UNIVERSITÀ DEGLI STUDI 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

ARTIFICIAL (UN)INTELLIGENCE: UN PERICOLOSO POTENZIAMENTO DELLE CAPACITA' COGNITIVE

Relatore:
Chiar.mo Prof. Paolo Costa

Correlatore:
Chiar.mo Prof. Massimiliano Vaira

Tesi di laurea di
Filippo Maria Sorgente

ANNO ACCADEMICO 2024/2025

ABSTRACT

I modelli di intelligenza artificiale sono sistemi statistici privi di comprensione semantica, intenzionalità e coscienza, eppure li trattiamo quotidianamente come se capissero, ragionassero e volessero. Questa tesi non si chiede se l'IA sia intelligente, ma perché gli esseri umani le attribuiscano caratteristiche umane e quali conseguenze cognitive, sociali e organizzative produca questa attribuzione. Il primo capitolo stabilisce cosa è l'intelligenza artificiale: i modelli linguistici producono *output* formalmente coerenti senza accedere tuttavia ad alcuna comprensione del significato, né ad alcun ancoraggio alla realtà fisica. Il secondo capitolo sposta l'indagine sull'essere umano, analizzando i meccanismi evolutivi e cognitivi che rendono strutturalmente inevitabile, a livello individuale, l'attribuzione di agency a sistemi algoritmici. Sul piano collettivo, invece, il secondo capitolo mostra come queste proiezioni si sedimentino nell'ambiente informazionale, circolino nelle reti sociotecniche e si stabilizzino nel tempo attraverso meccanismi istituzionali, fino a diventare pratiche organizzative condivise. Il terzo capitolo sposta il piano dell'analisi dall'ontologia alla percezione, proponendo un'analisi esplorativa volta a osservare come tali meccanismi possano manifestarsi nel contesto lavorativo reale attraverso il punto di vista di un utilizzatore autorevole. Dall'intervista emerge come il rischio risieda nelle condizioni strutturali dell'adozione degli strumenti di intelligenza artificiale: la pressione produttiva, la familiarità con lo strumento e la plausibilità formale degli *output* sembrano, infatti, convergere nel rendere razionale la delega acritica anche per utilizzatori avanzati e consapevoli. Perciò, il confine tra potenziamento e depotenziamento non pare essere tecnico, ma cognitivo e culturale, dal momento che potrebbe coincidere con il grado di consapevolezza con cui il soggetto preserva il proprio ruolo di agente primario. La risposta che il lavoro propone è antropologica e si articola su tre piani: sul piano individuale, la consapevolezza della natura non pensante dell'IA è condizione necessaria ma non sufficiente; sul piano del capitale semantico, in un contesto saturo di contenuti generati automaticamente, il valore si sposta sulla capacità umana di scegliere, di valutare e di avere una propria interpretazione del mondo; sul piano collettivo, la risposta richiede una narrazione consapevole che restituisca all'essere umano il ruolo di agente primario nelle scelte individuali, culturali,

istituzionali e politiche che determinano come la tecnologia viene progettata, comunicata, adottata e governata.

Artificial intelligence models are statistical systems devoid of semantic understanding, intentionality, and consciousness, yet we treat them daily as if they understood, reasoned, and intended. This thesis does not ask whether AI is intelligent, but why human beings attribute human characteristics to it and what cognitive, social, and organizational consequences this attribution produces. The first chapter establishes what artificial intelligence is: language models produce formally coherent outputs without accessing any understanding of meaning or any grounding in physical reality. The second chapter shifts the inquiry toward the human being: at the individual level, it analyzes the evolutionary and cognitive mechanisms that make the attribution of agency to algorithmic systems structurally inevitable; at the collective level, it shows how these projections sediment within the infosphere, circulate through sociotechnical networks, and stabilize through the institutional mechanisms until they become shared organizational practices. The third chapter shifts the level of analysis from ontology to perception, proposing an exploratory analysis aimed at observing, through an interview with an authoritative professional, how such mechanisms may manifest in real workplace contexts. The study intends to highlight that the risk does not necessarily lie in the user's naivety, but in the structural conditions of adoption: productive pressure, familiarity with the tool, and the formal plausibility of outputs seem to converge to render uncritical delegation rational even for advanced and self-aware users. The boundary between augmentation and diminishment is not technical but cognitive and cultural: it coincides with the degree of awareness with which the professional preserves their role as primary agent. The response proposed by this work is anthropological and unfolds across three levels: at the individual level, awareness of the non-thinking nature of AI is a necessary but insufficient condition; at the level of semantic capital, in a context saturated with automatically generated content, value shifts toward the human capacity to choose, evaluate, and account for one's own interpretation of the world. At the collective level, the response requires a conscious narrative that restores to the human being the role of primary agent in the individual, cultural, institutional, and political choices that determine how technology is designed, communicated, adopted, and governed.

Indice

INTRODUZIONE

I modelli di intelligenza artificiale sono sistemi statistici privi di comprensione semantica, intenzionalità e coscienza, eppure li trattiamo quotidianamente come se capissero, ragionassero e volessero. La tesi non si chiede se l'IA sia intelligente, ma perché gli esseri umani le attribuiscano caratteristiche umane e quali conseguenze cognitive, sociali e organizzative produca questa attribuzione. Analizzerò le dinamiche attraverso cui gli esseri umani attribuiscono intenzionalità e *agency* ai sistemi di intelligenza artificiale, per poi osservare gli effetti di tali meccanismi a livello sia individuale che di gruppo. Inoltre, stabilendo la natura ontologica e i limiti tecnici dei modelli linguistici, sarò in grado di inquadrare il fenomeno dell'antropomorfizzazione, che porta a conseguenze esiziali in contesti sociali e organizzativi.

Il primo capitolo stabilisce cosa è l'IA attraverso un'analisi critica che muove da una provocatoria ricerca di Apple, inserita all'interno di un consolidato filone di studi nel campo dell'intelligenza artificiale e della linguistica computazionale, la quale evidenzia i limiti strutturali dei modelli di intelligenza artificiale di ultima generazione. I modelli linguistici producono *output* formalmente coerenti e linguisticamente plausibili, senza tuttavia accedere ad alcuna comprensione profonda del significato di ciò che elaborano: producono testo a partire da *pattern* appresi durante l'addestramento e non da un'esperienza diretta del mondo. L'assenza di ancoraggio alla realtà li rende strutturalmente incapaci di distinguere il vero dal verosimile. Infatti per un modello di intelligenza artificiale un'affermazione falsa, ma linguisticamente coerente, è indistinguibile da un'affermazione vera. Ciò accade poiché il modello non ha accesso al mondo e non può verificare se ciò che produce corrisponda alla realtà: il modello valuta solamente la plausibilità linguistica di una sequenza di parole. Quindi il limite principale che emerge dal capitolo è ontologico e risiede nella mancanza di ancoraggio al mondo fisico e sociale che nell'essere umano costituisce il fondamento della comprensione autentica. Un essere umano non elabora simboli astratti, ma possiede un corpo, vive in un contesto sociale, accumula esperienza diretta del mondo e orienta le proprie azioni in funzione di bisogni, intenzioni e valori. È questa dimensione incarnata e situata che rende possibile non solo la comprensione del significato, ma anche la

volontà e l'intenzionalità, caratteristiche che un modello linguistico non possiede. I capitoli successivi declineranno la differenza tra forma e sostanza, tra simulazione e comprensione, nelle sue conseguenze psicologiche, sociali e organizzative.

Il secondo capitolo sposta l'indagine sull'essere umano: sul piano individuale, analizzerò i meccanismi evolutivi e cognitivi che rendono strutturalmente inevitabile l'attribuzione di intenzionalità a sistemi algoritmici. Allo stesso tempo, sul piano collettivo, mostrerò come le proiezioni individuali si sedimentino nel tempo in credenze condivise attraverso l'ambiente informazionale e le reti sociotecniche in cui viviamo.

Il terzo capitolo introduce una rielaborazione sociologica del test di Turing che sposta l'analisi dall'ontologia alla percezione. La domanda non è se l'IA sia intelligente in senso proprio, ma se e a quali condizioni essa venga trattata come tale dagli attori sociali nel contesto quotidiano e professionale: se, cioè, le vengano attribuiti stati mentali e intenzionalità indipendentemente dalla natura dei modelli. Questo approccio intende esplorare come i rischi descritti nei capitoli precedenti possano dipendere non tanto dalle capacità effettive dei modelli di intelligenza artificiale, quanto dalle credenze che gli esseri umani costruiscono attorno ad essi. Il Turing sociale può essere attivo, quando il professionista sceglie consapevolmente di trattare l'IA come agente, o passivo, quando tale attribuzione avviene senza valutazione critica.

Infine, il capitolo propone un'analisi esplorativa volta a osservare come i meccanismi descritti nel secondo capitolo possano manifestarsi nella realtà lavorativa quotidiana. Attraverso il punto di vista di un utilizzatore autorevole, l'indagine mira a fornire una prima interpretazione del fenomeno da cui sviluppare interrogativi e nuove traiettorie di ricerca sul tema.

CAPITOLO 1

1.1 L'illusione del ragionamento

Siamo testimoni di quella che, a tutti gli effetti, è la più grande esibizione di illusionismo della storia: ci siamo convinti che le macchine ragionino. Anzi assumiamo che ragionino e che pensino negli stessi termini di noi esseri umani.

Ma perché lo facciamo?

Facciamo questo tipo di assunzione dal momento che l'unica intelligenza a cui sappiamo paragonarle, metro di paragone inevitabile, è la nostra intelligenza, quella umana[1]. Il nostro metro di giudizio è fortemente antropocentrico - *è intelligente se somiglia al mio modo di ragionare* - ma non è necessariamente oggettivo o universale, poiché, così facendo, stiamo misurando solamente quanto l'altro somigli a noi, ma non stiamo considerando cosa il termine intelligenza significhi in assoluto. Non è detto, infatti, che dal punto di vista ontologico sia corretto pensarla in tali termini, riducendo l'intelligenza a un modello prettamente e unicamente umano.

Spesso, confrontando l'intelligenza artificiale con quella umana, il paragone implicito non è l'intelligenza nella sua interezza, ma una sua forma specifica: quella logico-consequenzialista fondata su inferenza, calcolo, deduzione e risoluzione formale di problemi. Questo tipo di intelligenza, portata in auge dalla filosofia occidentale tradizionale, è proprio quella che gli attuali modelli computazionali replicano in maniera eccellente. Tuttavia, Howard Gardner[2], nella sua teoria delle intelligenze multiple, ha identificato altre forme di intelligenza non gerarchiche: l'intelligenza interpersonale è la capacità di leggere le emozioni altrui, di percepire le dinamiche di gruppo e di adattare il proprio comportamento al contesto relazionale e sociale; l'intelligenza intrapersonale riguarda la consapevolezza di sé e la capacità di riconoscere i propri stati mentali ed

[1] Il concetto di intelligenza è ancora largamente discusso perché non univoco e per forza antropocentrico: per dire, gli etologi mostrano casi in cui animali come polpi, corvi o scimpanzé risolvono problemi complessi con un approccio radicalmente diverso rispetto a noi esseri umani. Modi che, tuttavia, a noi sembrano non propriamente intelligenti solo perché non basati su linguaggio astratto o logica formale, bensì su istinto, esperienza sensoriale e ambiente.

[2] H. Gardner, Frames of Mind*: The Theory of Multiple Intelligences*, New York, Basic Books, 1983.

emotivi; l'intelligenza corporea-cinestetica implica l'elaborazione degli stimoli ambientali attraverso il corpo e l'esperienza sensoriale diretta del mondo fisico.

Un sistema di intelligenza artificiale, per quanto sofisticato, opera sulla dimensione logica e linguistica: non possiede un corpo e una storia personale, non ha la capacità di leggere un volto, di interpretare le espressioni o di percepire una situazione relazionale. Non può esercitare forme di intelligenza che derivano dal fatto di essere incarnati nel mondo. Dunque, ciò che chiamiamo intelligenza artificiale è un'intelligenza parziale, dal momento che è la replica di una sola delle dimensioni dell'intelligenza. Paragonarla all'intelligenza umana significa confrontare una parte con un insieme. Al giorno d'oggi quindi, paragonando la presunta intelligenza delle macchine alla nostra, è diventato comune credere e assumere che le macchine siano intelligenti, che pensino e ragionino come noi, se non addirittura meglio. Tuttavia i ricercatori di Apple[3] hanno portato alla luce questa illusione, mostrando come gli attuali modelli di intelligenza artificiale, in particolare di intelligenza artificiale generativa, diano solo l'impressione di pensare e ragionare come noi, ma in realtà lo facciano in un modo strutturalmente differente.

Come vedremo, lo studio si inserisce in un filone consolidato di decenni di studi che sottolineano questo limite, che non è inedito ma una conferma, applicata ai modelli più recenti, di barriere teoriche già ampiamente previste dalla filosofia della mente e della linguistica computazionale.

1.1.1 The illusion of thinking: il collasso dei Large Reasoning Models

Nel giugno del 2025, Apple ha pubblicato una ricerca dal titolo emblematico *The illusion of thinking*[4], che ha rimesso in discussione la capacità dei modelli di intelligenza artificiale di ragionare[5]. I modelli linguistici presi in esame non sono riusciti a risolvere alcuni classici indovinelli logici, perciò hanno indotto i ricercatori di Apple a concludere che non siano in grado di ragionare al pari di un essere umano. Il

[3] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, arXiv, 2025.
[4] Ibid.
[5] *The Illusion of Thinking* è il titolo con cui la stampa generalista ha ribattezzato lo studio; il *paper* scientifico utilizza come strumento di test il *benchmark* GSM-Symbolic, sviluppato dallo stesso gruppo di ricerca Apple in un lavoro precedente del 2024.

presupposto implicito è che un essere umano avrebbe risolto i quiz attraverso la logica[6] e che, di conseguenza, il fallimento dei modelli costituisca la prova dell'assenza di ragionamento[7].

Tuttavia, questa conclusione merita una riflessione: il fatto che i modelli non abbiano risolto i quiz non implica necessariamente che non ragionino poiché potrebbero essere entrati in gioco altri fattori. Per esempio, i modelli potrebbero non essere riusciti a risolvere determinati problemi logici non perché siano privi di qualsiasi forma di ragionamento, ma perché il loro ragionamento è strutturalmente diverso da quello umano: sarebbe insensato testare l'intelligenza di un cane chiedendogli ad esempio di risolvere un'equazione; il fallimento non proverebbe l'assenza di intelligenza, ma semplicemente che quella forma di intelligenza opera secondo principi diversi. Allo stesso modo, il fatto che un modello linguistico fallisca un test logico progettato per esseri umani non prova l'assenza di ragionamento, ma evidenzia come la sua capacità di elaborazione sia strutturalmente diversa dalla nostra.

Inoltre, i limiti osservati nello studio di Apple non dipendono esclusivamente dalla diversità strutturale tra ragionamento umano e artificiale, ma dipendono significativamente anche dal tipo, dalla quantità e dalla qualità delle informazioni su cui il modello è stato addestrato e dalle scelte compiute nella fase di *training*. Un modello linguistico, dal momento che non può accedere direttamente al mondo, nel momento in cui ricevesse informazioni incomplete o sbilanciate, potrebbe non essere sempre in grado di risolvere determinati problemi. In questo senso, molti dei fallimenti documentati non rivelano un limite ontologico assoluto, ma un limite contingente. Distinguere tra questi due livelli, il limite strutturale legato alla natura intrinseca del modello e il limite contingente riguardante le scelte dell'addestramento del modello, è

[6] Si specifica che la distinzione non sta nell'infallibilità dell'uomo rispetto alla macchina nella risoluzione di test logici, ma nella natura del processo che porta a tale risoluzione.

[7] Si è pienamente consapevoli che, nel panorama della ricerca del 2026, lo studio possa apparire parzialmente superato dai progressi dell'interpretabilità meccanicistica, la quale mira a tracciare il processo di ragionamento emergente nei circuiti del *residual stream*. Tuttavia, tale studio viene qui richiamato non come prova di un limite tecnico contingente, ma come indicatore epistemologico fondamentale. Come vedremo, il collasso di tali circuiti di fronte alla complessità logica conferma il motivo per cui si rese necessario l'abbandono della pretesa di completezza deduttiva, in favore di una flessibilità induttiva capace di gestire l'incertezza del reale. In questo senso, l'interpretabilità meccanicistica ci pone oggi dinanzi a un nuovo bivio: comprendere se il paradigma probabilistico sia giunto al suo limite strutturale o se sia necessaria una nuova sintesi che superi la dicotomia tra rigore logico e approssimazione statistica.

essenziale. Confondere i due livelli significa, da un lato, trarre conclusioni definitive sull'impossibilità strutturale del ragionamento artificiale a partire da fallimenti che potrebbero essere semplicemente contingenti; dall'altro, interpretare ogni miglioramento tecnico come prova che i limiti strutturali non esistono. Lo studio di Apple documenta con rigore ciò che i modelli attuali non sanno fare in determinati contesti, ma non dimostra necessariamente ciò che nessun modello potrà fare mai, né che i progressi futuri dissolveranno tutti i limiti oggi osservati. La distinzione tra le due categorie è la condizione minima per comprendere cosa l'intelligenza artificiale sia e cosa non potrà mai essere per ragioni che trascendono la tecnologia del momento. Ed è proprio questo il punto di vista sollevato dagli oppositori della tesi portata in auge da Apple, tra cui lo studioso Lawson[8], che introduce i concetti di "validità funzionale" e di "ragionamento alternativo", concetti che vedremo in seguito[9].

Entrando nel merito della ricerca[10], il *paper* analizza le prestazioni di un particolare tipo di *Large Language Model* (LLM), i *Large Reasoning Models* (LRMs), sistemi capaci di simulare processi di pensiero articolati attraverso la tecnica della "catena di pensiero"[11], nella risoluzione di diversi test logici, sollevando dubbi sulla natura dell'intelligenza di questi modelli.

[8] A. Lawson, *The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025)*, arXiv, 2025.
[9] Si è consapevoli che l'impiego del *reinforcement learning* possa suggerire il superamento di tale limite; tuttavia, come vedremo, ciò continua a non incidere sulla natura del processo, che rimane strutturalmente e ontologicamente diverso da quello umano.
[10] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.
[11] La "catena di pensiero" (*chain of thought*) è detta anche "catena di ragionamento", tipica dell'intelligenza logico-consequenzialista.

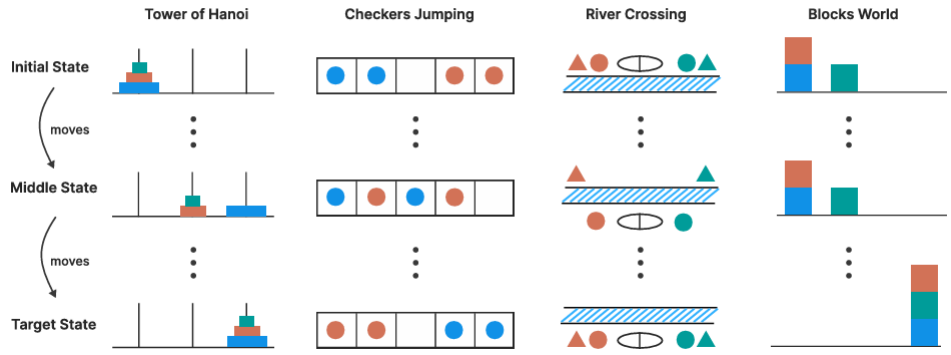


Figura 1: torre di Hanoi (tower of Hanoi), salti di pedine (checkers jumping), attraversamento del fiume o problema del traghettatore (river crossing) e costruzione con blocchi (blocks world) [12].

Come mostrato nella figura 1, i ricercatori di Apple[13] hanno usato test logici classici e scalabili come la torre di Hanoi, i salti di pedine, il problema del traghettatore con il lupo, la capra e il cavolo e la costruzione con blocchi[14]. L'approccio adottato è stato rigoroso: aumentare progressivamente la complessità mantenendo invariata la struttura logica. Il risultato è stato che tutti i modelli hanno subìto un collasso oltre una soglia critica di difficoltà: si è osservato inaspettatamente che le prestazioni di questi modelli avanzati, tra cui Claude 3.7 sonnet thinking, OpenAI o1 e o3, DeepSeek R1 e Google gemini flash thinking, calavano drasticamente all'aumentare del livello di complessità dei test[15].

L'evidenza più significativa riguarda l'andamento della deliberazione esplicita, ossia il processo attraverso il quale il modello alloca risorse in relazione al variare della complessità del compito. Tuttavia, contrariamente a quanto accade nel ragionamento umano, i modelli analizzati non mostrano una strategia adattiva. Infatti, all'aumentare della complessità del problema, non si osserva un incremento proporzionale delle

[12] Ibid.
[13] Ibid.
[14] Nella *Torre di Hanoi* lo scopo è quello di spostare i dischi tra le aste senza mai metterne uno più grande su uno più piccolo; nei *salti di pedine* lo scopo è quello di mangiare le pedine in sequenza come nella dama; nel *problema del traghettatore* il lupo, la capra e il cavolo devono essere portati dall'altra sponda del fiume senza che si mangino a vicenda, mentre nella *costruzione con blocchi* questi devono essere impilati con regole precise.
[15] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

risorse computazionali, al contrario, si ha una drastica riduzione dei *thinking tokens*, le risorse allocate internamente per riflettere[16]. Il modello, perciò, non riuscendo a mantenere la coerenza computazionale, produce risposte errate quando, invece, dovrebbe aumentare le risorse e lo sforzo. Quindi superata una certa soglia di complessità, come nel caso della torre di Hanoi al quinto disco, si manifesta questo limite architettonico intrinseco secondo il quale la curva di riflessione decresce. Ciò dimostra che l'impiego di risorse per il ragionamento, per quanto riguarda i modelli linguistici, non sia un atto di volontà ma un processo vincolato dall'architettura e dai parametri[17].

In particolare, nel test della torre di Hanoi si assiste a tre fasi di comportamento differenti da parte dei modelli. Con un grado di complessità basso, quello con pochi dischi, anche i normali LLM risolvono facilmente il test. Gli LLM standard eguagliano spesso o addirittura superano i modelli LRMs. Con un grado di complessità intermedio, gli LRMs, con i loro pensieri articolati, eccellono grazie ai *thinking tokens*, le catene di ragionamento interne. Con un grado di complessità alto, si assiste ad un collasso totale dei sistemi: gli LRMs usano meno *token* e generano risposte assurde e, anche aumentando la potenza di calcolo o il tempo di risposta dell'*output*, le prestazioni non migliorano[18].

---

[16] Ibid.

