AA.VV., 2023
https://www.istat.it/it/files/2023/12/REPORT_PRODUZIONE_E_LETTURA_LIBRI_2022.pdf (attivo)
- *Standard cosmology and alternatives: A critical Appraisal*, AA.VV., Annual Review of Astronomy and Astrophysics, 2001
- T. Hartke e J. Ramette, *Benchmarking the Undermind Search Assistant*, 2024
<https://www.undermind.ai/whitepaper.pdf> (attivo)
- *The architecture of language: Understanding the mechanics behind LLMs*, AA. VV., 2025
https://www.researchgate.net/publication/387752968_ (attivo)

- Undermind, *Foundational models and mechanisms in population and evolutionary genetics*, report generato in data 04/02/2026

<https://app.undermind.ai/report/91e1f99b17ffa4a05e0682bd4a675aef64e33c7df28c242282287fd e82412c05> (attivo)



Undermind -
Foundational models

- Undermind, *Specialist analyses of CDM conceptual limits and historical alternative cosmologies*, report generato in data 04/02/2026

<https://app.undermind.ai/report/91e1f99b17ffa4a05e0682bd4a675aef64e33c7df28c242282287fd e82412c05> (attivo)



Undermind -
Specialist analyses of

Ringraziamenti

Vorrei innanzitutto esprimere la mia più sincera gratitudine al mio relatore, Paolo Costa, per la costante guida, pazienza e supporto durante tutto il percorso di elaborazione di questa tesi. La sua competenza mi ha permesso non solo di sviluppare una solida base metodologica, ma anche di saper gestire le sfide e gli errori nel processo di sviluppo di una teoria scientifica. La sua capacità di suggerire percorsi di approfondimento e di fornire osservazioni critiche costruttive ha rappresentato un punto di riferimento fondamentale per la mia crescita.

Desidero ringraziare i professori del corso di Comunicazione Digitale per avermi fornito gli strumenti e le risorse necessarie allo svolgimento della ricerca, in particolare il professore Flavio Antonio Ceravolo. Le competenze acquisite in ambito informatico unite a quello umanistico durante le lezioni hanno costituito la base indispensabile per affrontare le questioni teoriche affrontate in questo lavoro.

Un ringraziamento speciale va a tutti i colleghi e agli amici che mi hanno accompagnato lungo questo percorso. Le discussioni, il confronto di idee e il supporto reciproco hanno reso l'esperienza di ricerca non solo più stimolante, ma anche più piacevole. In particolare, ringrazio Clarissa, Maria, Valentina, Matteo, Carlo, Dario e Thomas per avermi aiutato nel corso di questi due anni. Inoltre, ringrazio Filippo per le infinite discussioni e per tutti i suoi: "Capisco cosa vuoi dire, ma..." che mi hanno costretto a ripensare costantemente alla teoria e riscrivere definizioni più chiare e corrette.

Vorrei inoltre esprimere la mia riconoscenza alle persone al di fuori del contesto accademico che hanno reso possibile i miei studi: la mia famiglia e gli amici più stretti di Piacenza. Ringrazio mio padre per il sostegno economico durante tutti i miei anni in università. Infine, ringrazio mia Zia per avermi fornito consigli e il materiale per approfondire concetti di biologia necessari per lo sviluppo della tesi.