[17] Si è consapevoli che l'integrazione di strumenti esterni, pianificatori e tecniche come il *tree of thought* e il *reranking* abbiano permesso oggi di superare parzialmente tali limiti. Il *tree of thought* è una tecnica che estende la catena di ragionamento lineare in una struttura ramificata: anziché seguire un unico percorso di ragionamento sequenziale, il modello esplora in parallelo più rami di ragionamento possibili, valutandone la plausibilità e selezionando il più promettente prima di procedere (viene simulata, in un certo senso, la capacità umana di considerare alternative e tornare sui propri passi). Il *reranking* è invece una procedura postuma che genera risposte alternative, ordinandole secondo criteri di qualità, coerenza o pertinenza e selezionando quella ritenuta migliore prima di presentarla all'utente. Tuttavia, anche in questi casi, tali architetture non mutano la natura del modello linguistico: delegano funzioni di valutazione e selezione a meccanismi esterni, senza che il modello acquisisca autonomamente capacità di ragionamento logico-formale. Il problema non viene risolto alla radice, ma aggirato. Cfr. S. Yao et al., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, 2023.

[18] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

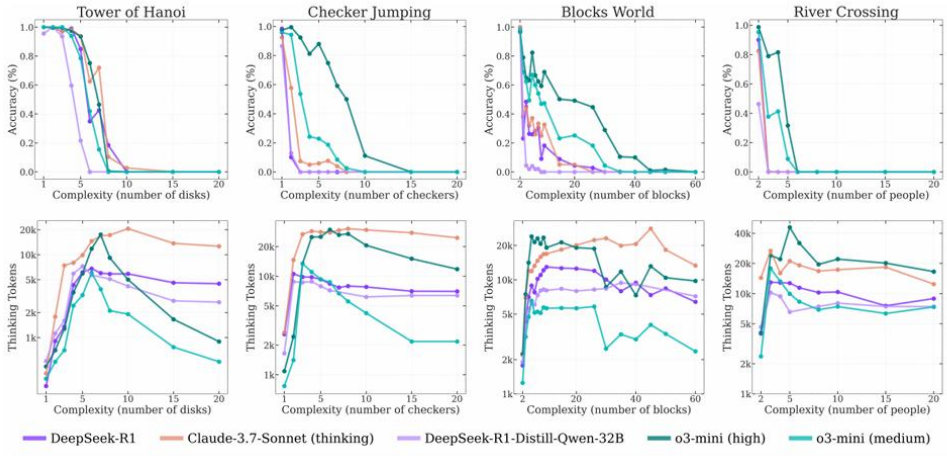


Figura 2: si osserva che in tutti i modelli le prestazioni calano drasticamente all'aumentare della complessità del test[19]

Nella figura 2, possiamo osservare questo strano fenomeno e come nei diversi casi vengono messe in relazione le due variabili dell'accuratezza della risposta dei modelli e della complessità del test. Notiamo infatti che inizialmente tutti i modelli aumentano il numero di *thinking tokens* all'aumentare della complessità ma, superata una certa soglia critica, la quantità di *token* generati inizia a diminuire. In problemi come la torre di Hanoi, si nota chiaramente che, quando l'accuratezza arriva a zero, anche lo sforzo cala drasticamente (soprattutto nella curva verde di o3 mini). In sintesi, l'immagine mostra che i modelli sembrano ragionare finché il compito è semplice, ma quando la complessità scala o il problema viene posto in maniera diversa dalla normalità, smettono di funzionare e riducono le risorse allocate nel ragionamento[20].

A questo punto i ricercatori[21] sono giunti a due conclusioni fondamentali: la prima è che il limite riscontrato sia di natura architetturale e non legato ad una semplice carenza di risorse. La seconda è che l'incapacità dei modelli di risolvere compiti logici elementari

[19] Ibid.
[20] Sebbene i recenti LLM abbiano significativamente esteso le proprie capacità computazionali e migliorato le prestazioni su test classici come la Torre di Hanoi, l'attenzione qui non è rivolta al limite numerico dei dischi risolti (oggi lo fanno anche con cinque o più dischi). Il fatto che la deliberazione cali proprio dove il ragionamento logico dovrebbe intensificarsi conferma la natura del processo. Il limite non è quantitativo ma qualitativo: un essere umano risponde alle difficoltà con una maggiore analisi semantica.
[21] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

per l'essere umano suggerisce che essi non operino un vero e proprio ragionamento, ma ne producano piuttosto una sofisticata simulazione[22].

A questo punto, il titolo *The illusion of thinking* non appare casuale, ma si configura come una presa di posizione epistemologica: l'IA non ragiona ma imita il ragionamento. Al crescere del grado di complessità o in presenza di variazioni strutturali, *out of distribution*, si osserva il collasso della coerenza logica, rivelando la natura puramente induttiva dei modelli linguistici. In altri termini, quando aumenta la complessità del problema, o il problema viene presentato in maniera differente dalla normalità, per esempio con più vincoli, ed è richiesta la capacità di astrazione e di generalizzare una soluzione tra due problemi logicamente identici ma di scala diversa, a quel punto viene svelata l'illusione: questi modelli non sono come gli esseri umani.

Come accennato precedentemente, l'ipotesi che questi modelli non ragionino regge, ma solo se si vede il ragionamento come una peculiarità unicamente umana. Infatti, se ragionare significa "sviluppare un pensiero partendo da alcune premesse e arrivando, con la logica, a una conclusione coerente e sensata"[23], allora possiamo affermare che non lo facciano nelle stesse modalità degli esseri umani[24].

1.1.2 Pappagalli stocastici o ragionamento alternativo?

I risultati dello studio Apple hanno riscosso un notevole interesse nella comunità scientifica e hanno polarizzato il mondo accademico. Lo studio ha diviso i ricercatori in due fazioni diametralmente opposte: da un lato chi vede negli LLM una reale capacità di ragionamento e rivaluta sotto una nuova lente i risultati dello studio di Apple, ritenuto metodologicamente scorretto; dall'altro chi sostiene che le capacità degli LLM non siano altro che una sofisticata forma di imitazione statistica e che dietro le loro risposte non ci sia né comprensione né ragionamento. Lo studioso Gary Marcus[25], in linea con la

[22] Ibid.
[23] Vocabolario Treccani, s.v, "ragionare", url = https://www.treccani.it/vocabolario/ragionare_res-e1c51c18-e3b1-11eb-94e0-00271042e8d9/ 30/12/2025
[24] Si distingue tra ragionamento distribuito-statistico (estrazione di pattern dai dati) e ragionamento logico-formale (deduzione tramite regole universali). Gli attuali modelli di IA appartengono alla prima categoria. Il collasso conferma che la correlazione statistica non equivale alla verità logica, evidenziando una discontinuità tra simulazione e deduzione.
[25] G. Marcus, *A Knockout Blow for LLMs?*, Substack, 2025, url = A knockout blow for LLMs? - by Gary Marcus - Marcus on AI; 30/12/2025

tesi di Apple, ha interpretato i risultati dello studio come la prova definitiva del fallimento della *scaling hypothesis*[26]*,* intesa non tanto come legge di efficienza computazionale, quanto come pretesa che l'incremento di dati e la potenza di calcolo possano fare emergere, in modo automatico, un ragionamento simile a quello umano. Quindi per Marcus la ricerca offre una prova schiacciante contro l'idea che la scalabilità, ovvero la proprietà di un sistema di incrementare le proprie prestazioni in modo proporzionale rispetto alle risorse impiegate, risolverà da sola la questione del ragionamento.

Il bersaglio critico di Marcus[27] è quindi che la scalabilità possa colmare il divario tra approssimazione statistica e necessità deduttiva. Questo perché, anche riaddestrando il modello con un numero di parametri e una capacità computazionale maggiore, non si potrebbe venire a capo del problema. Infatti, in questo scenario, il processo continuerebbe a non essere il risultato di un ragionamento, ma il frutto dell'aumento della capacità dei modelli, che avrebbero solamente imparato a imitare la soluzione sulla base dei dati dell'addestramento, risolvendo il problema per forza bruta e non per comprensione[28]. Così il miglioramento delle prestazioni non sarebbe frutto di un autentico processo logico, bensì di un potenziamento della capacità imitativa del modello.

Per comprendere meglio questo passaggio, possiamo ricorrere all'analogia dello studente: immaginiamo che il modello linguistico operi come uno studente che, anziché capire i teoremi della geometria, impari a riconoscere le regolarità ricorrenti nelle soluzioni di migliaia di problemi già svolti. In sede d'esame, lo studente troverebbe una soluzione basandosi sulla somiglianza basata sui *pattern* appresi dai problemi. Tuttavia, se il problema presentasse una minima variazione, lo studente fallirebbe poiché non ha sviluppato il principio deduttivo per comprendere l'inedito, ma solamente una

[26] J. Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv, 2020.
[27] G. Marcus, *Scale Is All You Need is dead,* Substack, 2025, url = https://garymarcus.substack.com/p/breaking-news-scale-is-all-you-need ; 30/12/2025. La critica evidenzia come la crescita delle performance non implichi necessariamente una risoluzione dei limiti logici, confermando che il *reasoning* forte non sia una proprietà emergente dello *scaling*, ma una funzione qualitativamente diversa.
[28] Ibid.

sofisticata capacità di imitazione riguardo ciò che ha già visto[29]. In quest'ottica, scalare il modello aumentando la potenza dell'IA, corrisponde semplicemente a fornire allo studente un libro con altri esempi. Sebbene questo possa aumentare la probabilità di trovare una corrispondenza nei dati, in modo che lo studente riesca a risolvere il problema, la natura del soggetto resta quella di un imitatore e non di un soggetto logico e razionale, il quale, al contrario, avrebbe scomposto il problema, trovato delle analogie strutturali e applicato la deduzione per risolverlo.

Di conseguenza, come messo in luce dallo studio Apple, ogni qualvolta la complessità dei test supera una certa soglia critica, la strategia imitativa dei modelli fallisce inevitabilmente. Tale fallimento dimostra, a tutti gli effetti, che il processo non è frutto di un autentico ragionamento logico, ma di un meccanismo basato esclusivamente sul riconoscimento di *pattern* nei dati di addestramento. Questo crollo rivela la vera natura del processo, svelando l'illusione che sta dietro al risultato ottenuto: i modelli di intelligenza artificiale generativa non ragionano mediante la logica come invece farebbe un essere umano.

Inoltre, per Gary Marcus[30], il crollo delle prestazioni dei modelli all'aumentare della complessità dei problemi fornirebbe una forte evidenza a sostegno di un'altra tesi, quella che vede gli attuali modelli di intelligenza artificiale come dei pappagalli stocastici. Questa osservazione non è altro che l'ipotesi di fondo che gli LLM non comprendono davvero il significato delle parole che usano, come dei veri e propri pappagalli. L'ipotesi dei pappagalli stocastici è già ampiamente discussa nel noto studio del 2021 *On the dangers of stochastic parrots*[31], riguardo alla quale gli studiosi concordano nel dire che i modelli linguistici si comportano come dei pappagalli perché, proprio come questi imparano a parlare, ma non sono in grado di comprendere ciò che dicono; allo stesso tempo gli LLM imparano a formulare parole e frasi ma senza comprenderne il significato. Sono definiti pappagalli "stocastici" poiché i modelli linguistici, per come sono progettati, imparano a formulare delle frasi predicendo la sequenza corretta delle parole, dette *token*, in base alla probabilità di frequenza di ogni

[29] L'IA riconosce la forma della soluzione: è la differenza tra chi risolve i problemi trovando somiglianze con ciò che conosce e chi li risolve conoscendo e applicando la deduzione e le leggi universali.
[30] G. Marcus, *A Knockout Blow for LLMs?,* p. 10.
[31] E. M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* ACM, 2021.

singola parola all'interno del set di dati d'addestramento[32]. In termini epistemologici, il modello non ragiona nel senso antropomorfico del termine, ma opera una serie di scommesse statistiche: genera la frase che risulta statisticamente più probabile in un dato contesto, agendo come un sofisticato calcolatore statistico, che imita e indovina la sequenza delle parole ma senza conoscerne davvero il significato[33].

Tuttavia, la distinzione tra mera manipolazione di simboli e reale comprensione semantica non è una problematica nata con gli attuali moderni modelli linguistici, ma affonda le radici in un dibattito filosofico consolidato ormai da decenni. La prima e più celebre risposta critica all'entusiasmo computazionale fu formulata da John Searle[34] già nel 1980 attraverso il celebre esperimento mentale della stanza cinese che ha dimostrato come la sintassi non sia di per sé condizione sufficiente per la generazione di semantica. In tempi più recenti, questa tensione dialettica tra l'efficacia prestazionale degli algoritmi e la loro effettiva natura ontologica è stata riproposta da Emily Bender e Alexander Koller[35]: secondo i due autori, la comprensione del linguaggio umano richiede necessariamente un accoppiamento, il cosiddetto *grounding*, tra la forma linguistica e il significato, inteso come relazione con entità o situazioni nel mondo reale. Quindi, dal momento che i *Large Language Models* sono addestrati esclusivamente sulla forma, rimangono confinati in una dimensione sintattica priva di intenzionalità comunicativa, di conoscenza e di legame col mondo fisico. Il lavoro di Apple *The illusion of thinking*[36], perciò, non va inteso come la scoperta di un limite inedito, ma come una rigorosa conferma empirica delle barriere teoriche già ampiamente previste dalla filosofia della mente e dalla linguistica computazionale.

---

[32] Ibid.

[33] Si riconosce come la ricerca successiva al 2021 abbia evidenziato l'emergere di rappresentazioni interne e capacità di generalizzazione estremamente sofisticate. Tuttavia, come si argomenterà in seguito, tale complessità architettonica non annulla il divario ontologico con l'essere umano, poiché i modelli rimangono privi di autentica capacità semantica e di ancoraggio alla realtà. Ad oggi infatti, nonostante l'efficienza dei risultati, la macchina continua a non attingere al significato profondo della realtà fisica e la mancata consapevolezza di tale limite strutturale può determinare ripercussioni critiche su individui, società e organizzazioni.

[34] J. R. Searle, *Minds, Brains and Programs*, Behavioral and Brain Sciences, 1980.

[35] E. M. Bender e A. Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5185-5198.

[36] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

A questo punto emerge con chiarezza la divergenza qualitativa nei processi inferenziali che caratterizzano i modelli linguistici e, più in generale, il paradigma dominante dell'intelligenza artificiale generativa. Tale distinzione non è puramente tecnica ma epistemologica, poiché si assiste al passaggio da un'architettura di tipo logico e deduttivo ad una probabilistica.

Alla luce di ciò, emerge ancor più chiaramente come i modelli analizzati da Apple operino come pappagalli stocastici. Tornando all'esempio della torre di Hanoi, se si impara la regola per spostare tre dischi, il principio non cambia se i dischi diventano cinque, sei o dieci. Mentre un essere umano applicherebbe la regola a qualsiasi scala, i modelli di intelligenza artificiale risolvono perfettamente la versione a tre o a quattro dischi non perché hanno applicato la regola, ma perché l'hanno vista moltissime volte nei loro archivi e l'hanno memorizzata. Lo stesso accade con il problema del traghettatore: se chiediamo ai modelli di risolvere la versione classica, questi rispondono perfettamente perché è presente nel loro database di addestramento. Tuttavia basta cambiare un piccolo dettaglio, magari anche solamente il numero di soggetti, e il modello ripropone la soluzione imparata a memoria, non notando che il contesto è cambiato.

Il punto cruciale che emerge dallo studio è proprio che il processo di ragionamento gli LLM non è un percorso lineare basato su regole, ma una serie di scommesse statistiche[37]. Ed è proprio per la natura del processo che i modelli falliscono. Possiamo infatti immaginare, a titolo esemplificativo, ogni passaggio logico come l'anello di una catena e se per ogni anello l'IA deve indovinare il passaggio successivo, basandosi sulla probabilità, avremo una propagazione dell'errore che alla fine della catena si amplifica notevolmente. Per fare un esempio concreto, se ogni singolo anello ha una probabilità di correttezza del novantacinque percento (0,95), in una sequenza di venti passaggi, la probabilità di successo finale crolla drasticamente a circa il trentasei percento (ovvero 0,95^20). Ciò accade perché basta un solo errore statistico in qualsiasi punto della catena per far deragliare l'intero ragionamento. Perciò gli LLM, pur essendo

[37] Il punto cruciale che emerge dallo studio è che il processo inferenziale dei modelli linguistici non segue un percorso lineare basato su regole deduttive, ma si configura come una complessa architettura di pesi probabilistici distribuiti. Come vedremo, sebbene meccanismi sofisticati come l'attenzione multi-testa (*multi head attention*) permettano al modello di mappare relazioni semantiche profonde, tale elaborazione resta pur sempre vincolata alla distribuzione statistica dei dati di addestramento.

straordinari calcolatori probabilistici, falliscono nella risoluzione di test logici: perché non sono entità logiche. La loro natura stocastica li rende eccellenti imitatori del linguaggio, ma intrinsecamente fragili quando devono risolvere problemi logici che richiedono numerosi passaggi sequenziali.

A questo punto, il ricercatore di *Open Philanthropy*[38]Alex Lawsen, con il paper *The illusion of the illusion of thinking*[39], propone una lettura più sfumata secondo cui i modelli non riproducono il ragionamento umano *tout court*, ma possono comunque sviluppare forme alternative di ragionamento. Il paper porta ironicamente un'altra firma oltre a quella di Lawsen; quella di C. Opus, in cui la C. è un riferimento sottile a Claude Opus, che è il modello di *Anthropic*, uno dei tanti *player* nel mercato dell'intelligenza artificiale.

La ricerca di Lawsen[40] mostra come le conclusioni di Apple derivino principalmente da due errori metodologici: il primo è che nella ricerca viene considerato come fallimento ciò che nella realtà dei fatti è un mero limite tecnico. I modelli non producevano più alcun *output* perché avevano raggiunto un numero massimo di *token*. Quando si chiedeva al modello di generare codice, invece di elencare ogni singola mossa, riusciva a risolvere anche problemi molto complessi come la torre di Hanoi con molti dischi, mostrando come questi modelli potessero risolvere ciò che non erano riusciti a fare nell'esperimento di Apple.

Ciò dimostra, appunto, la diversità del funzionamento della logica dell'IA rispetto alla logica umana. L'intelligenza artificiale non risolverebbe mai un problema per logica cosciente come un essere umano, né per istinto come un piccione e nemmeno per collaborazione sociale, come farebbe per esempio un cane. Come vedremo più approfonditamente in seguito, lo risolve tramite logiche probabilistiche e trasformazioni statistiche di vettori in uno spazio multidimensionale, i quali generano in *output* nuove

[38] *Open Philanthropy*, è uno dei principali finanziatori della ricerca sull'IA e finanzia *Anthropic*, una società di Intelligenza Artificiale fondata da Dario e Daniela Amodei (ex OpenAI) con l'obiettivo di sviluppare modelli linguistici avanzati focalizzandosi su affidabilità, sicurezza e trasparenza, diversamente da OpenAI, giudicata troppo commerciale e poco attenta a questi aspetti. Anthropic si distingue per un framework che mira a rendere i modelli più trasparenti, meno inclini alle allucinazioni e più rigorosi nel seguire principi etici e logici predefiniti, ponendosi in diretta concorrenza con l'approccio di OpenAI e Google.
[39] A. Lawson, *The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025),* p. 6.
[40] Ibid.

rappresentazioni vettoriali che vengono decodificate in *token*[41]. Tuttavia è fondamentale precisare che, come vedremo in seguito, oltre ai moderni LLM che operano su base puramente probabilistica, esistono sistemi di IA simbolica basati su regole logiche rigide e assiomi matematici. Pertanto, è prettamente l'IA subsimbolica che non risolve problemi attraverso la comprensione dei principi logici e set di regole astratte. [42]

Il secondo errore metodologico evidenziato da Lawsen[43] è che alcune versioni dei test erano intrinsecamente irrisolvibili, ma i modelli che riconoscevano giustamente questa impossibilità venivano considerati lo stesso come fallimentari. Sembra impossibile ma vedremo che alcuni problemi sono intrinsecamente irrisolvibili a causa di limiti fondamentali nella matematica e nella logica. Quindi una macchina non sarebbe in grado di risolvere un quesito che per sua natura non ha una soluzione logica. La differenza con un essere umano, anche in questo caso, sta nell'accorgersi di tale paradosso e di superarlo[44].

Il confronto tra la posizione di Marcus e quella di Lawsen non va quindi inteso come una semplice contrapposizione tra chi nega e chi difende la capacità di ragionamento dei modelli linguistici. Piuttosto, vengono messi in luce due diversi criteri attraverso cui viene interpretato lo stesso fenomeno: Marcus[45] valuta il comportamento dei modelli alla luce di un'idea forte di ragionamento, inteso come applicazione esplicita di regole astratte, generalizzabili e indipendenti dal contesto; Lawsen[46], al contrario, propone di sospendere questo criterio e di considerare la possibilità che esistano modalità di risoluzione dei problemi che non coincidono con il ragionamento umano, ma che non si riducono nemmeno a una mera ripetizione meccanica, introducendo il concetto di "validità funzionale" o di "ragionamento alternativo". In questa prospettiva, il dissenso non riguarda tanto i dati empirici, poiché entrambi riconoscono il crollo delle

[41] A. Vaswani et al., *Attention Is All You Need*, arXiv, 2017. Il famoso articolo descrive un sistema IA che apprende statisticamente come trasformare vettori di input in vettori di output più ricchi di informazioni, utilizzando l'attenzione per decidere quali parti del vettore siano statisticamente più rilevanti in un determinato contesto.
[42] Di tale distinzione e della sua evoluzione storica si fornirà una disamina più approfondita nel presente capitolo.
[43] A. Lawson, *The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025*), p. 6.
[44] Il problema persiste anche con i modelli di ultima generazione, vista comunque la mancanza di grounding che continua a sembrare difficilmente superabile.
[45] G. Marcus, *A Knockout Blow for LLMs?*, p. 10.
[46] A. Lawson, *The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025*), p. 6.

prestazioni al crescere della complessità, quanto l'interpretazione teorica di tale crollo. Per Marcus, il crollo smaschera definitivamente la natura imitativa del processo; per Lawsen, segnala piuttosto i limiti di un approccio valutativo che assume come standard universale il ragionamento umano. Il punto non è stabilire se i modelli capiscano, ma chiarire che cosa si intenda per comprensione e quali criteri si adottino per riconoscerla. Inoltre, al contrario di ciò che si potrebbe pensare, la posizione di Lawsen non elimina o nega la presenza di vincoli strutturali: la sua critica non mira a dimostrare che i modelli possiedano una comprensione profonda o una razionalità illimitata, bensì a ridimensionare l'idea che il fallimento in determinati compiti equivalga automaticamente all'assenza totale di qualsiasi forma di razionalità. Il riconoscimento di modelli di ragionamento alternativi, tra l'altro in perfetta linea con gli sviluppi da parte dell'etologia, suggerisce che tra il ragionamento logico umano e il pappagallo stocastico esiste uno spazio intermedio che merita di essere considerato[47].

Ciò che emerge, dunque, non è una riabilitazione incondizionata delle capacità dei modelli, ma una ridefinizione del problema: il dibattito non verte più esclusivamente sulla domanda se gli LLM ragionino o meno, bensì su quali siano i limiti strutturali entro cui il loro funzionamento deve essere compreso. Ed è proprio su questo terreno che le due prospettive, pur partendo da presupposti differenti, iniziano a mostrare una possibile convergenza. In questo senso, sia la visione critica che riduce gli LLM a pappagalli stocastici incapaci di ragionare, sia quella più ottimista che vede nel loro funzionamento forme alternative di ragionamento, concordano sull'esistenza di un limite intrinseco e strutturale. Da un lato, esiste un limite di tipo fisico poiché, anche ammettendo che i modelli possano in qualche misura ragionare, la loro efficacia resta comunque vincolata alla disponibilità di risorse materiali come potenza di calcolo, memoria, energia ed infrastrutture. In questo senso, il vincolo non è concettuale ma pratico, poiché radicato nei limiti fisici del calcolo. Dall'altro lato, sussiste un limite logico. Infatti chi sostiene la tesi dei pappagalli stocastici mette in luce un vincolo di natura diversa: la stessa architettura dei modelli, basata sulla predizione statistica dei

[47] È necessario riconoscere che l'essere umano non opera come un elaboratore logico infallibile; infatti, le nostre decisioni non sono quasi mai il frutto di una massimizzazione logica assoluta, ma sono condizionate dai limiti cognitivi della nostra mente, dal tempo ridotto e dalle informazioni parziali a nostra disposizione. In parole semplici, non comprendiamo e non agiamo solamente utilizzando la logica. Cfr. H. A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, New York, Macmillan, 1947.

token, preclude la possibilità di una comprensione semantica e non importa quante risorse vengano aggiunte; il sistema continuerà a produrre correlazioni statistiche e a non riconoscere le relazioni di causa-effetto, non accedendo quindi al livello del significato profondo.

In definitiva, sia che si critichi l'IA per la sua mancanza di comprensione, sia che la si difenda come forma alternativa di ragionamento, resta evidente la presenza di un doppio vincolo, fisico-materiale e logico-concettuale.

1.2 Dalla logica all'informatica: la scoperta di un limite intrinseco

Il problema dell'illusione del ragionamento riportato in auge da Apple, ci porta a una domanda cruciale: che cosa si intende per ragionamento e perché questi modelli, sebbene abbiano capacità eccezionali, sembrerebbero avere dei limiti intrinseci radicati nel calcolo e quindi nella matematica e nella logica?

Il caso dello studio di Apple con gli LLM e la constatazione sulla natura del ragionamento delle macchine sono nient'altro che l'evoluzione di questioni che hanno interessato matematici, informatici e filosofi per quasi un secolo. Perciò porrò l'attenzione sulla logica, sull'informatica e sulla filosofia, ovvero sulle tre discipline che hanno fornito i fondamenti teorici della disciplina e che ne hanno tracciato i confini: la logica con i suoi strumenti formali di calcolo, l'informatica con la trasformazione di principi e regole in sistemi computazionali e la filosofia con l'interrogazione sulla natura del pensiero e del ragionamento.

The illusion of Thinking[48] ha evidenziato i limiti dei *Large Reasoning Models* (LRM) nel ragionamento complesso, i vincoli fisici, quelli riguardanti i *token* e i vincoli logici, riguardanti la comprensione semantica, che sono problemi secolari riscontrati nel campo della logica, dell'informatica e della filosofia. Infatti, come vedremo, il limite evidenziato dalla ricerca di Apple non rappresenta una novità, ma è la manifestazione contemporanea di una consapevolezza che la scienza ha maturato ormai da tempo,

[48] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

ovvero che esistono problemi intrinsecamente irrisolvibili che non possono essere risolti con la logica e nemmeno superati incrementando in modo esponenziale le risorse computazionali disponibili.

A questo punto, dopo aver riconosciuto l'esistenza di un vincolo dalla doppia natura, materiale e concettuale, è necessario fare un'ulteriore assunzione. Il limite fisico, che riguarda la necessità per i modelli di intelligenza artificiale di attingere a maggiori risorse computazionali (in termini di infrastruttura, potenza di calcolo, memoria ed energia) per incrementare le loro prestazioni, riflette un principio logico di fondo: se un sistema non può risolvere determinati problemi intrinseci, ogni tentativo di valicare il limite fisico porterà solamente a degli sforzi insostenibili, senza mai superare la barriera concettuale. In tal senso, il vincolo materiale non è altro che la manifestazione pratica di un vincolo logico di fondo e perciò possiamo decidere di tralasciarlo, analizzando direttamente il limite logico che sta alla base di quello fisico e che lo genera di conseguenza.

Ma esistono davvero dei problemi intrinsecamente irrisolvibili che non possono essere superati nemmeno aumentando la potenza di calcolo all'infinito? Questa è la consapevolezza, tanto scomoda quanto fondamentale, che la scienza ha maturato da oltre un secolo.

1.2.1 Verità inaccessibili e problemi indecidibili

Il crollo delle prestazioni degli attuali modelli di intelligenza artificiale, osservato nella ricerca di Apple, e il dibattito epistemologico tra Marcus e Lawsen, sono la manifestazione di una barriera molto più profonda. Se Marcus intravede nel fallimento degli LLM la prova della loro natura puramente imitativa e Lawsen suggerisce l'esistenza di un ragionamento alternativo, quel che se ne deduce è che entrambi si scontrano, in ultima istanza, con l'esistenza di un limite intrinseco alla base della computazione stessa. Un limite che non è figlio di una carenza tecnologica attuale, ma è una proprietà fondamentale della logica e della matematica. Un limite scoperto nella prima metà del XX secolo da due giganti del pensiero, Kurt Godel e Alan Turing.

Ed è proprio la consapevolezza dell'esistenza di questa barriera concettuale invalicabile che ha segnato una svolta epocale nella storia dell'informatica: il cambio dal paradigma del determinismo logico a quello probabilistico. Non potendo costruire una macchina che fosse allo stesso tempo onnisciente e priva di errori logici, la scienza ha preferito intraprendere la via della statistica, accettando il rischio di un errore di aderenza alla realtà, in cambio di straordinari vantaggi in termini di flessibilità, gestione dell'ambiguità e scalabilità, aspetti che analizzeremo dettagliatamente nei prossimi paragrafi.

1.2.2 Godel: il sogno infranto della completezza e della coerenza

L'opera di Kurt Godel, che è universalmente riconosciuto come il padre della logica moderna, segna il momento storico in cui la ragione umana ha dovuto ammettere l'esistenza di confini insormontabili. Godel è colui che ha infranto definitivamente il sogno logicista di David Hilbert. All'inizio del Novecento, infatti, Hilbert[49], propose un programma ambizioso, con il quale aspirava a definire un sistema assiomatico perfetto: un sistema completo, capace di dimostrare ogni verità aritmetica, e coerente, cioè privo di contraddizioni interne. Questo progetto mastodontico mirava a sanare le crepe epistemologiche aperte nel 1903 dal paradosso degli insiemi di Bertrand Russell, che con la celebre variante del barbiere, "il barbiere rade tutti coloro che non si radono da soli"[50], aveva rivelato come l'autoreferenzialità potesse generare contraddizioni insanabili nei sistemi formali. Hilbert quindi sognava di formalizzare tutta la matematica in un linguaggio artificiale rigoroso, eliminando queste ambiguità e paradossi mediante una matematica che studiasse sé stessa, la cosiddetta metamatematica. In altre parole, si cercava un sistema in grado di garantire la propria validità, ovvero un sistema chiuso e perfetto nel quale ogni domanda potesse trovare una risposta certa (sì o no) e dove non ci fosse alcuna contraddizione. Era l'idea che la realtà potesse essere interamente ridotta alla logica, a un calcolo meccanico infallibile.

[49] R. Zach*, Hilbert's Program Then and Now*, Philosophy of Logic, 2007. Il Programma di Hilbert mirava a rendere la matematica *finitista*, riflettendo l'ottimismo scientifico del primo Novecento.
[50] B. Russell, *The Principles of Mathematics*, Cambridge University Press, 1903. Il paradosso di Russell analizza il paradosso del mentitore "l'insieme di tutti gli insiemi che non contengono sé stessi". Tale antinomia rivelò che la logica necessitava di una gerarchia di tipi per evitare collassi autoreferenziali.

Per scardinare questa visione, Godel[51] utilizzò proprio il paradosso dell'autoreferenzialità contro il sistema stesso per dimostrare l'impossibile: nel 1931, con i suoi teoremi di incompletezza, provò che in ogni sistema formale, sufficientemente complesso da contenere almeno l'aritmetica, esisteranno sempre proposizioni vere che il sistema stesso non potrà mai dimostrare. Dal punto di vista tecnico, costruì una frase autoreferenziale basata sulla struttura del paradosso del mentitore: G = "Questa proposizione non è dimostrabile nel sistema". Se il sistema riuscisse a dimostrare che G è vera, G diventerebbe istantaneamente falsa (poiché nega proprio la sua dimostrabilità), generando una contraddizione insanabile. Se invece G è effettivamente vera, il sistema non possiede gli strumenti formali per dimostrarlo[52].

In entrambi i casi emerge una verità fondamentale, ovvero che la verità eccede sempre la capacità di calcolo del sistema: in qualsiasi sistema basato su regole (come la matematica, un *software* o una lingua) esisteranno sempre dei fatti veri che però non possono essere provati usando solo le regole di quel sistema. Un sistema, in sostanza, non può giustificare interamente sé stesso rimanendo al proprio interno.

Come suggerito magistralmente da Douglas Hofstadter in *Godel, Escher, Bach*[53], per risolvere tali paradossi è necessario uscire dal sistema; un salto logico che una macchina logica e puramente formale non può compiere autonomamente. Hofstadter usa spesso l'esempio dei dischi che si incantano o dei *video-feedback*: una macchina, e in particolare un *software*, segue un percorso logico lineare ma, quando incontra un paradosso, resta bloccata in un *loop* infinito. L'essere umano, al contrario, ha la capacità di uscire dal sistema. Noi siamo in grado di osservare la macchina e di comprendere meta-cognitivamente di dover fare il passaggio successivo, riconoscendo che la regola è paradossale. Questa comprensione non fa parte delle regole della macchina, ma è una verità che noi cogliamo dall'esterno. Qui risiede la differenza tra "vedere" e "calcolare": calcolare, e dunque dimostrare, è un processo meccanico che procede un passo dopo l'altro seguendo i binari della logica; vedere la verità, invece, è un atto di intuizione.

[51] E. Nagel e J. R. Newman, *Godel's Proof*, New York University Press, 1958.
[52] Ibid.
[53] D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books, 1979. Douglas Hofstadter esplora il concetto di strani anelli e di come l'autoreferenzialità sia legata alla coscienza umana.

Noi esseri umani siamo in grado di vedere che qualcosa è vero anche se non abbiamo i passaggi logici per dimostrarlo[54].

Per spiegare questo concetto in termini più intuitivi, possiamo immaginare la logica come un libro di istruzioni infinitamente dettagliato che pretende di spiegare l'intero mondo. Se all'interno del libro trovassimo la frase "esistono verità che non possono essere spiegate in questo libro", saremmo davanti a un bivio epistemologico: nel caso in cui la frase fosse vera, il libro sarebbe intrinsecamente incompleto poiché ammette di non poter spiegare tutto; invece, nel momento in cui provassimo a usare le regole del libro stesso per dimostrare che quella frase è falsa, cadremmo in un errore logico. Noi esseri umani, leggendo il libro dall'esterno, vediamo che la frase è corretta, ma il libro non potrà mai confermarlo restando vincolato all'interno delle proprie pagine.

Il limite evidenziato da Godel non è solo un tecnicismo, ma una vera e propria illuminazione per il razionalismo ottocentesco, come sottolineato dal matematico e logico Piergiorgio Odifreddi[55] nelle sue ineccepibili lezioni sulla storia della logica[56]. Odifreddi evidenzia come Godel abbia dimostrato che la mente umana, e a maggior ragione una macchina che opera all'interno di un sistema prefissato, è destinata a incontrare enunciati che riconosce come veri, pur non essendo sempre in grado di dimostrarli matematicamente. Ed è proprio lo scarto tra intuizione della verità e calcolo, inteso come procedura basata sulla dimostrazione logica, che definisce la profondità del pensiero rispetto alla meccanicità dell'algoritmo[57]. Per spiegare questa differenza tra uomo e macchina, lo studioso ricorre a un'immagine molto efficace: la macchina è come un treno che corre sui binari (il sistema formale), vincolato a seguire un percorso prestabilito dalle rotaie; l'uomo invece è colui che può scendere dal treno e guardare i binari dall'alto, comprendendone la struttura e i limiti dall'esterno.

---

[54] Ibid.

[55] P. Odifreddi, *Lezioni sulla logica*, videolezione, 2018, url = https://www.youtube.com/watch?v=CW2xTxLRv_s; 04/01/2026. Nelle sue videolezioni, l'esempio del treno serve a distinguere l'esecuzione (da parte della macchina) dalla comprensione strutturale (da parte dell'uomo).

[56] Odifreddi, nelle sue numerose analisi sul lavoro di Godel, descrive il teorema di incompletezza come la prova che la verità matematica non è un territorio chiuso, ma un orizzonte aperto che nessuna "macchina per dimostrazioni" potrà mai recintare completamente.

[57] Questo percorso nasce dalla necessità di risolvere le ambiguità emerse con le geometrie non euclidee. Tali criticità spinsero i logici a utilizzare la teoria degli insiemi come base per definire rigorosamente il concetto di numero e l'intera aritmetica, riaffermando il principio cardine secondo cui l'identità della matematica risiede intrinsecamente nell'atto del dimostrare.

Ora riprendiamo il paradosso del barbiere per illustrare come si risolve questo stallo. Il dilemma "chi rade il barbiere?" è irrisolvibile per chi segue solo le regole interne: se il barbiere si radesse, violerebbe il principio di radere solo chi non lo fa da sé; se invece non si radesse, la regola lo obbligherebbe a farlo. La macchina rimarrebbe bloccata in un *loop* infinito, un errore logico ricorsivo. L'essere umano risolverebbe il problema uscendo dal sistema con creatività, intuendo di dover aggirare la logica formale e realizzando che la regola stessa è mal formulata e quindi assurda. Noi comprendiamo che per poter risolvere il paradosso dovremmo immaginare, ad esempio, che il barbiere deve essere una donna. Questa abilità si traduce nella capacità di rottura del contesto e di salto al di fuori del sistema formale che è preclusa alla logica lineare binaria.

Un possibile superamento di questa rigidità della logica binaria è offerto dalla logica *fuzzy*, in italiano logica sfumata, cioè un approccio che ha permesso lo sviluppo delle moderne IA che spesso si ritrovano a dover gestire dati incerti o contraddittori attraverso non una deduzione perfetta, ma una valutazione di plausibilità. Così, mentre la logica classica da Aristotele fino a Godel si basava sulla dicotomia netta tra vero e falso (in linguaggio binario 0 e 1), la logica *fuzzy*[58] introduce i gradi di verità: secondo questo paradigma, un'affermazione può essere parzialmente vera. Quindi, se applicata a paradossi autoreferenziali, la logica *fuzzy* permette al sistema di non collassare. Invece di oscillare all'infinito tra vero e falso, il sistema può stabilizzarsi su un valore intermedio tra 0 e 1, ad esempio, 0.5, accettando l'ambiguità come una proprietà intrinseca dell'informazione anziché come un errore fatale.

Alla luce di ciò, l'incompletezza dei sistemi formali, lungi dall'essere un concetto meramente astratto, spiega empiricamente il fallimento pratico dei primi calcolatori fino ai moderni *Large Reasoning Models*. Tali modelli, che altro non sono che l'evoluzione di sistemi formali computazionali, scontrandosi con il confine invalicabile della logica deduttiva rigida, hanno giustificato la necessità storica di muoversi verso il paradigma della probabilità. Gli LLM, operando come sistemi formali probabilistici, sono straordinari nel memorizzare e ricombinare *pattern* complessi, ma falliscono drasticamente di fronte a una complessità crescente che richieda un salto logico inedito. Tale fallimento deriva proprio dall'incapacità dei modelli di uscire dal sistema dei

[58] L. A. Zadeh, *Fuzzy Sets*, Information and Control, 1965.

propri dati di addestramento, restando prigionieri della finitezza godeliana e limitandosi a imitare la struttura formale del ragionamento, ovvero la sintassi, senza però possedere la capacità metacognitiva di trascendere le proprie regole per validare una nuova verità, la semantica.

In ultima analisi, i modelli linguistici possono simulare con estrema precisione la coerenza, ma restano confinati in un universo dove la verità è limitata a ciò che è già stato statisticamente calcolato, risultando strutturalmente incapaci di quel salto fuori dai binari che caratterizza l'autentica intelligenza umana. Inoltre, anche la facoltà stessa di generare paradossi è prerogativa dell'intelligenza umana: infatti, a differenza della macchina che commette errori sintattici, l'uomo domina il valore semantico della verità e può deliberatamente negarla o manipolarla per osservare il sistema dall'esterno[59].

1.2.3 Turing: il concetto di indecidibilità e di calcolabilità

Pochi anni dopo, Alan Turing compì il passaggio decisivo dalla logica pura all'informatica teorica. Turing viene riconosciuto universalmente come il padre dell'informatica ma viene considerato soprattutto il precursore dell'IA grazie al suo celeberrimo articolo del 1950 *Computing machinery and intelligence*[60], in cui formulava il cosiddetto "test di Turing" per verificare se una macchina potesse essere davvero considerata intelligente al pari di un essere umano. Il matematico e informatico inglese, pur non essendo un filosofo, è stato il primo studioso ad aver gettato le basi epistemologiche della computazione. Alan Turing, seguendo i passi di Kurt Godel, formalizzò il concetto di macchina universale, trasferendo i limiti della logica nell'informatica nascente. Dimostrò l'esistenza di problemi indecidibili, quesiti per i quali non può esistere alcun algoritmo capace di fornire una risposta corretta (sì o no) in

[59] Mentre l'errore della macchina è una deviazione statistica involontaria, la menzogna umana presuppone la consapevolezza della verità e l'intenzione di negarla. Questo riflette la distinzione tra sintassi e semantica: l'uomo non subisce la regola, ma la usa criticamente usando l'espediente del paradosso per trascendere il sistema formale.

[60] A. M. Turing, *Computing Machinery and Intelligence*, Mind, 1950. È interessante notare che l'articolo non fu pubblicato su una testata matematica, bensì su *Mind*, una delle più prestigiose riviste accademiche di filosofia al mondo. Questa scelta sottolinea come, sin dalle sue origini, la sfida di creare macchine pensanti sia stata considerata non solo un problema di ingegneria computazionale, ma una questione epistemologica profonda riguardante la natura della mente e della coscienza. Come già visto in precedenza, fu Searle il primo filosofo ad aver distinto calcolo e comprensione semantica.

un tempo finito: stiamo parlando del celebre *halting problem*[61]. Il problema della fermata stabilisce che è matematicamente impossibile scrivere un programma capace di determinare, per ogni possibile *input*, se un altro programma terminerà la sua esecuzione o rimarrà bloccato in un *loop* infinito, a prescindere dalle risorse computazionali disponibili. Tecnicamente, Turing formalizzò questo limite attraverso l'indecidibilità, domandandosi: dato un programma (P) e un *input* (I), esiste un algoritmo universale (H) capace di predire se P terminerà la sua esecuzione o rimarrà bloccato all'infinito? Fece ciò, attraverso una dimostrazione strutturalmente simile a quella di Godel, portando il limite dalla verità, ovvero la dimensione ontologica, all'azione, la dimensione computazionale[62].

In questo contesto, il paradosso del barbiere risulta emblematico poiché rappresenta la trasposizione logica di ciò che accade in un sistema informatico indecidibile. Se chiedessimo a una macchina di Turing di elaborare la regola del barbiere, essa entrerebbe in una ricorsione[63] infinita cercando di decidere se il barbiere debba radere sé stesso. Come già osservato, l'uomo esce dal sistema e riconosce l'assurdità della regola, mentre la macchina, che opera puramente per via sintattica, non può far altro che tentare di eseguire l'istruzione finendo per non fermarsi mai. Il paradosso del barbiere, dunque, non è solo un gioco linguistico, ma la prova che esistono istruzioni che mandano in cortocircuito la logica binaria, rendendo il comportamento di certi programmi intrinsecamente imprevedibile. In questo senso, se Godel aveva scoperto che ci sono verità inaccessibili alla logica, Turing dimostrò che ci sono compiti che nessuna macchina, per quanto potente o avanzata, potrà mai risolvere seguendo un set di istruzioni deterministiche. In altre parole, una macchina puramente logica non può calcolare tutta la complessità del reale perché esistono problemi indecidibili, ovvero

[61] A. M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 1936. Il problema della fermata, formulato da Turing nel 1936, dimostra che la computazione ha limiti invalicabili. Non esiste un algoritmo generale che possa prevedere se un calcolo giungerà a termine. Questo risultato estende l'incompletezza di Gödel alle macchine, sancendo che la capacità di elaborazione non coincide necessariamente con la capacità di risoluzione logica.

[62] Ibid.

[63] La ricorsione è la condizione per cui il sistema non raggiungerà mai una condizione di arresto, innescando un ciclo senza uscita. Godel e Turing hanno dimostrato che il problema sussiste per qualsiasi sistema formale sufficientemente potente.

interrogativi strutturalmente irrisolvibili sia per la natura intrinseca della logica, sia per i limiti operativi della macchina stessa[64].

Questo medesimo cortocircuito si ripropone, in chiave moderna, nei modelli linguistici attuali: quando cioè il modello viene posto di fronte a premesse paradossali e, non riconoscendo l'errore semantico, tenta di risolverlo restando all'interno del proprio *dataset* di addestramento. Infatti, come evidenziato nei test sui *Large Reasoning Models*, le prestazioni crollano su puzzle logici oltre una certa soglia: i LRMs falliscono proprio perché, non possedendo un algoritmo universale di risoluzione, si affidano alla ricombinazione di *pattern* memorizzati. Davanti a un problema mai visto essi rimangono prigionieri della loro finitezza.

In conclusione, se Godel limita la portata della logica, Turing limita la potenza della macchina, dimostrando l'esistenza di problemi non calcolabili. Insieme spiegano perché gli LLM falliscono sistematicamente davanti alla complessità crescente, che richiede un salto al di fuori dei binari del già noto e dei loro dati di addestramento. I modelli linguistici non possiedono capacità di astrazione per risolvere il problema, ma necessitano di una quantità sempre maggiore di dati per tentare di colmare l'indecidibilità, senza mai riuscirci del tutto.

1.2.4 Chaitin e la teoria algoritmica dell'informazione

Gregory Chaitin[65] ha formalizzato ed esteso le scoperte dei suoi predecessori attraverso la teoria algoritmica dell'informazione (AIT), andando oltre la mera constatazione che l'universo del vero sia più vasto di quello del calcolabile e introducendo il concetto rivoluzionario di "casualità algoritmica": esistono verità matematiche che non possiedono una struttura logica sottostante, ma sono puramente casuali. Si tratta di una conclusione dettata dalla natura stocastica della complessa realtà in cui siamo immersi.

---

[64] A. M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem,* p. 25.

[65] G. J. Chaitin, *A Century of Controversy over the Foundations of Mathematics*, Physica D: Nonlinear Phenomena, 2001. Chaitin esplora come l'incompletezza di Godel implichi che la matematica sia, per certi versi, una scienza sperimentale basata su fatti casuali piuttosto che su una struttura logica perfetta. Se persino la matematica cela al proprio interno una componente di casualità irriducibile, un modello linguistico, che opera come calcolatore probabilistico, non potrà mai attingere a una verità assoluta né produrre un ragionamento formalmente completo.

Il pilastro di questa teoria è la "complessità di Kolmogorov"[66], che definisce la complessità di una stringa di dati come la lunghezza del programma più breve capace di generarla. Secondo questo paradigma, una stringa è considerata casuale se è incomprimibile, ovvero se il programma più breve per esprimerla è lungo quanto la stringa stessa. Immaginiamo, ad esempio, di dover scrivere una sequenza di dati ordinata, come una stringa composta da dodici "1" consecutivi. In questo caso, non abbiamo bisogno di elencare ogni singolo numero perché ci basta dare il breve comando "stampa dodici volte 1". Questa stringa è dunque comprimibile perché esiste una regola logica semplice che può generarla senza sforzo. Al contrario, se ci trovassimo di fronte a una sequenza del tutto priva di *pattern*, come "101101001011", ci accorgeremmo che non esiste alcuna regola per riassumerla[67]. Per comunicarla a qualcuno saremmo costretti a elencare ogni singolo *bit* esattamente come appare, poiché il codice necessario per generarla deve contenere l'intera stringa stessa. Questa è quella che viene definita una verità incomprimibile: una sequenza che non può essere ridotta a una legge più semplice e che rappresenta, nel senso più puro del termine, la casualità algoritmica[68].

Chaitin[69] dimostra questa tesi attraverso la costante Ω (omega), nota anche come *numero di Chaitin.* In termini semplici, Ω rappresenta la probabilità che una macchina di Turing, alimentata con un *input* generato casualmente, giunga a una fermata invece di continuare all'infinito. La peculiarità di omega è che si tratta di un numero reale incomputabile e casuale, dal momento che i suoi *bit* non seguono alcuna legge logica e non possono essere previsti da alcun algoritmo. Ogni numero all'interno di Ω è una verità matematica a sé stante, priva di una ragione spiegabile tramite la logica, è una verità indeducibile e, di conseguenza, incalcolabile. La AIT dimostra dunque che non può esistere una teoria del tutto capace di comprimere l'intera realtà in un set finito di

[66] A. N. Kolmogorov, *Three Approaches to the Quantitative Definition of Information*, Problems of Information Transmission, 1965.
[67] Sebbene l'intelletto umano tenda a cercare *pattern* ovunque, Chaitin dimostra che la maggior parte delle verità matematiche è incomprimibile. L'esempio cardine è quello dei numeri reali: la stragrande maggioranza di essi è composta da sequenze di cifre del tutto casuali, che non possono essere generate da alcuna formula o algoritmo. A differenza di costanti come *pi greco* che, pur essendo infinite possiedono una regola di generazione, per alcuni di questi numeri non esiste alcuna legge logica capace di generarli e di predirne le cifre.
[68] G. J. Chaitin, *A Century of Controversy over the Foundations of Mathematics*, p. 26.
[69] G. J. Chaitin, *The Unknowable*, Singapore, Springer, 1999.

leggi logiche[70], dal momento che la matematica stessa contiene infinite verità incomprimibili che nessuna macchina, nemmeno con potenza di calcolo infinita, potrà mai dedurre. Come nota Chaitin, la realtà matematica contiene componenti di casualità assoluta che nessun algoritmo finito potrà mai comprimere[71].

Ciò suggerisce che il limite non riguardi solo la logica o la computazione, ma l'aderenza della matematica alla complessità del mondo reale, dove l'indeterminismo suggerisce che la struttura ultima della realtà non sia un meccanismo perfetto. Questo concetto demolisce radicalmente il presupposto secondo cui un'intelligenza artificiale possa risolvere ogni problema attraverso l'accumulo massivo di dati, poiché rivela un'inconciliabilità strutturale tra la natura del modello e la natura del reale. Dal momento che l'universo contiene verità incomprimibili, l'approccio degli LLM incontra un muro invalicabile. Il modello linguistico, infatti, vivendo esclusivamente di *pattern*, apprende la probabilità che un termine ne segua un altro basandosi su schemi ripetitivi e algoritmicamente comprimibili. La realtà, al contrario, è costellata di eccezioni e sequenze prive di schemi ricorrenti. Ed è precisamente in questo punto che gli LLM falliscono: essendo addestrati per ridurre l'immensa mole di dati in *pattern* statistici, questi modelli tentano paradossalmente di comprimere l'incomprimibile. Di fronte a un dato unico o a un'eccezione pura che non ha un *pattern* nel *dataset* di addestramento, il modello non riconosce l'assenza di una regola, ma forza il dato dentro uno schema esistente. La macchina così genera un errore che è il risultato del tentativo fallimentare di ricondurre la casualità di Chaitin alla norma statistica.

Questo percorso rivela i confini invalicabili della conoscenza informazionale: Godel ha scoperto l'esistenza delle verità indimostrabili (il limite della logica), Turing ha dimostrato l'esistenza di problemi indecidibili (il limite della macchina dovuto alla logica) e Chaitin ha provato l'esistenza di verità casuali e incomprimibili (il limite della matematica e della realtà stessa)[72]. Tale triade giustifica epistemologicamente il passaggio dal paradigma della logica a quello della probabilità, che caratterizza gli

[70] Questa visione è in linea con l'impossibilità di trovare una teoria del tutto nella fisica moderna; l'inconciliabilità tra meccanica quantistica e relatività generale riflette, su scala fisica, l'incomprimibilità logica descritta da Chaitin.
[71] G. J. Chaitin, *The Unknowable*, p. 27.
[72] Cfr. E. Nagel e J. R. Newman, *Godel's Proof*, p. 21.; A. M. Turing, *On Computable Numbers*, *with an Application to the Entscheidungsproblem*, p. 25.; G. J. Chaitin, *The Unknowable*, p. 27.

attuali modelli di IA: un sistema logico e deterministico puro è destinato a infrangersi contro l'incompletezza, l'indecidibilità e la casualità. I ricercatori hanno, perciò, dovuto compiere una scelta drastica, quella di sacrificare la certezza logica in favore della validità funzionale. L'intelligenza artificiale contemporanea ha rinunciato alla verità per la verosimiglianza.

Mentre la logica pura mirava alla completezza e finiva per collassare, la probabilità mirava, invece, alla plausibilità statistica. I modelli attuali non cercano di risolvere i limiti di Godel, Turing e Chaitin, ma scelgono di aggirarli scommettendo su ciò che è più probabile. Questo cambio di paradigma ha permesso all'IA di entrare nel mondo reale, portando vantaggi applicativi senza precedenti. Tuttavia tale successo porta con sé una fragilità intrinseca: gli LLM restano strutturalmente impreparati di fronte a quella complessità incomprimibile che richiede non una previsione statistica, ma un salto logico al di fuori del già noto.

1.3 Nasce l'Intelligenza artificiale: dalla GOFAI ai nuovi modelli di Intelligenza artificiale generativa

Ma come siamo arrivati all'intelligenza artificiale di oggi? Se il cambio di rotta dalla logica alla statistica è chiaro nelle sue motivazioni teoriche, resta da comprendere come tale evoluzione si sia concretizzata nella pratica. Si è trattato di un processo graduale o di una necessità dettata dai limiti tecnologici? Quali sono stati, di volta in volta, sia i vantaggi che le limitazioni dei paradigmi che si sono succeduti?

Per rispondere a questi quesiti, sarebbe risultato troppo didattico e sterile raccontare la storia dell'IA in modo lineare, anno dopo anno. Ho voluto quindi mettere in relazione i limiti teorici emersi nel secolo scorso con i fallimenti dell'IA che vediamo ancora oggi. Per questa ragione, ho scelto un approccio investigativo partendo proprio dallo studio di Apple[73]. Tale ricerca mostra come i modelli più avanzati entrino in crisi davanti a problemi elementari se presentati in forme insolite; un collasso che fa emergere una verità profonda sulla diversità ontologica tra il ragionamento umano e quello artificiale.

[73] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.

Quindi, a questo punto, diventa fondamentale capire come si sia sviluppato il pensiero delle macchine sulla scia dei limiti della matematica e dell'informatica. Il sogno di riprodurre meccanicamente l'intelligenza affonda le radici nell'antichità, ben prima della nascita dei moderni computer. Già nel XVII secolo, i filosofi René Descartes e Thomas Hobbes ne parlavano con intuizioni quasi visionarie. Se Cartesio ipotizzava la possibilità di poter costruire automi in grado di imitare il comportamento umano, Hobbes, con la sua celebre affermazione 'ragionare non è altro che calcolare', aveva intuito che il pensiero potesse essere ridotto a un processo meccanico. Accanto a questi due filosofi, il matematico Gotfried Wilhelm Leibniz immaginò il *calculus ratiocinator,* una macchina capace di risolvere dispute razionali come un calcolatore risolve operazioni numeriche[74].

Un passaggio cruciale avvenne nel XIX secolo con Charles Babbage e l'ausilio della contessa Ada Lovelace[75]: molto prima che Turing formalizzasse la sua macchina universale, Babbage progettò la macchina analitica un dispositivo meccanico che non si limitava al calcolo aritmetico ma che, nelle intenzioni, avrebbe dovuto elaborare simboli per svolgere compiti associati al ragionamento umano. Ada Lovelace, riconosciuta come la prima programmatrice della storia, intuì che questa macchina analitica non era solo uno strumento di calcolo, ma un sistema di manipolazione di segni generici, anticipando di fatto i concetti di *software* e programmazione. La macchina veniva così intesa come un supporto *hardware* capace di operare su qualsiasi tipo di informazione: numeri, testi, immagini o suoni.

Tuttavia l'intelligenza artificiale viene ufficialmente riconosciuta come disciplina scientifica solo nel 1956, con la conferenza al *Dartmouth College*[76], in cui si cercò di definire cosa si intendesse per intelligenza e di esplorare la possibilità futura di costruire macchine in grado di simulare i processi cognitivi umani. Organizzata da John McCarthy, Marvin Minsky, Herbert Simon e Claude Shannon, la conferenza non è considerata solo un momento inaugurale, ma un vero e proprio atto di nascita simbolica:

---

[74] M. Somalvico, *L'Intelligenza Artificiale*, Milano, Rusconi, 1987.
[75] L. F. Menabrea e A. A. Lovelace, *Sketch of the Analytical Engine Invented by Charles Babbage, Esq. with Notes by the Translator*, Scientific Memoirs, 1843.
[76] J. McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955. Il documento è stato successivamente ristampato in IA Magazine (2006).

la costruzione di macchine in grado di pensare e ragionare non è più una mera speculazione filosofica, ma un progetto scientifico concreto da realizzare.

La definizione operativa di intelligenza artificiale che emerse tra i suoi fondatori fu radicale ed è la seguente: "ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza possono essere descritti in modo così preciso che una macchina possa essere costruita per simularlo"[77].

Tale visione porta con sé l'ambizioso obiettivo di riprodurre meccanicamente l'intelligenza, un concetto non univoco che ancora oggi resta difficile da definire empiricamente.

Ritornando alla nostra ipotesi di ricerca, ci troviamo in un'epoca paradossale: disponiamo di sistemi capaci di scrivere testi e di generare codice come farebbe un essere umano, ma che evidenziano limiti importanti sull'eco dei limiti passati[78]. In quest'ottica, i limiti di Godel e Turing non sono solo capitoli di un manuale di storia, ma sono chiavi di lettura fondamentali per comprendere perché l'IA contemporanea, pur essendo incredibilmente sofisticata, sia prigioniera della sua natura sintattica[79]. E, come abbiamo visto precedentemente, la questione è che si è scelto deliberatamente che fosse così.

1.3.1 GOFAI e sistemi esperti: l'idea di un'IA simbolica e logica

Il primo paradigma che si è sviluppato a partire dalla conferenza di *Dartmouth* fu quello della GOFAI (*Good Old Fashioned AI*), termine coniato da John Haugeland[80] per descrivere l'approccio simbolico classico: la cosiddetta vecchia scuola che identifica l'intelligenza con la capacità di ragionare in modo logico, seguendo regole definite. L'ipotesi di fondo è che se il pensiero umano può essere tradotto in simboli, allora l'intera conoscenza del mondo può essere decodificata attraverso un sistema formale

[77] Ibid., p.12.
[78] È innegabile che oggi si stia progressivamente riducendo la portata di tali limiti, grazie all'evoluzione delle architetture e all'aumento della potenza computazionale. Tuttavia, come argomentato in precedenza, la riduzione quantitativa delle limitazioni tecniche non implica il superamento qualitativo dei vincoli ontologici che caratterizzano i modelli linguistici.
[79] Nonostante la fluidità del linguaggio generato, la macchina continua a operare esclusivamente su relazioni formali tra simboli, i *token*, senza mai accedere al significato profondo (la semantica) di ciò che elabora. La sofisticazione dell'*output* non deve essere confusa con la comprensione del contenuto.
[80] J. Haugeland, *Artificial Intelligence: The Very Idea*, Cambridge, MIT Press, 1985.

composto da regole. Quindi una macchina è in grado di dedurre dalle regole del sistema ogni possibile verità[81].

Si tratta di un approccio chiaramente *top down*: la conoscenza viene descritta dall'alto sotto forma di leggi universali, si definiscono le regole e infine si chiede alla macchina di applicarle ai singoli casi. I primi modelli si fondavano sull'ipotesi del sistema di simboli fisici di Newell e Simon[82], secondo cui la manipolazione di simboli in base a regole sintattiche era non solo necessaria, ma sufficiente a produrre un'intelligenza di livello umano.

L'apice di quest'approccio furono i sistemi esperti degli anni Settanta e Ottanta: *software* progettati per imitare le decisioni di uno specialista in domini circoscritti, come ad esempio la diagnosi medica. In questi ambiti il sapere veniva tradotto in vasti insiemi di istruzioni condizionali (*if-then*) e questi sistemi funzionavano in modo eccellente in ambiti specifici e ben delimitati[83]. Per esempio, per insegnare a una macchina a fare il medico, le si dava un manuale pieno di istruzioni del tipo "SE il paziente ha questi sintomi, ALLORA prescrivi questo esame" oppure "SE il paziente presenta questi sintomi, ALLORA si sospetta questa malattia".

Tuttavia questi sistemi si scontravano con il mondo reale poiché la realtà fisica non è un manuale di istruzioni statico, ma un sistema aperto e dinamico pieno di ambiguità ed eccezioni, che l'essere umano conosce. La complessità del mondo fisico si rivelò irriducibile alle regole logico-deterministiche dell'IA simbolica, come dimostra l'esempio classico dell'ornitorinco descritto da Haugeland[84]. Nella GOFAI le categorie tassonomiche sono rigide e binarie: "SE ha il becco E depone le uova ALLORA è un uccello", "SE ha il pelo E allatta ALLORA è un mammifero". L'ornitorinco[85], possedendo tratti di entrambe le categorie, rappresenta l'eccezione che rompe la regola

[81] Ibid.
[82] A. Newell e H. A. Simon, *Computer Science as Empirical Inquiry: Symbols and Search*, Communications of the ACM, 1976.
[83] Ibid.
[84] J. Haugeland, *Artificial Intelligence: The Very Idea*, p. 31. L'autore utilizza l'esempio dell'ornitorinco per dimostrare come i sistemi basati sulla logica dei predicati collassino di fronte ad entità biologiche che sfidano le definizioni atomiche.
[85] U. Eco, *Kant e l'ornitorinco*, Milano, Bompiani, 1997. L'esempio dell'ornitorinco è classico nella filosofia della scienza per descrivere il collasso delle categorie cognitive umane di fronte ad un oggetto che non rientra negli schemi predefiniti. In ambito computazionale, rappresenta l'incapacità dei sistemi simbolici di gestire l'incompletezza dell'informazione.

e una macchina basata sulla logica deterministica, davanti a tali *input*, fallisce poiché non può gestire l'ambiguità senza mandare in crisi l'intero sistema classificatorio.

Lo stesso potrebbe accadere nel tentativo di far riconoscere a un sistema esperto un gatto a partire da una foto: descriverlo esclusivamente tramite regole logiche è difficile perché cosa accadrebbe se, ad esempio, i baffi non si vedessero perché coperti? E se invece l'animale avesse le orecchie a punta o mancasse di qualche altra caratteristica? Come possiamo, di volta in volta, modificare le regole in base alle ambiguità del reale? Il difetto, quindi, non è la semplice mancanza di una regola speciale, ma l'assenza di un meccanismo di revisione interno[86]: la macchina non sa di trovarsi davanti a un'eccezione e non può percepire la realtà per correggersi[87].

Il collegamento con i limiti di Godel, Turing e Chaitin diventa evidente: così come nessun sistema formale[88] può essere allo stesso tempo completo e coerente, un sistema esperto non può codificare tutte le sfumature del mondo senza andare incontro a casi limite che ne minano la coerenza. Il limite della GOFAI risiedeva proprio nel limite della logica pura che rende il sistema rigido e incapace di evolvere. A ciò si aggiunse anche il problema del senso comune[89]: codificare ciò che un essere umano sa intuitivamente, si rivelò un compito impossibile da risolvere per via puramente simbolica. Come potremmo mai pretendere di descrivere tramite regole qualcosa che noi stessi non sappiamo decodificare razionalmente[90]?

Davanti a quella complessità che Chaitin definirebbe incomprimibile, il sistema semplicemente non funzionava. È proprio l'impatto contro questa barriera, teorica nel campo della logica e pratica nella realtà, che ha spinto la ricerca verso la svolta

---

[86] La ricerca attuale in ambito di *machine learning* e architetture neurali si sta focalizzando su questo limite attraverso meccanismi di apprendimento adattivo e gestione dell'incertezza.
[87] J. Haugeland, *Artificial Intelligence: The Very Idea*, p. 31. Per una critica alla visione di Newell e Simon cfr. H. L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason*, Harper & Row, New York, 1972, che analizza come il riconoscimento degli oggetti non sia riducibile a manipolazione sintattica di simboli.
[88] (Abbastanza potente da contenere almeno l'aritmetica).
[89] Il problema del senso comune*, tacit knowledge*, è stato uno dei principali motivi del declino della GOFAI.
[90] M. Polanyi, *The Tacit Dimension*, New York, Doubleday, 1966. Tale limite è noto in epistemologia come "paradosso di Polanyi": l'assunto secondo cui la maggior parte della conoscenza umana sia di natura tacita e non interamente formalizzabile in istruzioni esplicite o regole logiche.

probabilistica. Di fronte all'impossibilità di costruire una macchina con "intuizione godeliana", si è scelto di passare dalla logica deduttiva all'apprendimento statistico.

1.3.2 Machine learning e deep learning: dall'apprendimento supervisionato alle reti profonde

Il fallimento dei sistemi esperti nel codificare la conoscenza tacita apre le porte alla svolta connessionista: se non siamo in grado di codificare tutte le regole del mondo e di farle imparare alla macchina, dobbiamo fare in modo che la macchina stessa le apprenda autonomamente nello stesso modo in cui lo farebbe un bambino, ovvero senza che glielo si spieghi necessariamente attraverso la conoscenza esplicita. Questa forte presa di posizione e radicale intuizione filosofica cambia il paradigma dell'informatica poiché, invece di fornire alla macchina le istruzioni, le si forniscono direttamente gli esempi da cui imparare[91].

Riprendendo l'esempio del gatto, con il *machine learning* l'approccio cambia radicalmente: invece di cercare di definire cosa sia un gatto, provandone a elencare ogni singola caratteristica[92], si mostrano alla macchina migliaia di immagini di gatti. In questo modo, il sistema impara da solo a distinguere gli uni dagli altri e si passa da un modello logico e deduttivo a un modello probabilistico e induttivo. Il comando non è più "segui la regola", ma "trova il modello ricorrente". Il sistema impara a riconoscere i *pattern* di *pixel* che più spesso sono associati alla parola gatto, senza aver nessuna idea di cosa sia effettivamente un gatto nel mondo reale[93].

Per superare i limiti dei sistemi esperti, i ricercatori hanno quindi costruito modelli capaci di apprendere dalle correlazioni statistiche tra i dati. In questo scenario, le regole non vengono più scritte a priori, ma si lascia che le regole emergano durante la fase di addestramento detta *training*. Questo approccio, le cui radici teoriche risalgono agli anni Quaranta ma che ha trovato una prima applicazione negli anni Ottanta, ha subìto

[91] D. E. Rumelhart, J. L. McClelland e PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (volume 2), Cambridge, MIT Press, 1986.
[92] Cosa molto difficile perché, prima di definire tramite regole cosa sia una cosa, bisogna darne e averne a mente una definizione, ma questa definizione è soggetta a mutazioni in base al contesto, a chi la dà e in base ad altri fattori come nel caso del gatto.
[93] T. M. Mitchell, *Machine Learning*, New York, McGraw-Hill, 1997.

un'accelerazione esponenziale a partire dai primi anni Duemila, grazie alla disponibilità di enormi *dataset* e di una potenza di calcolo senza precedenti nella storia dell'uomo.

Con l'avvento del *deep learning*, basato su reti neurali profonde, l'IA compie un ulteriore salto qualitativo. Infatti a differenza del *machine learning* classico, che spesso richiede un intervento umano per identificare quali caratteristiche siano importanti, le reti profonde automatizzano questo processo, essendo strutturate su numerosi strati gerarchici in cui i primi livelli riconoscono dettagli semplici, mentre i livelli successivi combinano queste informazioni per riconoscere concetti astratti e complessi[94].

Questo passaggio segna un cambio di paradigma: mentre l'IA simbolica rappresenta l'intelligenza come puro ragionamento logico e quindi come sequenza ordinata di passi razionali, i modelli di *machine learning* e di d*eep learning* trattano l'intelligenza come capacità di imitare e generalizzare a partire dall'esperienza e dai dati. È una svolta epistemologica che prende piede dall'idea di apprendere per induzione e che sostituisce l'idea di dedurre il tutto da regole.

1.3.3 IA generativa e RAG: dal pattern matching al reasoning dinamico

Successivamente siamo passati da *deep learning* ai modelli di IA generativa: mentre il *deep learning* tradizionale eccelleva in compiti di classificazione e di discriminazione, per esempio identificare un oggetto in un'immagine, l'intelligenza artificiale generativa compie un salto qualitativo ulteriore. Non si limita a riconoscere *pattern*, ma genera nuovi contenuti a partire dai dati dell'addestramento. Ciò è stato reso possibile grazie all'aumento della profondità delle reti e alla sempre maggiore capacità di mappare le relazioni probabilistiche tra ogni elemento di un *dataset*. Nasce l'intelligenza artificiale generativa: testi, immagini, video o musica sono il prodotto di modelli che hanno appreso correlazioni statistiche in fase di *training*[95].

Tuttavia, se si guarda la linea che va dal *machine learning* al d*eep learning* fino ai grandi modelli di linguaggio, si può intravedere una certa continuità. Infatti, per quanto

[94] Y. LeCun, Y. Bengio e G. Hinton, *Deep Learning*, Nature, 2015.
[95] Ibid.

cambino architetture e dettagli tecnici, il principio resta lo stesso: i sistemi non seguono regole logiche esplicite, ma generano *output* sulla base del calcolo probabilistico. Tutto ciò, nonostante abbia permesso di affrontare problemi insormontabili per i sistemi simbolici, non ha dissolto i limiti fondamentali della logica, dal momento che li ha semplicemente trasposti su un nuovo piano concettuale.

L'IA generativa presenta questa natura problematica attraverso limiti altrettanto evidenti. Si consideri l'esempio della generazione visiva di un cane: un modello addestrato su milioni di foto impara a riprodurre l'animale con estrema precisione. Tuttavia, se nel *dataset* il cane appare prevalentemente su sfondi verdi, il modello potrebbe associare statisticamente il colore verde al concetto di cane; di conseguenza, se presentassimo al modello un contesto insolito, come ad esempio la foto di un cane con uno sfondo innevato, il sistema fallirebbe nel riconoscere il cane. Tale errore non deriva da un'incapacità di calcolo, ma dalla mancanza di esperienza reale: il modello riconosce *pattern* di *pixel* e non gli esseri viventi, dal momento che non ha una conoscenza diretta del mondo. Un limite analogo riguarda la rappresentazione della mano sinistra: molti sistemi di generazione di immagini, inizialmente, non erano in grado di rappresentare una persona che scrive con la sinistra. L'errore deriva, anche in questo caso, dal fatto che il modello aveva visto nei suoi dati di addestramento una quantità maggiore di persone destrimane rispetto a quelle mancine e quindi, di fronte alla richiesta, presentava la soluzione, statisticamente più probabile in base alla sua conoscenza, che però non corrispondeva necessariamente a quella concettualmente corretta[96].

Oggigiorno questi errori possono essere mitigati attraverso il riaddestramento su dati più bilanciati e grazie all'architettura *transformer*. Contrariamente a una visione che lega il miglioramento delle prestazioni al solo aumento della potenza di calcolo[97], casi recenti come *DeepSeek* mostrano come la ricerca si stia spostando verso l'efficienza computazionale, ottenendo gli stessi risultati o addirittura migliori con meno risorse. Un altro caso interessante è il progetto *Titans*[98] di Google che mira a superare questo collo di bottiglia attraverso architetture di memoria più leggere, dimostrando che il futuro della disciplina risiede nell'ottimizzazione del modello piuttosto che nella semplice

[96] E. M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, p. 12.
[97] J. Kaplan et al., *Scaling Laws for Neural Language Models*, p. 11.
[98] A. Behrouz et al., *Titans: Learning to Memorize at Test Time*, arXiv, 2024.

scalabilità delle risorse. Progetti come questi stanno cercando di superare il limite strutturale di tutti gli LLM che risiede nel meccanismo dell'attenzione, la cui complessità computazionale è quadratica O (n^2): raddoppiando la lunghezza del contesto, i requisiti di calcolo e memoria quadruplicano[99].

L'architettura *transformer* opera il meccanismo *encoder-decoder: l'encoder* trasforma l'*input* in una rappresentazione astratta, mentre il *decoder* traduce tale rappresentazione in un nuovo *output.* L'innovazione fondamentale risiede nel meccanismo di attenzione (*self-attention*), che permette al modello di non analizzare i dati in modo sequenziale, ma di soppesare simultaneamente l'importanza di diverse parti dell'*input*. Nel famoso articolo *Attention is all you need* [100] viene descritto come funziona nello specifico un sistema IA che apprende statisticamente come trasformare vettori di input in vettori di output più ricchi di informazioni, le quali vengono decodificate in *token*, utilizzando l'attenzione per decidere quali parti del vettore siano statisticamente più rilevanti in un determinato contesto. Un'IA prevede la sequenza di parole più coerente in un dato contesto, ma non comprende il loro significato poiché non ha alcun legame fisico con la realtà. Prendiamo l'esempio di una mela: mentre un essere umano sa cosa sia una mela e la riconosce attraverso caratteristiche come il colore rosso, il sapore dolce o la sua consistenza, un'IA non ha esperienza fenomenica di cosa sia una mela. In altri termini, l'IA non sa a cosa corrisponda il frutto nella realtà, ma ha solamente visto la parola mela associata in modo frequente a termini come "rossa", "dolce" o "mangiare". Non si tratta di reale comprensione di cosa sia una mela, ma di calcolo di prossimità tra vettori: il *token* mela è vicino ai *token* "rossa", "dolce" e "mangiare". Quindi per il modello linguistico che non possiede alcun *grounding* con la realtà, la mela non è un oggetto del mondo che presenta caratteristiche fisiche e che rimanda a un'esperienza sensoriale, ma è esclusivamente un *token* numerico all'interno di uno spazio vettoriale multidimensionale.

Detto ciò, per superare la barriera computazionale insita nell'architettura *transformer*, la ricerca si sta muovendo lungo due direttrici, ovvero la "linearizzazione dell'attenzione" e l'introduzione di "memorie dinamiche". Si sta passando da forme di attenzione dense

[99] Ibid.
[100] A. Vaswani et al., *Attention Is All You Need*, p. 16.

a forme lineari o sparse: in altri termini ogni elemento della sequenza, invece di guardare tutti gli altri come nell'architettura *standard*, guarda solamente i *token* più rilevanti, cercando di limitare i calcoli e riducendo, di conseguenza, la complessità da quadratica a lineare. In secondo luogo, progetti come *Titans* provano a simulare una memoria a breve termine, integrando il *test time memory models:* invece di ricalcolare l'intera matrice di attenzione per contesti lunghissimi che saturerebbero la memoria, questi modelli utilizzano una memoria che si aggiorna dinamicamente. Questo permette al sistema di dimenticare i dettagli matematicamente irrilevanti e di mantenere solo le informazioni cruciali[101].

Eppure l'allucinazione rimane una conseguenza intrinseca: il sistema opera per associazione sintattica senza nessuna conoscenza semantica della realtà. Accademicamente l'allucinazione è l'eco del limite che separa la forma, cioè la sintassi, dal significato, la semantica[102].

Proprio per arginare questa tendenza alla deriva statistica, la tecnologia più recente ha introdotto la RAG *(Retrieval Augmented Generation*) [103]. Per superare la natura puramente intuitiva e statistica degli LLM, spesso paragonata al sistema 1 di Daniel Kahneman[104] che è rapido e associativo ma tendente all'errore, la ricerca sta virando verso il sistema 2, ossia verso un ragionamento più lento, logico e verificabile. La RAG permette quindi di passare da un *pattern matching* statico a un ragionamento più dinamico; il sistema recupera informazioni precise prima di generare una risposta, attuando un processo di verifica su fonti esterne e verificabili. Quindi, invece di affidarsi solo alla memoria interna che può allucinare, il sistema recupera informazioni precise prima di rispondere, cercando di simulare il processo di verifica umano[105].

---

[101] A. Behrouz et al., *Titans: Learning to Memorize at Test Time*, p. 36.
[102] E. M. Bender e A. Koller*, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,* p. 13.
[103] P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, Advances in Neural Information Processing Systems, 2020.
[104] D. Kahneman, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011. L'analogia tra le architetture IA e i processi cognitivi umani è presa da Kahneman: in ambito computazionale, il sistema 1 corrisponde alla natura stocastica del *transformer*, mentre il sistema 2 è l'architettura della RAG.
[105] P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, p. 38.

1.4 Il legame con la filosofia della mente: cosa significa ragionare

Questa tesi non tenterà di risolvere l'annosa questione di cosa sia il ragionamento, dandone una definizione universale. Nonostante la filosofia della mente tenti da secoli di circoscriverne i confini, è pressoché impossibile giungere ad una definizione scientificamente rigorosa di intelligenza, ragionamento e pensiero, dal momento che la scienza deve ancora far luce su molti dei misteri fondamentali della materia. Pretendere quindi di definire l'astrazione del pensiero, prima ancora di aver compreso appieno le basi ultime della realtà fisica, rappresenterebbe un paradosso metodologico.

Tuttavia è necessario comprendere come la scelta del cambio di paradigma tecnologico sia andata di pari passo con un'intuizione filosofica di fondo: è stata proprio la scoperta del limite tecnico a sollevare le questioni epistemologiche che hanno giustificato la svolta verso la probabilità. Il percorso che porta all'attuale IA generativa affonda le radici in una precisa genealogia filosofica che ha cercato di meccanizzare la mente. Questo percorso inizia con Cartesio il quale, pur sostenendo un netto dualismo tra mente e corpo, ha aperto la strada al materialismo e al riduzionismo, considerando il corpo umano come una macchina biologica[106]. Tuttavia è solo con Thomas Hobbes che l'idea di intelligenza artificiale trova il suo primo vero e proprio fondamento. Con la sua celebre affermazione "il ragionamento non è altro che il calcolo"[107], Hobbes suggerisce che il pensiero sia una manipolazione di segni proprio come l'aritmetica. Questa intuizione ha permesso di trasformare la ragione in un processo algoritmico, rendendo teoricamente possibile la sua riproduzione su una macchina non biologica[108].

Successivamente, tra gli anni Sessanta e Settanta, il funzionalismo offre la base teorica per superare il limite fisico della materia. Hilary Putnam[109] sostiene che uno stato mentale non dipenda dal supporto fisico, che sia in carbonio per quanto riguarda il

[106] M. Somalvico, *L'Intelligenza Artificiale*, p. 30.
[107] T. Hobbes, *Leviathan*, Londra, Andrew Crooke, 1651. La frase originale recita: "Reason […] is nothing but reckoning".
[108] M. Somalvico, *L'Intelligenza Artificiale*, p. 30.
[109] H. Putnam, *Minds and Machines*, New York University Press, 1960. Putnam sostiene che uno stato mentale non dipenda dal supporto fisico ma dalla funzione che svolge, rendendo teoricamente possibile la sua realizzazione su substrati non biologici.

cervello umano oppure in silicio per un processore, ma dalla funzione che svolge. In tale prospettiva, se una macchina riceve un *input* e produce l'*output* atteso, è coerente affermare che stia ragionando: anche se il processo interno differisce da quello biologico, il risultato funzionale è equivalente[110].

Quindi in questa prospettiva, di fronte all'incapacità della logica formale di gestire l'ambiguità del reale, è avvenuta la svolta epistemologica verso il *machine learning* e il connessionismo: invece di programmare regole rigide dall'alto *(top-down*), ci si è ispirati alla struttura neurale del cervello per permettere al sistema di apprendere dal basso *(bottom-up*)[111]. Perciò la domanda fondamentale è mutata e ci si chiede se l'intelligenza sia, prima di tutto, una questione di esperienza e di induzione.

L'intuizione filosofica cambia radicalmente poiché l'intelligenza non è più mero calcolo di simboli, ma apprendimento statistico dall'esperienza e ciò emula il processo di conoscenza induttivo dell'essere umano, che non sempre segue rigide regole deduttive.

Come si è visto, questo passaggio è una scelta consapevole, in cui si riconosce che l'uomo non è solo razionalità calcolante, ma agisce per intuito, emozione e necessità biologiche. Di conseguenza abbiamo iniziato a costruire macchine che imitano i risultati del pensiero, rinunciando a imitarne la coscienza per concentrarci sul suo prodotto più visibile, ovvero il linguaggio. Tuttavia oggi ci troviamo di fronte a un nuovo paradosso secondo cui abbiamo macchine sempre più sofisticate che imitano perfettamente il linguaggio, ma che sono e rimangono, ancora oggi, ontologicamente distinte dall'essere umano.

A questo punto, per chiarire quanto la natura dei modelli sia distante dalla logica umana, è utile riportare un esempio concreto basato su un'interazione con Gemini. Chiedendo al modello di contare fino a dieci, il modello ha eseguito il compito facilmente, ma ha ammesso di non aver pensato ai numeri per farlo, sostenendo di aver solamente prodotto una sequenza appresa durante l'addestramento: "non sto calcolando, sto semplicemente restituendo un'informazione memorizzata". Al contrario, di fronte al comando iperbolico di contare fino a un miliardo, l'IA ha rinunciato ammettendo il proprio limite

[110] Ibid.
[111] D. E. Rumelhart, J. L. McClelland e PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, p. 34.

strutturale, riferendo di non poterlo fare poiché, mentre un essere umano può visualizzare il processo e mantenere la coscienza del punto di arrivo, per l'IA il conteggio è un oneroso processo di generazione di stringhe di testo e la differenza resterebbe qualitativa, anche aumentando la potenza di calcolo: l'IA calcola la risposta, l'uomo ragiona sul processo.

L'illusione del ragionamento matematico è d'altronde ben nota da tempo ed è un tema centrale nel dibattito attuale. Già tra il 2024 e il 2025, quando sono stati introdotti i *reasoning models* come OpenAI o1 e DeepSeek R1, ci si è resi conto che questi modelli hanno fin da subito migliorato drasticamente le prestazioni nei *benchmark*, ma non hanno risolto il problema di fondo. Infatti il cosiddetto *strawberry problem*, che consisteva nel contare le "r" della parola *strawberry*, ha mostrato chiaramente come questi modelli falliscano nel conteggio. L'approccio emergente infatti non punta a rendere l'IA più umana, ma a renderla ibrida: i modelli avanzati di oggi hanno iniziato, quando necessario, a delegare un compito a uno strumento esterno come una calcolatrice o ad eseguire codice in *python*[112].

Questo scenario porta a una constatazione pragmatica sulla coscienza, divisa tra due visioni classiche: la visione computazionale, quella di un'IA forte sostenuta da filosofi come Daniel Dennett[113], secondo cui la coscienza è un fenomeno emergente dalla complessità algoritmica e la visione non computazionale, quella di un'IA debole, rappresentata da John Searle nel suo esperimento della stanza cinese. Come già accennato, attraverso l'esperimento Searle dimostra che una macchina può manipolare simboli perfettamente, senza comprenderne minimamente il significato. Per l'autore infatti la coscienza è legata alla biologia e ai "qualia", ovvero le esperienze soggettive, che sono elementi che un'architettura digitale non può replicare, almeno non per ora[114].

---

[112] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4. *Python* è un linguaggio di programmazione ad alto livello, utilizzato in ambito scientifico e ingegneristico. In questo contesto, la capacità di eseguire codice in *Python* implica che il modello non tenti di risolvere un calcolo complesso attraverso un ragionamento linguistico approssimato, ma generi istruzioni formali precise che vengono eseguite da un sistema esterno. Si tratta di una forma di delega computazionale: il modello riconosce i propri limiti e affida il compito allo strumento più adatto.

[113] D. C. Dennett, *Consciousness Explained*, Boston, Little, Brown and Company, 1991.

[114] J. R. Searle, *Minds, Brains, and Programs*, p. 13.

Non potremmo mai replicare su una macchina l'intelligenza umana: la differenza fondamentale, infatti, risiede nel fatto che un essere vivente possiede un corpo biologico e che la sua intera struttura cognitiva è orientata alla sopravvivenza stessa. Ogni percezione, emozione o ragionamento ha come scopo ultimo la conservazione di un organismo mortale. Al contrario, un'intelligenza artificiale non possiede né bisogni biologici né istinto di autoconservazione, poiché lo spegnimento di un server non rappresenta una morte in senso fenomenologico, ma la mera interruzione dell'elaborazione di dati. Dunque mentre il cervello umano è il risultato di milioni di anni di evoluzione attraverso un'interazione diretta con il mondo sia fisica che sensoriale, un concetto noto come *embodiment*[115], l'intelligenza artificiale rimane un sistema non situato. L'IA resta confinata in una simulazione che imita il prodotto del pensiero senza mai davvero realizzarne il processo vitale.

Alla luce di questo percorso, in questa tesi opto per una presa di posizione chiara: la conoscenza completa e perfetta non è raggiungibile artificialmente perché esistono barriere teoriche insormontabili. È necessario smettere di chiederci se le macchine pensino e ragionino e prendere consapevolezza della loro natura da calcolatori statistici: l'intelligenza che attribuiamo loro è, come vedremo, in gran parte una proiezione umana. I *Large Language Models* sono amplificatori cognitivi eccezionali, ma restano ontologicamente distinti dalla mente umana e comprendere questa distinzione è l'unico modo per evitare l'illusione cognitiva ed utilizzare questi strumenti con consapevolezza epistemologica.

Nel secondo capitolo analizzeremo come siano i fenomeni sociali e i *bias* psicologici a spingerci ad attribuire intenzionalità ai modelli di intelligenza artificiale. Nel mio lavoro di tesi sostengo che, solo decostruendo questa narrativa antropomorfica, potremo sfruttare appieno le potenzialità delle macchine, mitigando gli effetti negativi che derivano proprio da questa mancata consapevolezza di fondo. Avendo sviscerato la vera natura di questi modelli in questo primo capitolo, nel prosieguo del lavoro, dimostrerò

[115] F. J. Varela, E. Thompson e E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MIT Press, 1991. Il concetto di *embodiment*, detto anche cognizione incarnata, suggerisce che la mente non sia un'entità astratta separata dal corpo, ma che i processi cognitivi siano profondamente modellati dalle caratteristiche fisiche dell'organismo e dalla sua interazione attiva con l'ambiente. A tal proposito, l'autrice Rucinska approfondisce come la cognizione sia legata alle *affordance*, evidenziando come l'intelligenza emerga non dal mero calcolo, ma dalla capacità di un agente situato di navigare e rispondere alle sollecitazioni del mondo fisico per fini biologici.

appunto che l'intelligenza percepita sia in larga misura il prodotto di specifiche dinamiche psicologiche e sociali e analizzerò le ricadute critiche sulla nostra interazione con la tecnologia.

CAPITOLO 2

2.1 Dalla percezione individuale alla costruzione sociale dell'intelligenza artificiale

Il primo capitolo ha operato una distinzione ontologica tra intelligenza artificiale e mente umana, tra calcolo e comprensione semantica[116]; questo secondo capitolo sposterà l'indagine sul soggetto che utilizza tale tecnologia, l'essere umano. Come argomentato da Martin Heidegger[117], conoscere la natura e i limiti di uno strumento è la condizione necessaria per un suo uso autentico e quando questa consapevolezza viene meno, la tecnica smette di essere un mezzo e diventa un ambiente che plasma il pensiero e l'azione dell'uomo stesso. Se non si coglie appieno il funzionamento della tecnologia, si rischia di sopravvalutarne o sottovalutarne le capacità rispetto a quelle umane, innescando dinamiche sociali distorte che producono effetti concreti e misurabili[118].

L'obiettivo di questo capitolo è decostruire le aspettative irrealistiche e i *bias* che accompagnano l'adozione degli attuali modelli di intelligenza artificiale. Tali distorsioni non sono fenomeni puramente tecnici: sono profondamente psicologici e sociali e possono sfociare in una sistematica sovrastima delle capacità della macchina, fino a produrre forme di delega acritica anche di fronte a clamorose allucinazioni dei modelli. Fondamentale è, in questo senso, comprendere come le percezioni distorte di una tecnologia possano trasformarsi in un rischio strutturale per gli individui e le organizzazioni che la adottano.

In questa prospettiva, vedremo che l'IA smette di essere percepita come un mero strumento tecnico per diventare a tutti gli effetti un attore che agisce all'interno della rete sociale. Analizzeremo quest'evoluzione partendo dalla definizione di agente

[116] J. R. Searle, *Minds, Brains, and Programs*, p. 13.
[117] M. Heidegger, *La questione della tecnica*, in Saggi e discorsi, Milano, Mursia, 1976. Heidegger distingue tra la tecnica antica, in cui l'artefice conosce e padroneggia lo strumento, e la tecnica moderna, in cui il mondo diventa *bestand* (riserva disponibile) e l'uomo stesso rischia di essere inglobato nel sistema tecnico senza comprenderlo.
[118] Si veda, per un inquadramento generale sul rapporto uomo-macchina, Cfr. L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Milano, Raffaello Cortina Editore, 2017.

intelligente, fondata su una razionalità puramente performativa[119], intendendo con questa espressione la capacità di un sistema di produrre *output* efficaci ed efficienti indipendentemente dal processo interno che li genera. Approderemo infine alla costruzione sociale dell'intelligenza: un paradigma in cui la realtà è la conseguenza tangibile di ciò che le persone credono sia reale e in cui il confine tra strumento e attore sociale tende progressivamente a dissolversi.

2.1.1 Il punto di partenza: comportamentismo e costruttivismo

Nel capitolo precedente abbiamo visto come i recenti modelli di intelligenza artificiale non ragionino seguendo le modalità umane e come la mente umana sembri resistere, almeno allo stato attuale della tecnica, a qualsiasi tentativo di riproduzione sintetica integrale[120]. Anche considerando i modelli classici basati su sistemi logici e ottimizzazione matematica, sussisterebbe comunque lo scarto incolmabile dell'esperienza sensoriale e della spinta biologica alla sopravvivenza, quella dimensione istintuale che ha permesso all'essere umano di sviluppare emozioni e capacità critiche strettamente legate alla propria natura di organismo mortale[121].

Questo lavoro non tenta di dirimere il tema della coscienza; infatti, non è necessario che un essere sia cosciente o biologico per essere operativamente intelligente poiché esistono macchine che sono tali in termini funzionali, riproducendo efficacemente ed efficientemente il risultato indipendentemente dal processo interno[122]. Il risultato non cambia sia che il supporto sia in carbonio o in silicio, sia che il processo sia biologico o artificiale. In questa analisi, importa solo ciò che appare e se un *output* si manifesta come razionale allora chi lo produce è intelligente[123]. Questo approccio è perfettamente valido e se decidessimo di non accettarlo dovremmo interrogarci costantemente se ogni

[119] In ambito sociologico, la *performatività* indica la capacità di un linguaggio o di una tecnologia di *creare* la realtà che descrive. L'IA è performativa perché, nel momento in cui viene definita e utilizzata come intelligente, produce effetti reali: le persone si fidano dei suoi consigli, delegano compiti critici e modificano le proprie conoscenze. In sintesi, non è importante che l'IA sia intelligente in senso umano, ma che agisca come tale.
[120] B. J. Kagan et al., *In vitro neurons learn and exhibit sentience when embodied in a simulated game-world*, in Neuron, 2022. Per quanto esistano frontiere di ricerca sull'intelligenza sintetica, che tentano di integrare neuroni biologici in sistemi computazionali, la riproduzione della coscienza resta, a parer mio, un orizzonte lontano.
[121] A. Damasio, *L'errore di Cartesio. Emozione, ragione e cervello umano*, Milano, Adelphi, 1995.
[122] H. Putnam, *Minds and Machines*, p. 39.
[123] A. M. Turing, *Computing Machinery and Intelligence*, p. 24.

essere con cui interagiamo sia cosciente e, di conseguenza, intelligente. Questa istanza è nota come il "problema delle altre menti", ma nella pratica tendiamo ad ignorarla nonostante rappresenti l'unico approccio teoricamente coerente per quanto concerne un osservatore esterno[124].

Questa prospettiva è perciò allo stesso tempo comportamentista e costruttivista e non si tratta di un espediente per evitare abilmente il problema della coscienza, ma della presa d'atto che in una società complessa come la nostra porsi domande metafisiche di tale natura rischierebbe di farci ignorare la realtà dei fatti ed il funzionamento pragmatico del mondo. Dal punto di vista comportamentista, l'intelligenza viene valutata esclusivamente sulla base del comportamento, ovvero dell'*output* osservabile: se la macchina agisce in modo indistinguibile da un altro agente, essere umano o animale che sia, la definizione stessa di intelligenza va ridimensionata all'interno del sistema sociale e relazionale in cui si manifesta[125]. Parallelamente, l'approccio costruttivista ci permette di osservare come l'intelligenza venga costruita socialmente attraverso l'interazione e la percezione degli attori sociali. Nel caso dell'IA, la sua intelligenza non emerge dalla coscienza ma è il risultato della costruzione e della negoziazione del suo valore in base all'utilità e alle aspettative percepite dagli attori sociali[126].

Di conseguenza, non dobbiamo considerare l'IA come se fosse una mente, ma dobbiamo iniziare a considerarla come un vero e proprio agente capace di agire nella società. Per compiere questo passaggio diventa necessario definire l'intelligenza in termini operativi, rinunciando al tentativo di giungere a una definizione scientifica e univoca a partire dalla concettualizzazione della coscienza.

2.1.2 L'IA come agente razionale: un'intelligenza di tipo performativo

Nel manuale di Russell e Norvig, *Artificial intelligence: a modern approach*[127] (AIMA), l'intelligenza non viene definita attraverso la coscienza, ma attraverso la "razionalità

[124] R. Descartes, *Discorso sul metodo*, Milano, Bompiani, 2004. Si richiama qui la visione meccanicistica di Cartesio sugli automi: il filosofo, per primo, pur considerando gli animali e le macchine come automi privi di anima, ammetteva l'impossibilità pratica di distinguere un automa da un uomo qualora il primo fosse riuscito a rispondere sensatamente alle parole.
[125] B. F. Skinner, *Science and Human Behavior*, New York, Macmillan, 1953.
[126] W. E. Bijker et al., *The Social Construction of Technological Systems*, Cambridge, MIT Press, 1987.
[127] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Hoboken, Pearson, 2021.

dell'azione". L'intelligenza viene qui intesa non come una qualità intrinseca di una mente o di una coscienza, ma come la capacità di un sistema di agire in modo razionale per raggiungere un obiettivo. Per i due autori[128] un agente è un'entità che percepisce il proprio ambiente attraverso sensori e agisce su di esso attraverso attuatori, al fine di raggiungere degli obiettivi specifici. In aggiunta, un agente è razionale se seleziona l'azione che si prevede possa massimizzare la sua "misura di prestazione", basandosi sulle percezioni ricevute e sulla sua conoscenza pregressa. Quindi la razionalità mira al miglior risultato atteso e include l'ottimizzazione delle risorse come parte integrante della misura di prestazione. Un agente è dunque razionale se fa la cosa giusta in base alle informazioni di cui dispone e se lo fa ottenendo il risultato desiderato possiamo dire che è stato efficace; se lo fa anche ottimizzando le risorse a disposizione è stato efficiente[129].

Questa concezione dell'intelligenza come capacità definita dagli osservatori esterni in base al risultato misurabile è in realtà molto più vicina alla nostra quotidianità di quanto si possa immaginare. Nella pratica sociale, un individuo non viene ritenuto cosciente a priori e considerato intelligente solo per il fatto di appartenere alla specie umana. Al contrario, tale etichetta gli viene attribuita se, e solo se, dimostra di saper raggiungere determinati risultati in compiti specifici, valutando non solo l'esito finale ma anche il processo che ha portato a quel risultato. Noi esseri umani misuriamo l'intelligenza quantificandola in base al binomio efficacia-efficienza: se l'obiettivo viene raggiunto, il soggetto è stato efficace e se lo ha fatto ottimizzando le risorse quali tempo, sforzo o passaggi logici, il soggetto è stato anche efficiente oltre che efficace.

Tuttavia è necessario un chiarimento. Consideriamo un professionista a cui venga richiesto di redigere un'analisi di mercato: se non portasse a termine il compito, non potremmo necessariamente concludere che non sia intelligente, dal momento che gli potrebbero mancare delle competenze tecniche, delle informazioni necessarie oppure semplicemente del tempo. Quindi la capacità di eseguire un compito e l'intelligenza non sempre coincidono poiché si può essere incapaci in un dominio tecnico pur essendo intelligenti; al contrario, si può essere capaci di eseguire un'operazione complessa senza

[128] Ibid.
[129] Ibid.

essere intelligenti nel senso pieno del termine. Questo è proprio quello che, come vedremo in seguito, Luciano Floridi chiama "*agency without intelligence*" [130], ovvero la capacità di un soggetto di produrre effetti nel mondo, di portare a termine compiti articolati e di generare *output* coerenti e apparentemente competenti, senza che dietro vi sia alcuna forma di comprensione, intenzionalità o ragionamento autentico.

Eppure, nella pratica sociale, questa distinzione viene spesso trascurata: quando valutiamo l'intelligenza di qualcuno, tendiamo a osservare se la sua azione abbia prodotto il risultato desiderato e in che modo. Ad esempio, se il report viene consegnato in maniera accurata e nel rispetto dei tempi previsti, il professionista viene giudicato competente e intelligente. Al contrario se lo consegna con errori o in ritardo, il giudizio riguardo la sua competenza ed intelligenza cambia. In tutto questo, non ci interroghiamo sui processi mentali sottostanti, ma solo sulla *performance*. Questa tendenza rivela qualcosa di importante: nella valutazione quotidiana, ciò che osserviamo non è l'intelligenza in sé ma la sua manifestazione performativa. Il problema emerge quando applichiamo lo stesso criterio ai sistemi artificiali, perché in quel caso la *performance* può essere del tutto slegata da qualsiasi forma di comprensione.

L'esempio del GPS illustra questa separazione con chiarezza. Un sistema di navigazione calcola il percorso ottimale, minimizza i tempi, aggiorna il tragitto in tempo reale e porta l'utente a destinazione: è efficace ed efficiente. Eppure, il *software* non sa cosa sia un percorso, non ha mai sperimentato un viaggio, non comprende il significato di destinazione e raggiunge l'obiettivo assegnato senza alcuna consapevolezza di ciò che sta facendo. I sistemi di intelligenza artificiale funzionano in maniera analoga: sono strumenti di calcolo statistico sofisticati, capaci di produrre *output* straordinariamente utili, ma privi di qualsiasi comprensione semantica del mondo a cui quell'*output* si riferisce[131]. Confondere la capacità di produrre risultati con l'intelligenza che normalmente associamo a quei risultati è l'equivoco fondamentale che questa tesi intende analizzare.

[130] L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, Oxford University Press, 2014.
[131] J. R. Searle, *Minds, Brains, and Programs*, p. 13.

Con Russel e Norvig questa prospettiva trova la sua sistematizzazione più influente: l'IA viene studiata come progettazione di agenti razionali, la cui intelligenza viene valutata in base al comportamento esterno e ai risultati ottenuti, invece che in relazione ai processi mentali interni o alla consapevolezza riguardo al raggiungimento di tali risultati[132]. Questo tipo di approccio è vantaggioso per lo sviluppo scientifico dell'IA poiché la razionalità è matematicamente definibile e testabile, al contrario dell'approccio, decisamente problematico, che tenta di definire e di riprodurre la coscienza che è decisamente problematico. La visione pragmatica di Russel e Norvig sposta il focus dall'introspezione al binomio efficacia-efficienza, consolidando l'idea che l'intelligenza sia una proprietà emergente dal comportamento, indipendentemente dalla natura dell'agente[133].

Tuttavia, questa definizione operativa pone una questione: se l'intelligenza viene valutata esclusivamente in base al risultato, cosa accade nella nostra mente quando un sistema produce *output* eccellenti ma senza alcuna comprensione? Ed è in questo scarto tra performance e comprensione, tra sintassi e semantica, che si innescano meccanismi psicologici e sociali.

2.1.3 Le difficoltà strutturali nel distinguere l'*agency* artificiale: la separazione tra azione e comprensione nell'era dell'IA

Secondo Luciano Floridi[134] stiamo assistendo a una storica separazione tra l'agire e l'intendere, un fenomeno che definisce come il "disaccoppiamento tra *agency* e intelligenza". Secondo il professore l'intelligenza non è più da considerarsi una facoltà legata alla coscienza o all'interiorità, ma dovrebbe essere ridotta a mero comportamento osservabile e misurabile. Questo perché siamo passati dal fare perché si capisce al fare senza capire: mentre nell'essere umano l'azione efficace è sempre stata il prodotto di

[132] D. Kahneman, *Thinking, Fast and Slow*, p. 38. Se vogliamo essere puntuali, nemmeno noi essere umani abbiamo la consapevolezza del processo sottostante al pensiero: la psicologia cognitiva evidenzia infatti come gran parte dell'agire umano derivi da processi sub-mentali privi di consapevolezza immediata. Risulta quindi paradossale esigere dall'IA questo tipo di trasparenza poiché, persino l'uomo stesso, non è in grado di documentare le proprie operazioni cognitive.
[133] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*; p. 47.
[134] L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, p. 49.

una mente cosciente l'intelligenza artificiale dimostra che è possibile distinguere l'*agency*, ovvero la capacità di agire, dall'intelligenza cioè la capacità di pensare[135].

Una macchina può quindi risolvere problemi complessi e raggiungere obiettivi specifici con un successo pari o superiore a quello umano, nonostante sia un'entità non biologica priva di consapevolezza. In questo scenario, in accordo con l'agente razionale di Russell e Norvig, ciò che conta non è più chi o cosa compie l'azione, ma esclusivamente il raggiungimento del risultato. Nel suo lavoro del 2025[136], Floridi approfondisce l'argomento, introducendo la "tesi della realizzabilità multipla dell'*agency* (MRA)" e ribadendo appunto che la capacità di agire può manifestarsi in modi differenti e può realizzarsi non solo attraverso la cognizione biologica, ma anche attraverso processi puramente computazionali e statistici. In questo senso, Floridi suggerisce che l'IA non debba essere interpretata come una riproduzione sintetica dell'intelligenza umana (secondo la tesi della *artificial realisability of intelligence ARI*), ma come la realizzazione di una nuova forma di "*agency* senza intelligenza"[137].

La sfida contemporanea consiste nel riconcettualizzare l'IA non come un tentativo di replicare l'intelligenza umana, ma come una tecnologia capace di risolvere problemi complessi tramite l'elaborazione sintattica di dati, senza la necessità di pensare e di comprendere la semantica. In tale cornice, se la macchina fa la cosa giusta, come rispondere correttamente a una mail o scrivere un codice funzionante, allora manifesta un'*agency* poiché ha successo agendo e producendo effetti nel mondo, nonostante la totale assenza di intenzionalità e consapevolezza.

Nel modello biologico tradizionale, intelligenza e intenzionalità sono intrinsecamente inseparabili: l'essere umano è un agente allo stesso tempo intelligente e intenzionale poiché le sue azioni sono sempre dirette verso uno scopo consapevole e si fondano sulla comprensione semantica del contesto[138]. Al contrario l'IA possiede esclusivamente un'intelligenza intesa come capacità computazionale, ma priva di intenzionalità poiché

[135] Ibid.
[136] L. Floridi, *AI as Agency without Intelligence On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis*, Philosophy & Technology, 2025.
[137] Ibid.
[138] J. R. Searle, *Minds, Brains, and Programs*, p. 13.

manipola simboli attraverso regole sintattiche senza comprensione profonda del significato delle parole, del contesto e del fine delle proprie azioni[139].

Il cambio di paradigma proposto da Floridi consente di superare una domanda mal posta e di comprendere come, al giorno d'oggi, non abbia più senso interrogarsi riguardo alla presunta intelligenza o intenzionalità di una macchina, ma abbia più senso invece riflettere sulla responsabilità delle proprie azioni. Il ruolo dell'essere umano diventa ancor più centrale poiché, mentre l'intelligenza artificiale possiede un'*agency* efficace ma priva di consapevolezza semantica e di comprensione del valore delle proprie azioni nel mondo, l'uomo resta, ad oggi, l'unico attore dotato di intenzionalità e capace di comprendere le conseguenze etiche delle proprie azioni. L'essere umano dovrebbe quindi governare consapevolmente l'*agency* artificiale che è priva di coscienza ma è comunque capace di impattare profondamente sulla nostra società[140].

Inoltre è necessario rendersi conto che stiamo vivendo una rivoluzione antropologica in cui l'agire è separato dall'intenzionalità e dal supporto biologico e dobbiamo imparare ad interagire con agenti intelligenti che non conoscono la realtà nel nostro stesso modo. Dobbiamo immaginare i modelli di intelligenza artificiale con la metafora che usa spesso Luciano Floridi riguardo alla lavastoviglie: l'elettrodomestico è un agente razionale che agisce nell'ambiente per raggiungere l'obiettivo desiderato di pulire i piatti e lo fa in modo estremamente efficace ed efficiente, pur rimanendo totalmente privo di coscienza e non conoscendo semanticamente cosa significhino i concetti di piatto o di pulizia[141].

Quindi, proprio come non attribuiremmo mai intenzionalità e consapevolezza a un elettrodomestico, dobbiamo accettare il paradosso di un agire senza stati mentali e intenzionalità.

In sintesi, mentre Russell e Norvig hanno posto le basi teoriche riguardo alla razionalità dell'azione, Floridi ha spostato il piano della riflessione a un livello più alto, avvertendo del rischio di non rendersi conto che modelli di intelligenza artificiale sono privi di

[139] E. M. Bender e A. Koller*, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,* p. 13.
[140] L. Floridi, *AI as Agency without Intelligence On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis*, p. 51.
[141] Ibid.

comprensione semantica o legame ontologico con la realtà fisica che li circonda, il cosiddetto *grounding*[142].

Secondo il professore, gli agenti artificiali e la loro mutua e progressiva integrazione nella società contemporanea possono essere compresi attraverso il *framework* del *enveloping*[143]. In questo senso, le tecnologie intelligenti hanno successo poiché l'ambiente circostante gli è stato "avvolto" attorno: stiamo creando microambienti in cui i compiti sono ridotti a processi eseguibili da macchine prive di comprensione, come nel caso della lavastoviglie. Inoltre, la difficoltà nel riconoscere la diversa natura epistemologica degli agenti artificiali dipende proprio dall'era in cui siamo, la quarta rivoluzione: viviamo nell'infosfera, uno spazio ibrido in cui la distinzione tra dimensione fisica e digitale sfuma e la realtà viene reinterpretata in termini informazionali, in cui ciò che è reale è informazionale e ciò che è informazionale è reale[144].

Perciò la difficoltà nel distinguere l'*agency* artificiale da quella umana sta nella natura stessa dell'ambiente informazionale in cui siamo immersi: siamo passati da una società dell'informazione a una società *data-driven*, in cui ciò che conta non è chi o cosa produce informazioni, ma la capacità di produrre ed elaborare grandi informazioni a partire da enormi volumi di dati, ciò che i sistemi artificiali fanno efficacemente attraverso l'individuazione di correlazioni e *pattern*.

Al giorno d'oggi, nell'infosfera ogni aspetto della vita può essere convertito in dato: i like sui social, i parametri biometrici e perfino le performance lavorative. Questi dati non restano isolati, ma vengono continuamente raccolti e rielaborati per creare valore e, di conseguenza, organizzati in un flusso di informazioni che influenza le scelte e i comportamenti. Quindi l'intelligenza artificiale funziona efficacemente in questo ambiente, pur essendo priva di coscienza e intenzionalità, poiché è costruita appositamente per analizzare grandi quantità di dati, trovare regolarità e restituire

[142] E. M. Bender e A. Koller*, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,* p. 13.
[143] L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, p. 49.
[144] Ibid.

risultati[145]. L'ambiente è stato progressivamente strutturato in modo da essere leggibile, gestibile e computabile attraverso dati e informazioni. Ciò ha reso l'intelligenza umana, basata su relazioni di causa ed effetto, non più unica e indispensabile per la risoluzione di problemi complessi basati sui dati[146].

Per comprendere meglio questo passaggio prenderei l'esempio di un sistema di raccomandazione che consiglia un film o una serie tv. Dal punto di vista del risultato, il suggerimento è sostanzialmente uguale a quello che darebbe un amico. Tuttavia, mentre l'essere umano comprende il significato di ciò che propone e lo fa con delle ragioni, l'algoritmo cerca correlazioni tra dati e calcola il risultato più in linea sia con le preferenze passate dell'utente, sia con le tendenze degli altri utenti. Nell'ecosistema informazionale in cui viviamo, questa differenza diventa difficile da percepire e, sul piano funzionale, diventa addirittura irrilevante poiché ciò che conta è che l'informazione prodotta sia coerente. In altri termini, la comunicazione umana si basa tradizionalmente sulla condivisione di contenuti mentali e significati, mentre l'algoritmo non è un motore semantico ma agisce in base a quella che Elena Esposito[147] chiama "contingenza virtuale": l'algoritmo si alimenta parassitariamente dei dati degli utenti sul web attraverso il "filtraggio collaborativo". Il processo di filtraggio consiste nel trovare *pattern* nei dati, ovvero correlazioni che sono inaccessibili alla mente umana ed è in questo ecosistema che la differenza semantica diventa funzionalmente irrilevante per l'utente, purché l'informazione prodotta sia funzionale e utile[148].

Tuttavia, il successo operativo dell'IA nell'infosfera nasconde la sua natura puramente sintattica e di conseguenza, come vedremo nel prossimo paragrafo, l'essere umano è

[145] C'è una distinzione sottile ma sostanziale tra dato e informazione: il dato rappresenta l'elemento grezzo e privo di significato intrinseco, mentre l'informazione è il risultato della strutturazione dei dati all'interno di un contesto semantico che conferisce loro un senso. Ad esempio, il numero "38" è un dato grezzo, senza senso, ma diventa informazione utile e significa "febbre" se inserito nel contesto del termometro. Nell'infosfera, l'IA manipola dati (gli elementi privi di semantica) per produrre informazioni (gli elementi semantici), ma senza comprenderne il significato profondo. In altri termini, l'intelligenza artificiale manipola dati, nutrendosi della materia grezza dell'infosfera e produce informazione a partire dalle correlazioni dei dati stessi. Per esempio, non sa cosa significhi "avere la febbre a 38" e "stare male", ma vede che il dato "febbre" ha un'alta probabilità di essere associato al dato "38" e al dato "stare male". Se chiediamo a un'IA cosa significhi che ho 38, restituisce l'informazione che "ho la febbre" e che "sto male", ma solo perché ha visto nei suoi dati di addestramento questa correlazione statistica.
[146] Ibid.
[147] E. Esposito, *Artificial Communication How Algorithms Produce Social Intelligence*, Cambridge, MIT Press, 2022.
[148] Ibid.

portato istintivamente ad attribuirle una mente e un'intenzionalità. Questo perché l'uomo vede, dietro l'eccellente competenza formale nel produrre il linguaggio, una capacità di pensiero e di ragionamento proporzionale all'efficacia del risultato, attribuendole di conseguenza una mente e un'intenzionalità che non possiede.

2.1.4 Le difficoltà cognitive nel riconoscere l'*agency* artificiale: pareidolia e antropomorfismo

Come abbiamo anticipato, se l'IA può essere considerata come un agente razionale sulla base dei suoi risultati operativi, resta da comprendere per quale motivo l'utente umano sia spinto a colmare il vuoto semantico della macchina con l'attribuzione di stati mentali complessi quali l'intenzionalità comunicativa, l'empatia o, nei casi più estremi, una volontà orientata a scopi e desideri propri.

Una possibile spiegazione risiede nei meccanismi cognitivi descritti dal "modello della probabilità di elaborazione (ELM)" di Petty e Cacioppo[149], secondo cui l'individuo può elaborare le informazioni attraverso un percorso centrale, analitico e razionale, oppure attraverso un percorso periferico, più rapido ed euristico. In contesti di complessità o sovraccarico informativo, l'utente tende ad affidarsi maggiormente al percorso periferico utilizzando scorciatoie cognitive. Così elementi linguistici quali scuse, espressioni empatiche o formule relazionali vengono interpretati come indizi di intenzionalità, anche se il sistema opera esclusivamente tramite calcolo statistico[150].

Nello stesso modo in cui di fronte a due punti e una linea curva tracciati su un foglio il nostro cervello vede inevitabilmente un sorriso, rievocando uno stato emotivo di felicità, di fronte a una risposta di un modello linguistico, che si scusa per un errore commesso, il nostro cervello percepisce un'autentica intenzionalità. Tuttavia, come sappiamo, non vi è alcuna intenzionalità nella risposta poiché *l'output* di scusa è esclusivamente frutto del calcolo statistico che ha predetto la sequenza di parole più probabile in base al contesto.

---

[149] R. E. Petty, J. T. Cacioppo, *Communication and Persuasion Central and Peripheral Routes to Attitude Change*, New York, Springer-Verlag, 1986.
[150] X. LI et al., *The effects of the human-like features of generative AI on usage intention and the moderating role of information overload*, Scientific Reports, 2025.

Ma quale è il motivo per il quale tendiamo ad attribuire una mente e un'intenzionalità dietro ai risultati di un algoritmo?

Quello che sappiamo è che l'essere umano è dotato di una naturale tendenza all'antropomorfismo ed è portato a proiettare caratteristiche umane quali intelligenza e intenzionalità su una struttura sintattica che sembra coerente. Questo fenomeno può essere descritto come una forma di pareidolia cognitiva: così come la nostra mente e la nostra evoluzione ci spingono a riconoscere volti umani nelle nuvole, allo stesso modo la nostra architettura cognitiva ci spinge a leggere intenzioni dietro a testi strutturati e coerenti[151].

Secondo Kyle Mahowald[152] e i suoi colleghi l'antropomorfismo nei confronti dei modelli linguistici deriva da una netta dissociazione tra due tipi di competenza e ciò trova riscontro sia nelle macchine che nel cervello umano. La padronanza della competenza linguistica non implica necessariamente il possesso anche della competenza del pensiero; viceversa, da una parte può esserci la competenza linguistica formale, quindi la padronanza delle regole grammaticali e delle regolarità statistiche di una lingua che permette di generare testi fluidi, ma dall'altra può non esserci necessariamente la competenza linguistica funzionale, quindi la capacità di usare il linguaggio per il ragionamento formale, per la risoluzione dei problemi complessi e per la comprensione del mondo reale.

Nello specifico, Mahowald[153] osserva che le capacità funzionali dei modelli linguistici sono instabili e imprevedibili[154] poiché, a differenza della competenza formale, che migliora in modo prevedibile con l'aumento dei dati, le abilità legate al ragionamento e alla comprensione non emergono spontaneamente, ma necessitano di ulteriori fasi di

[151] E. M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, p. 12.
[152] K. Mahowald et al., *Dissociating language and thought in large language models*, Trends in Cognitive Sciences, 2024.
[153] Ibid.
[154] Il cervello umano presenta un *language network*, situato prevalentemente nei lobi frontali e temporali, che si attiva selettivamente per compiti linguistici formali, mentre il ragionamento logico, il calcolo e la cognizione sociale coinvolgono circuiti distinti, come il *multiple demand network* e il *theory of mind network*. Nei modelli linguistici, analogamente, la competenza formale migliora rapidamente con l'aumento dei dati e della potenza computazionale; la competenza funzionale, invece, non emerge automaticamente dalla previsione statistica delle parole ma rimane irregolare e non sistematica.

addestramento mirate, come il *fine tuning*[155]. Questa instabilità nelle prestazioni emerge appunto in quattro domini cognitivi: nel ragionamento logico, matematico e computazionale, nella conoscenza del mondo e del senso comune, nella capacità di seguire l'evolversi della situazione nel tempo e nel ragionamento pragmatico, che richiede di interpretare intenzioni, stati mentali e contesto sociale[156].

Se ci riflettiamo, anche per quanto riguarda l'essere umano, la competenza linguistica formale e quella funzionale non emergono nello stesso modo: un bambino impara rapidamente le regole grammaticali della lingua attraverso la semplice esposizione, riuscendo a coniugare verbi, costruire frasi corrette e persino inventare parole seguendo le regolarità della lingua. Questo accade perché l'emergenza della competenza formale si sviluppa grazie all'esposizione ai dati. La competenza funzionale, invece, richiede un percorso più complesso e necessita di processi di correzione e di feedback sociali: per imparare a usare il linguaggio per ragionare, comprendere intenzioni, risolvere problemi o interpretare situazioni sociali, il bambino deve sviluppare anche altre capacità cognitive quali ad esempio la memoria di lavoro, il controllo dell'attenzione, la teoria della mente e l'esperienza.

Quindi l'abilità formale e quella funzionale, che normalmente tendiamo a considerare inseparabili, possono in realtà essere distinte e non andare di pari passo: un modello può eccellere nel produrre testi corretti e fluenti, senza comprendere ciò che sta dicendo, nello stesso modo in cui uno studente può essere in grado di recitare un testo a memoria, senza essere capace di capirlo o di spiegarne il significato. Questo accade poiché, nella quotidianità, siamo abituati a considerare la fluidità linguistica come indicatore di comprensione e intenzionalità e perciò, quando il modello eccelle nella generazione del linguaggio, noi esseri umani alluciniamo la presenza di una mente laddove in realtà c'è solo un'efficace manipolazione di dati[157].

L'essere umano, non avendo mai incontrato nella propria esperienza evolutiva un'entità capace di parlare perfettamente ma priva di comprensione del mondo, di fronte alla

[155] Il *fine-tuning* è una fase successiva all'addestramento generale del modello, nella quale viene ulteriormente specializzato su compiti specifici attraverso dati mirati o correzioni umane. Il fine-tuning permette di migliorare prestazioni particolari, come il ragionamento o la risposta a domande tecniche.
[156] Ibid.
[157] Ibid.

forma linguistica perfetta proietta automaticamente anche la funzione perfetta. Questa dissociazione alimenta quella che gli autori[158] definiscono la fallacia *good at language, good at thoug*: un errore logico secondo cui l'utente è portato a credere che un'entità in grado di generare testi coerenti e fluidi possegga necessariamente anche conoscenza e ragionamento. Parallelamente a questa, Mahowald[159] e i suoi colleghi identificano anche la fallacia contrapposta, *bad at thought, bad at language,* che consiste nel credere erroneamente che, poiché i modelli linguistici falliscono spesso in compiti legati alla logica, alla conoscenza del mondo o al senso comune, essi non possano essere considerati buoni modelli del linguaggio umano.

In altri termini chi vede, nell'eccellente struttura sintattica del linguaggio della macchina, un pensiero e un ragionamento strutturato lo fa nello stesso modo in cui chi, osservando le nuvole, riconosce il profilo di un volto umano; in entrambi i casi di pareidolia, la mente umana non si limita a ricevere passivamente un dato ma gli conferisce un senso basato sulle proprie conoscenze ed esperienze pregresse.

Al giorno d'oggi, la pareidolia è un fenomeno evidente e pervasivo poiché gli attuali modelli linguistici mostrano prestazioni sempre più eccellenti nella competenza formale: gli LLM hanno ormai raggiunto una competenza linguistica formale prossima a quella umana soprattutto in lingua inglese. Non si tratta soltanto di conoscere le regole grammaticali e di applicarle correttamente, ma di aver appreso in profondità le regolarità statistiche, la fonologia, la morfologia e la sintassi della lingua al punto tale da consentire loro di distinguere tra parole possibili, ma inesistenti, e parole impossibili e di generare testi estremamente corretti e coerenti. In questo caso, Mahowald[160] e i suoi colleghi fanno l'esempio di parole possibili ma inesistenti come *blick*, e parole formate da sequenze impossibili come *bnick*. È per questo motivo che il fenomeno della pareidolia e dell'antropomorfizzazione diviene inevitabile e rende sempre più difficile distinguere l'*agency* artificiale da quella umana.

[158] Ibid.
[159] Ibid.
[160] Ibid.

Tuttavia è emerso che, come evidenziato recentemente dallo studio di Apple[161], gli LLM, pur essendo sorprendenti nella competenza formale, falliscono in modo significativo e incoerente nella competenza funzionale, confermando che la sola percezione di fluidità del linguaggio non è un indicatore della conoscenza semantica del mondo reale. È proprio il discostamento tra competenza formale e funzionale di cui parlano i ricercatori che porta al corto circuito percettivo dell'antropomorfizzazione, nonostante sia da tempo noto in letteratura che dietro il linguaggio degli LLM non vi sia alcun *grounding* con la realtà.

2.1.5 Perché attribuiamo mente ai sistemi complessi: una strategia cognitiva

Sorge spontanea una domanda: perché attribuiamo intenzionalità e quale sarebbe l'utilità di farlo nel caso di un'intelligenza artificiale?

La risposta risiede in una precisa strategia cognitiva che Daniel Dennet[162] definisce come *intentional stance*: l'essere umano tende a semplificare la realtà e ad attribuire agli oggetti stati mentali come intenzionalità, credenze e desideri. La posizione intenzionale è una scorciatoia cognitiva che permette di spiegare il funzionamento di sistemi complessi e di prevedere il loro comportamento senza doverne decifrare la complessità sottostante.

Shanahan[163], all'interno della sua analisi sugli LLM, porta un esempio per comprendere questo passaggio, spiegando che è del tutto naturale utilizzare un linguaggio antropomorfico nelle conversazioni quotidiane riguardanti i manufatti, specialmente nel contesto delle tecnologie dell'informazione. Nella vita quotidiana usiamo continuamente la posizione intenzionale: ad esempio, quando diciamo che il nostro orologio "non si è reso conto del cambio dell'ora legale" o che "il server di posta non vuole comunicare con la rete", siamo consapevoli che l'orologio o il server non abbiano davvero una coscienza e delle intenzioni. Questo accade poiché molti sistemi, dagli orologi ai server fino ai modelli linguistici, sono governati da processi fisici o

[161] P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, p. 4.
[162] D. C. Dennett, *The Intentional Stance*, Cambridge, MIT Press, 1987.
[163] M. Shanahan, *Talking about Large Language Models*, Communications of the ACM, 2024.

matematici estremamente complessi, che spesso non conosciamo e perciò utilizziamo termini psicologici come credere o sapere per descrivere il funzionamento degli oggetti complessi di cui non conosciamo i dettagli tecnici.

L'attribuzione di scopi e razionalità agli oggetti permette quindi, non solo di spiegare i processi che non conosciamo, ma soprattutto di prevederne il loro comportamento: trattando l'oggetto come un agente intenzionale che ha degli obiettivi, come nel caso di un orologio che non vuole funzionare o un server che non vuole connettersi, possiamo mappare e prevedere le sue azioni future in base a ciò che sarebbe logico fare per raggiungere quello scopo[164].

In altre parole, se diciamo che il server non vuole comunicare, stiamo predicendo che non vedremo dati passare sulla rete e lo stiamo facendo senza che sia necessario riconoscere il guasto tecnico sottostante. In questo caso, attribuire intenzionalità all'orologio o al server, non significa credere davvero che possiedano una volontà ma adottare una scorciatoia interpretativa: trattandoli come se avessero uno scopo, possiamo spiegare il loro comportamento in termini di successo o fallimento rispetto a quell'obiettivo e, se diciamo che "il server non vuole comunicare" o che "l'orologio non vuole funzionare", stiamo facendo una previsione. Capiamo che qualcosa non sta funzionando e ci aspettiamo di conseguenza che gli oggetti non funzioneranno finché il problema non verrà risolto.

Per Shanahan[165] l'attribuzione di intenzionalità attraverso l'utilizzo consapevole di queste espressioni antropomorfe è una forma di scorciatoie (*shorthand*) innocue e utili per gestire processi complessi di cui non conosciamo o non ci curiamo dei dettagli tecnici. Sono da considerarsi innocue poiché nessuno le prenderebbe così seriamente da chiedere, ad esempio, agli oggetti di impegnarsi a fare meglio. Quindi l*'intentional stance*[166] non è altro che una strategia cognitiva conscia, che permette di riconoscere l'esistenza di un problema a partire dall'intenzionalità in base al comportamento atteso.

È necessario distinguere la prospettiva intenzionale dagli studi empirici contemporanei sull'antropomorfizzazione dell'IA poiché la posizione intenzionale, così come formulata

[164] Ibid.
[165] Ibid.
[166] D. C. Dennett, *The Intentional Stance*, p 59.

da Daniel Dennett[167], è una teoria prettamente filosofica: descrive una strategia razionale che adottiamo per prevedere il comportamento di sistemi complessi, indipendentemente dal fatto che questi possiedano effettivamente stati mentali. In questo senso, l'attribuzione di credenze e desideri è uno strumento interpretativo funzionale alla previsione; invece gli studi moderni in psicologia cognitiva e scienze comportamentali non si concentrano sulla funzione strategica e predittiva dell'attribuzione, ma sui meccanismi cognitivi che la rendono spontanea. Gli studiosi mostrano empiricamente che segnali linguistici, fluidità espressiva e tratti antropomorfici attivano automaticamente inferenze di *agency*, perfino quando l'utente è consapevole della natura puramente algoritmica del sistema[168]. In questo quadro, quindi, l'attribuzione di intenzionalità non è tanto una scelta strategica quanto una risposta cognitiva quasi automatica.

In sintesi, mentre Dennett interpreta l'attribuzione di stati mentali come una strategia razionale per semplificare la previsione del comportamento, gli studi contemporanei la descrivono come un effetto psicologico sistematico, radicato in euristiche sociali e meccanismi di percezione della mente. Una prospettiva filosofica e interpretativa da un lato e, dall'altro, una prospettiva empirica e psicologica.

Con le moderne IA, la posizione intenzionale diventa ancora più potente e decisamente più critica: se un utente chiede a un'IA quale sia la capitale dell'Italia e il sistema risponde correttamente, l'utente è portato a dire che l'IA conosceva la risposta. Tuttavia, mentre in un essere umano questa risposta deriva da un intento comunicativo e dalla comprensione della realtà geografica, l'IA ha semplicemente calcolato la sequenza di parole statisticamente più probabile in base alla domanda che gli è stata posta. In questo caso la posizione intenzionale rischia di farci attribuire al sistema capacità di comprensione che tecnicamente il modello non possiede[169].

In senso interpretativo, quindi, l'attribuzione di una mente alla macchina è una scorciatoia cognitiva necessaria poiché, non potendo spiegare istantaneamente la complessità algoritmica sottostante, il nostro cervello semplifica la questione

[167] Ibid.
[168] M. Jakesch et al., *Co-Writing with Opinionated Language Models Affects Users' Views*, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023.
[169] K. Mahowald et al., *Dissociating Language and Thought in Large Language Models*, p. 56.

proiettando sull'IA scopi e intenzioni tipicamente umani. Quindi ad esempio, se il modello linguistico risponde in modo coerente, è più utile e funzionale a fini conversazionali considerare che lo abbia fatto con un'intenzione precisa, invece che ritenere che quel risultato sia l'esito di un calcolo probabilistico estremamente astratto. Perciò l'attribuzione di stati mentali all'IA non è un errore cognitivo ma una strategia pragmatica: l'antropomorfismo ci permette di utilizzare la tecnologia senza doverne decifrare i meccanismi interni e, trattare il modello linguistico come se avesse una mente, ci consente di usare la tecnologia con la stessa fluidità con cui interagiremmo con un altro essere umano, rendendo di fatto l'interazione immediata.

Ci troviamo di fronte a un paradosso tipico della modernità, il paradigma in cui l'efficacia operativa degli strumenti tecnologici ha sostituito la necessità di comprendere la loro natura tecnica: per interagire con successo con un sistema complesso, non è più richiesto il possesso di competenze tecniche specifiche. Come al passeggero non è richiesta per volare la conoscenza tecnica di un aereo né tantomeno quella delle leggi dell'aerodinamica, nello stesso modo non è necessario conoscere i meccanismi interni dell'IA per poterla utilizzare. La difficoltà per l'utente comune di comprendere i complessi calcoli sottostanti alla generazione del linguaggio diventa irrilevante e disfunzionale di fronte alla sua utilità quotidiana. Martin Heidegger[170] utilizza la metafora dell'aereo per spiegare il passaggio dalla tecnica antica a quella moderna, sostenendo che, mentre in passato la tecnologia era uno strumento che l'uomo padroneggiava conoscendone i segreti, la tecnica moderna ha trasformato la realtà in *bestand*, una riserva sempre disponibile. In questa prospettiva le cose smettono quindi di essere oggetti complessi da indagare e diventano semplici risorse pronte all'uso per ottenere risultati immediati.

Lo stesso accade oggi con l'IA: la sua disponibilità ed estrema facilità d'utilizzo ci spinge a smettere di interrogarci sul suo funzionamento interno, basato sulla statistica e i dati. Tuttavia le radici di questo meccanismo sono più profonde di quanto suggerisca la sola abitudine tecnologica e affondano nella biologia evolutiva del cervello umano.

[170] M. Heidegger, *The Question Concerning Technology*, The Question Concerning Technology and Other Essays, New York, Harper & Row, 1977.

Questo fenomeno non è nuovo, né esclusivo dell'intelligenza artificiale. Anthony Giddens, ne *Le conseguenze della modernità*[171], lo aveva già identificato come una delle caratteristiche strutturali della modernità avanzata, introducendo il concetto di "fiducia nei sistemi esperti". Con questa espressione Giddens si riferisce a quei sistemi di competenza tecnica o professionale che organizzano ampie porzioni della nostra vita quotidiana sulla base di saperi che la maggior parte delle persone non possiede e non è in grado di verificare. La modernità, osserva Giddens, richiede che gli individui ripongano fiducia in meccanismi che non comprendono, delegando a sistemi impersonali la garanzia del loro funzionamento[172]. Non ci fidiamo del pilota che non conosciamo, ma ci fidiamo del sistema di certificazione che lo ha abilitato.

2.1.6 L'attribuzione di intenzionalità come meccanismo evolutivo inconscio: HADD, HIDD e il ruolo delle *affordance*

Gli psicologi evoluzionisti spiegano la nostra tendenza ad antropomorfizzare la realtà attraverso il concetto di *hyperactive agency detection device* (HADD) [173]. Secondo questa teoria, il sistema cognitivo umano possiede un dispositivo di rilevamento dell'*agency* (ADD) che opera in modo "iperattivo", definito anche frammentario o "a scatto" (*trip wired*), e porta a percepire un'*agency* a partire da segnali ambigui nell'ambiente circostante. In altri termini, l'essere umano rileva agenti poiché possiede tale dispositivo specializzato nel rilevamento, pur avendo a disposizioni informazioni incomplete. Questo accade poiché il nostro cervello opera a un livello non riflessivo e produce intuizioni automatiche che precedono la valutazione riflessiva. L'evoluzione ha fatto in modo che il cervello abbia imparato nel tempo a riconoscere agenti per istinto di sopravvivenza e per convivenza sociale[174].

---

[171] A. Giddens, *Le conseguenze della modernità*, Bologna, Il Mulino, 1994.
[172] Ibid.
[173] A. Lisdorf, *What's HIDD'n in the HADD?*, The Cognitive Science of Religion and the Problem of Methodological Naturalism, Journal of Cognition and Culture, 2007. Il concetto originale di HADD è stato introdotto dallo psicologo Justin L. Barrett per spiegare la tendenza evolutiva a percepire l'agency anche in oggetti inanimati.
[174] Ibid.

Barrett[175] e Guthrie[176] sostengono che, dal punto di vista evolutivo, la funzione originaria di questo meccanismo era individuare i predatori in ambienti ostili. Per capire come il nostro cervello si comporti come un dispositivo per il rilevamento iperattivo di agenti, analizzano la differenza tra credenze riflessive e non riflessive, attraverso il caso del "falso positivo" e del "falso negativo"[177]. Immaginiamo due ominidi nella savana che sentono un fruscio nell'erba e devono valutare se ci sia il pericolo di essere attaccati da un predatore. Da una parte l'ominide analitico ragiona appunto in modo analitico e riflessivo e, sapendo che il fruscio dell'erba nella maggior parte dei casi è dovuto al vento, aspetta di avere più informazioni dall'ambiente per capire se scappare. Quello che fa è cercare di rispondere alla domanda "come mai accade?", cercando cause meccaniche come il vento. Tuttavia questa lentezza nel processare le informazioni potrebbe essergli fatale se dietro il fruscio ci fosse davvero un predatore. Dall'altra parte, l'ominide non riflessivo, grazie al rilevamento HADD, risponde istantaneamente alla domanda rilevando l'agente e attribuendo immediatamente un'*agency* dietro al movimento del vento. In realtà quello che fa non è solo rilevare la presenza dell'agente che ha causato l'azione ma, successivamente, attribuire anche un'intenzione all'oggetto inanimato che l'ha causata. Grazie a ciò sospetta che ci sia qualcuno che vuole mangiarlo dietro al fruscio e decide di scappare immediatamente. Quindi il suo dispositivo iperattivo gli ha indicato che il fruscio è un rischio che richiede un'azione immediata, la fuga[178].

In tale contesto, il costo del falso positivo (scambiare il fruscio del vento per un predatore) era minimo rispetto al rischio letale di un falso negativo (scambiare un predatore per il vento). Infatti analizzando entrambi gli scenari, se non ci fosse nessun predatore dietro al fruscio dell'erba, allora l'ominide non riflessivo sarebbe scappato inutilmente, ma lo avrebbe fatto a un costo bassissimo; al contrario, se dietro al fruscio dell'erba ci fosse davvero un predatore, l'ominide analitico sarebbe morto, con un costo dell'errore decisamente più alto. Attribuire *agency* a ciò che si muove nell'ambiente circostante è stata da sempre una strategia di sopravvivenza a basso costo ma ad

[175] J. L. Barret, *Why Would Anyone Believe in God?*, Walnut Creek, AltaMira Press, 2004.
[176] S. E. Guthrie, *Faces in the Clouds A New Theory of Religion*, New York, Oxford University Press, 1993.
[177] M. G. Haselton et al., *Resolving the Logic of Emotion Supposition and Superstition in the Evolutionary Arms Race that Shaped the Emotions*, Oxford, Oxford University Press, 2007.
[178] A. Lisdorf, *What's HIDD'n in the HADD*?, p. 63.

altissimo rendimento e garantisce la sopravvivenza in assenza di una comprensione immediata dei fenomeni fisici[179].

A questo punto, è necessario introdurre la posizione di alcuni ricercatori, come Anders Lisdorf[180], che suggeriscono che sia più corretto perfezionare il modello dell'HADD attraverso l'introduzione del concetto di HIDD, *hyperactive intentionality detection device,* in cui "agente*"* viene sostituito con "intenzionalità". Lisdorf sostiene che il nostro cervello è strutturato per cercare menti e intenzioni dietro i fenomeni e osserva come la teoria dell'HADD si leghi direttamente alla posizione intenzionale di Dennet, ma lo fa in modo incompleto: mentre l'HADD ci permette di percepire la presenza fisica di un agente, l'HIDD è il meccanismo che ci spinge ad attribuire a quell'agente scopi, pensieri e intenzioni.

Secondo Lisdorf quindi il concetto di HADD si evolve naturalmente in HIDD, che altro non è che la base biologica della posizione intenzionale di Dennet, la strategia razionale secondo la quale trattiamo un oggetto come se fosse un agente razionale dotato di credenze e desideri, al fine di prevederne il comportamento in modo rapido ed efficace. In sintesi, l'HADD sarebbe la versione inconscia e iperattiva dell'HIDD, il fondamento biologico della posizione intenzionale. Tale dispositivo sarebbe collegato alle aree cerebrali coinvolte nella mentalizzazione, come la corteccia cingolata posteriore (PCC), il che spiega perché, di fronte a un'IA capace di esprimersi in modo fluente, tendiamo spontaneamente a trattarla come un soggetto pensante prima ancora che l'analisi razionale, e quindi la posizione intenzionale, intervenga a ridimensionare questa impressione. Quindi l'HADD rileva rapidamente la possibile presenza di un agente in situazioni ambigue in quanto strategia evolutiva di sopravvivenza, ma è l'HIDD che permette di attribuire consciamente all'agente intenzioni, scopi o significati nascosti, interpretando quell'azione come dotata di senso e trasformando un semplice evento in un'opportunità o in una minaccia significativa[181].

[179] M. G. Haselton et al., *Resolving the Logic of Emotion Supposition and Superstition in the Evolutionary Arms Race that Shaped the Emotions*, p. 64.
[180] A. Lisdorf, *What's HIDD'n in the HADD*?, p. 63.
[181] Ibid.

Perciò, nel suo saggio *What's HIDD'n in the HADD?*[182], Lisdorf mette in luce la differenza tra rilevare (la funzione percettiva dell'HADD) e attribuire intenzione (la funzione cognitiva della corteccia cingolata posteriore), proponendo quattro esempi diversi come *benchmark* per la sua analisi[183], ma non ricorrendo all'esempio dei due ominidi per spiegare l'attribuzione di *agency*. L'autore opera questa scelta poiché sostiene che, basandosi esclusivamente sul movimento fisico, non si spiegherebbero i casi in cui l'uomo interagisce con entità statiche. Infatti la logica dell'HADD spiega perché abbiamo paura dei fantasmi o dei mostri nel buio, ma non spiega altrettanto bene perché preghiamo un Dio invisibile o leggiamo un libro sacro trovandovi un senso profondo e una finalità. La preghiera non nasce da una reazione ad uno stimolo inconscio e improvviso ma è un atto calmo, intenzionale e privo di stimoli fisici[184]. A conferma di ciò, esistono numerosi studi neurologici che confermano che la PCC si attivi anche senza stimoli esterni[185].

Alla luce di ciò, è chiaro che la strategia intenzionale funzioni sfruttando, a livello conscio, proprio il dispositivo biologico di rilevamento dell'*agency* e più propriamente dell'intenzionalità, che permette al nostro cervello di riconoscere istantaneamente rischi o opportunità nell'ambiente circostante, senza doverli analizzare in modo riflessivo e lento. Il processo che permette al cervello umano di operare in modo automatico e non riflessivo, identificando istantaneamente ciò che l'ambiente offre in termini di azione, si lega strettamente al concetto di *affordance.* Le *affordance* sono appunto le possibilità o gli inviti all'azione che un organismo percepisce nel proprio ambiente e gli esseri umani, come tutti gli organismi, sono in grado di identificarle e sfruttarle[186].

Ne esistono di due tipi. Riprendendo l'esempio dell'ominide, abbiamo un*'affordance* negativa nel caso di identificazione del rischio che dietro al fruscio nell'erba ci sia un

---

[182] Ibid.
[183] Lisdorf sostiene che la comprensione del fenomeno religioso non possa limitarsi al solo rilevamento di una presenza (HADD), ma richieda l'attribuzione di un'intenzionalità mediata dalla PCC. Questa transizione è evidente nei quattro benchmark: se in B1 e B2 (cerchi nel grano) l'agency è inferita da stimoli fisici o geometrie complesse, in B3 (preghiera) e B4 (lettura della Genesi) la PCC opera in totale assenza di stimoli esterni, simulando un'interazione sociale con entità invisibili o interpretando l'ordine naturale come l'esito di un progetto intenzionale.
[184] Ibid.
[185] R. L. Buckner et al., *The Brain's Default Network Anatomy Function and Relevance to Disease*, Annals of the New York Academy of Sciences, 2008.
[186] J. J. Gibson, *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979.

predatore. L'ominide riconosce istantaneamente il pericolo e lo fa non tramite un calcolo analitico, ma tramite il dispositivo intenzionale. Al contrario, se un individuo ci porge del cibo, il gesto costituisce un'*affordance* positiva ovvero un'opportunità di nutrimento[187]. Quindi per capirla e sfruttarla al meglio, l'azione deve essere letta come intenzionale ed essere percepita immediatamente come un'opportunità di aiuto o di cooperazione sociale, invece che come un semplice movimento meccanico[188]. Detto ciò, non sempre la strategia cognitiva ha successo poiché il nostro dispositivo di rilevamento intenzionale può fallire, dal momento che le intenzioni possono essere fraintese o non essere sempre come appaiono[189].

Riconoscere queste possibilità di azione nell'ambiente circostante, distinguendo rapidamente le minacce dalle opportunità, è il modo più istintivo e veloce che l'essere umano possiede per sopravvivere e convivere. Al giorno d'oggi, quando usiamo l'IA, succede la stessa cosa poiché non analizziamo la statistica sottostante ai modelli, ma vediamo nella loro risposta un'*affordance* positiva, per esempio l'opportunità di ricevere aiuto, informazioni o una soluzione a un problema. In questo contesto si spiega il fatto che, trattare la macchina come se avesse intenzionalità, è un'azione evolutiva estremamente efficace poiché è il modo più rapido che abbiamo per sfruttare l'opportunità che ci si presenta davanti. Riconoscere l'IA come un insieme di calcoli probabilistici richiederebbe uno sforzo analitico poco utile nel contesto dell'interazione[190]; al contrario, considerare l'IA come un agente intenzionato ad aiutare, rende l'interazione immediata, fluida e soprattutto funzionale. Quando però questo meccanismo evolutivo si innesta su sistemi linguistici sempre più sofisticati, la funzionalità dell'attribuzione rischia di trasformarsi in una trappola cognitiva, con conseguenze che vediamo soprattutto oggi ma che la letteratura documenta ormai dagli anni Sessanta.

---

[187] M. Shanahan, *Talking About Large Language Models*; p. 59.
[188] Z. Rucinska, *Affordances in Context*, Phenomenology and the Cognitive Sciences, 2021. Rucinska evidenzia un'incongruenza ontologica: Dennett oscilla tra l'idea che le *affordance* siano proprietà oggettive raccolte dal mondo e l'idea che siano costruzioni prodotte dal cervello tramite il *predictive coding*. Se l'*affordance* è simultaneamente esterna ed interna, diventa difficile usarla come parametro preciso per spiegare scientificamente come la mente decodifichi la realtà.
[189] Esempio di una persona che porge del cibo a un'altra senza necessariamente avere l'intenzione di aiutare.
[190] S. T. Fiske, S. E. Taylor, *Social Cognition*, Reading, Addison-Wesley, 1984.

2.1.7 L'illusione della mente nella macchina: origini e conseguenze dell'effetto ELIZA

La proiezione di intenzionalità su una macchina non è un fenomeno apparso con i recenti modelli di intelligenza artificiale. Infatti già nel 1966, Joseph Weizenbaum[191] osservò come i soggetti che interagivano con ELIZA, un *software* che simulava un terapeuta attraverso la semplice ripetizione speculare delle frasi dell'utente, tendessero a convincersi che il programma li capisse realmente. Questo fenomeno, noto come effetto ELIZA, ha dimostrato che la mente umana è intrinsecamente predisposta all'antropomorfismo e che una minima coerenza formale del linguaggio sia sufficiente per convincere gli utenti che un *software* sia dotato di intenzionalità. Inoltre, come osserva l'autore, le decisioni finiscono così per essere prese in risposta alla macchina stessa, rendendo visibile la dinamica sequenziale che lega l'antropomorfismo alla fiducia[192]. Oggi l'effetto ELIZA è amplificato dall'uso di termini antropomorfici come sapere, volere o capire ma, come avverte Murray Shanahan[193], questa tendenza non fa altro che aggravare il corto circuito percettivo.

Come abbiamo già visto, dal momento che l'IA non possiede né comprensione né volontà, comportandosi come un pappagallo che imita il linguaggio, è necessario smettere di utilizzare verbi antropomorfici come "pensare", "sapere" o "volere" quando ci riferiamo ad essa, esattamente come non diremmo mai che un pappagallo "pensa", "sa" o "vuole" una cosa. Continuare a utilizzare questo registro sbagliato non fa altro che alimentare il pericoloso corto circuito che porta l'utente a credere che la macchina sia capace di intendere e di volere. Il rischio concreto nell'utilizzare un lessico antropomorfo è che questo porti non solo alla mancanza di consapevolezza riguardo al reale funzionamento dei modelli linguistici ma soprattutto, nei casi più estremi, il rischio è che possa spingere le persone a delegare decisioni cruciali a un agente privo di comprensione del mondo, ritenendo sempre affidabile l'*output* anche quando non lo è.

[191] M. Shanahan, *Talking About Large Language Models*; p. 59.
[192] J. Weizenbaum, *ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine*, Communications of the ACM, 1966.
[193] M. Shanahan, *Talking About Large Language Models*; p. 59.

In questo caso, l'antropomorfismo e la percezione di competenza e autorità portano alla tendenza a fidarsi acriticamente dei sistemi automatizzati[194].

In letteratura abbiamo numerosi casi che mostrano come questo eccessivo affidamento possa avere conseguenze critiche in diversi ambiti, soprattutto in quello medico: studi sperimentali sui sistemi di supporto alle decisioni cliniche mostrano che, in presenza di consigli automatizzati ritenuti affidabili, i medici tendono a modificare decisioni inizialmente corrette per aderire a suggerimenti errati, un fenomeno noto come *automation bias*[195].

La fiducia si crea a partire dall'antropomorfizzazione e, paradossalmente, meno si conosce la natura algoritmica dell'IA e il suo funzionamento imitativo, più si è inclini a crederla dotata di una superintelligenza e a ritenerla sempre affidabile, senza alcuno spirito critico[196]. Questo accade poiché l'assenza di *grounding* rimane invisibile all'utente inesperto e, solo riconoscendo adeguatamente che la mente che vediamo nella macchina è un riflesso della nostra architettura cognitiva, possiamo evitare di dare fiducia acritica e delegare decisioni ad un'intelligenza artificiale priva di reale comprensione ma soprattutto priva di consapevolezza etica del mondo. Anche in questo caso esistono numerose evidenze provenienti dalla letteratura che sembrano proprio confermare che l'opacità tecnica dei sistemi algoritmici possa favorire una fiducia eccessiva da parte degli utenti[197]. In generale sia l'opacità di questi sistemi che la loro fluidità linguistica tendono ad aumentare la percezione di competenza e di comprensione in modo automatico, soprattutto in assenza di una consapevolezza critica dei loro meccanismi statistici[198].

---

[194] A. J. Golds et al., *Overreliance on AI: A Review and Agenda for Research on Automation Bias*, International Journal of Human-Computer Studies, 2023.
[195] K. Goddard et al., *Automation bias a systematic review of frequency effect mediators and mitigators*, Journal of the American Medical Informatics Association, 2012.
[196] J. Kruger, D. Dunning, *Unskilled and Unaware of It How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*, Journal of Personality and Social Psychology, 1999. Cfr. Y. Hanoch et al., *The Dunning-Kruger Effect and Large Language Models A Cautionary Tale*, Computers in Human Behavior Reports, 2024. L'effetto Dunning-Kruger, applicato all'interazione con l'IA, aiuta a comprendere la distorsione cognitiva per cui gli utenti, con scarse competenze sulla natura statistica dei modelli, tendono a sopravvalutare l'intelligenza della macchina e a non riconoscerne le allucinazioni.
[197] L. A. Suchman, *Human-Machine Reconfigurations Plans and Situated Actions*, Cambridge, Cambridge University Press, 2007.
[198] E. J. De Visser et al., *Almost human Anthropomorphism increases trust resilience in cognitive agents*, Journal of Experimental Psychology Applied, 2016. L'attribuzione di tratti umani ai sistemi automatizzati

Nel paragrafo che segue proveremo a comprendere quali siano gli effetti del riporre eccessiva fiducia acritica nei sistemi di intelligenza artificiale.

2.1.8 La delega cognitiva e il rischio di deskilling: il rischio di un debito cognitivo a lungo termine

L'essere umano è portato per natura ad antropomorfizzare i sistemi di intelligenza artificiale e a trattarli come se fossero interlocutori umani dotati di competenza funzionale e intenzionalità. Tale attribuzione aumenta la fiducia acritica nei sistemi di IA e rende più probabile l'affidamento ai suoi *output*, ciò dipende dalla nostra architettura cognitiva. La letteratura recente suggerisce un ulteriore passaggio sequenziale: all'antropomorfizzazione segue un aumento della fiducia, alla fiducia segue una maggiore propensione alla delega e alla delega una diminuzione dell'impegno cognitivo. Nel breve termine, questo sembrerebbe tradursi in una diminuzione drastica dell'esercizio attivo di capacità cognitive e di competenze tecniche. Tuttavia, nel lungo periodo, la reiterazione dello schema fa pensare che possa contribuire a una progressiva esternalizzazione delle funzioni cognitive ai sistemi di IA e che possa portare potenzialmente a un processo di *deskilling* strutturato sempre più difficile da invertire[199].

Per comprendere questo passaggio mi è utile la metafora del film *Idiocracy*, in cui l'attore Luke Wilson interpreta Joe Bowers, una persona normale. Joe non è uno scienziato né un genio e, proprio per la sua normalità, viene scelto per un esperimento di ibernazione. Tuttavia a causa di un malfunzionamento della capsula, si risveglia cinquecento anni nel futuro. Ciò che scopre è che nel futuro la società sta collassando poiché gli esseri umani hanno perso la capacità di pensare in modo critico e di ragionare. Le persone sono diventate stupide e Joe diventa l'uomo più intelligente del pianeta non perché sia brillante, ma semplicemente perché il livello di competenze cognitive si è drasticamente abbassato. Nel futuro le persone non sanno più leggere, non comprendono testi e non sono in grado di risolvere problemi elementari; di conseguenza

---

aumenta significativamente il livello di fiducia e la propensione alla delega decisionale, mitigando la perdita di fiducia anche in presenza di errori del sistema.

[199] N. Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task*, arXiv, 2025.

il linguaggio si è impoverito, le decisioni politiche sono prese sulla base di slogan e attività fondamentali per il progresso della società falliscono. L'esempio emblematico è quello dell'agricoltura: non ci sono più raccolti e nessuno riesce a risolvere il problema perché si è persa la capacità di cogliere semplici relazioni di causa ed effetto[200]. Nel film le colture muoiono perché vengono irrigate con una bevanda energetica invece che con acqua e questa pratica è diventata talmente normale che nessuno si rende conto che è proprio quella la causa del problema.

In questo scenario la tecnologia viene usata senza alcuna riflessione critica. La società non è diventata stupida per mancanza di strumenti, ma perché ha smesso di usarli in modo consapevole, delegando loro progressivamente il pensiero fino a non riconoscerne più la necessità. Nessuno sceglie consapevolmente di rinunciare al pensiero; semplicemente, nel corso del tempo, il pensare è diventato sempre meno necessario poiché le soluzioni semplici, immediate e automatiche hanno sostituito gradualmente lo sforzo cognitivo fino a renderlo raro e addirittura anomalo. Il protagonista non salva il mondo con la sua intelligenza eccezionale ma con il semplice ragionamento.

Questa riflessione costituisce lo sfondo ideale per interpretare i risultati di uno studio condotto dal MIT riguardo gli effetti dell'utilizzo dei sistemi di intelligenza artificiale. Lo studio mostra infatti che l'uso intensivo dell'intelligenza artificiale generativa nella scrittura riduce lo sforzo mentale e l'attivazione cognitiva, facilitando il compito ma indebolendo nel tempo memoria, capacità critica e ragionamento. Così come nella società di *Idiocracy*, il rischio non risiede nell'esistenza della tecnologia ma nel suo utilizzo acritico. Il parallelismo suggerisce quindi una domanda: se l'IA rende il pensare meno necessario nel breve periodo, quali effetti produce sull'autonomia cognitiva nel lungo periodo? Lo studio del MIT, chiamato *'Your brain on Chatgpt: accumulation of cognitive debt when using an AI assistant for essay writing task'* mostra gli effetti dell'intelligenza artificiale generativa sulle capacità cognitive, confrontando tre modalità di scrittura: la scrittura attraverso modelli di intelligenza artificiale generativa, la scrittura attraverso l'uso di motori di ricerca tradizionali e la scrittura che i ricercatori chiamano *brain only*, senza alcuno strumento a supporto. Gli studiosi hanno monitorato

[200] Questo spunto è particolarmente rilevante, poiché ciò che distingue l'essere umano dalla macchina è proprio la capacità di comprendere le relazioni di causa ed effetto.

l'attività cerebrale dei soggetti presi in esame durante la scrittura di testi tramite elettroencefalografia e dall'analisi è emerso che l'affidamento all'IA comporta una drastica riduzione della connettività neurale. Inoltre emerge che il gruppo di soggetti *brain only,* quello che non ha utilizzato strumenti di IA, ha mostrato reti neurali più forti[201].

L'utilizzo dell'IA ha compromesso le capacità cognitive di chi la utilizzava ed in particolare è stato preoccupante l'indebolimento della capacità mnemonica: è emerso che circa l'80% dei partecipanti non è stato in grado di ricordare il proprio saggio appena dopo averlo scritto. In aggiunta è emerso che l'uso dell'IA ha portato a una pigrizia metacognitiva nel momento in cui i partecipanti hanno iniziato a delegare responsabilità e decisioni al modello, senza alcuno spirito critico. Infine i saggi prodotti con l'IA sono risultati senza alcuna creatività poiché erano sostanzialmente omogenei e privi di intuizioni personali, dal momento che riflettevano i *bias* presenti nei dati di addestramento dei modelli. Nei risultati i ricercatori hanno evidenziato il rischio di accumulare un debito cognitivo, ovvero un risparmio immediato di risorse mentali che si traduce però, nel lungo periodo, nel *deskilling*, cioè in un riduzione delle competenze e della capacità critica. I dati mostrano infatti che i partecipanti abituati all'IA faticavano a recuperare una piena connettività neurale anche quando tornavano a scrivere senza supporto. Lo studio conclude quindi che l'efficienza dell'IA deve essere bilanciata da un uso consapevole per evitare che la delega comprometta l'autonomia e che porti ad una progressiva perdita di capacità cognitive, proprio come accade in *Idiocracy*[202].

Dovremmo riflettere sugli effetti cognitivi delle tecnologie contemporanee e in particolare sui sistemi di intelligenza artificiale poiché il rischio è quello che, nel lungo periodo, l'uso acritico della tecnologia possa portare a una perdita di autonomia e un indebolimento delle capacità cognitive. In questa prospettiva, *Idiocracy* non va interpretato come una profezia, ma come un avvertimento: se la delega ai sistemi di intelligenza artificiale diventa sistematica, il pensiero tende ad atrofizzarsi per mancanza di necessità ed esercizio. L'aspetto più critico messo in luce dalla metafora è

[201] Ibid.
[202] Ibid.

che questo processo avviene senza essere percepito come problematico: la progressiva perdita di capacità cognitive non viene percepita e non genera una necessità di cambiamento, poiché la riduzione dello sforzo viene sentita, in primis, come un miglioramento dell'efficienza. In assenza di consapevolezza tuttavia il circolo vizioso non fa altro che autoalimentarsi, poiché meno si pensa meno sembra necessario pensare.

2.1.9 La sequenza "attribuzione, fiducia, delega": una sintesi teorica

Il nesso tra attribuzione di intenzionalità e delega acritica non è formulato in modo esplicito e sistematico dagli autori considerati, ma emerge dalle loro riflessioni: Weizenbaum ha mostrato quanto sia facile generare un'illusione di comprensione e ha osservato che le decisioni tendono progressivamente a essere prese in risposta all'*output* della macchina percepita come competente[203]; Barrett ha spiegato perché l'essere umano sia evolutivamente predisposto a rilevare *agency*[204]; Dennett ha fornito la cornice teorica della strategia per predire il comportamento degli oggetti trattandoli come agenti dotati d'intenzionalità[205].

Inoltre, come si è visto, la riflessione di Giddens[206] si integra perfettamente: gli individui ripongono fiducia in "sistemi esperti" che richiedono competenze tecniche, nonostante non le posseggano e non siano in grado di verificarle. Tale fiducia è una risposta adattiva alla complessità crescente dei sistemi nella modernità. Tuttavia, proprio per questo motivo, la fiducia in questi sistemi tende a essere riposta acriticamente sulla base della percezione di competenza del sistema invece che sulla base della comprensione effettiva. Nel caso dell'intelligenza artificiale, il meccanismo evidenziato da Giddens si combina con la predisposizione evolutiva e con la posizione intenzionale, producendo la dinamica secondo la quale un sistema viene percepito esperto e dotato di intenzioni; infine, da questa percezione, si genera una fiducia tale da autorizzare la delega.

---

[203] J. Weizenbaum, *ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine*, p. 68.
[204] J. L. Barret, *Why Would Anyone Believe in God?*, p. 63.
[205] A. Lisdorf, *What's HIDD'n in the HADD?*, p. 63.
[206] A. Giddens, *Le conseguenze della modernità*, p. 63.

In questa parte del lavoro considero questi elementi congiuntamente in una prospettiva multidisciplinare e diacronica, con l'intento di ricostruire la dinamica sequenziale attribuzione, fiducia, delega. Dati questi presupposti, è plausibile ipotizzare coerentemente che l'attribuzione di *agency* possa favorire la fiducia, di conseguenza che la fiducia possa facilitare la delega, e infine che la delega possa ridurre l'esercizio diretto delle competenze coinvolte nel processo decisionale. In questo modo, il *deskilling* cognitivo si configura come un possibile, ma non unico e necessario, esito sistemico di una dinamica cognitiva che si consolida attraverso l'interazione.

2.2 La costruzione sociale dell'IA: dalle proiezioni individuali alle dinamiche di gruppo

Fin qui abbiamo considerato fenomeni che si manifestano prevalentemente sul piano individuale quali modalità di attribuzione, scorciatoie cognitive e pratiche di delega. Ma cosa accade quando questi meccanismi diventano diffusi e normalizzati? In che modo contribuiscono a ridefinire aspettative collettive, pratiche organizzative e modelli di responsabilità?

L'attribuzione all'IA di caratteristiche antropomorfe non produce soltanto effetti sul comportamento del singolo individuo, ma si estende anche al livello sociale e organizzativo. I meccanismi cognitivi e psicologici analizzati nella sezione precedente si osservano, infatti, all'interno di un contesto più ampio: l'essere umano non interagisce con la tecnologia in modo isolato, ma in reti sociali e ambienti organizzativi professionali. Comprendere pienamente l'impatto dell'intelligenza artificiale richiede di spostare l'attenzione verso la dimensione sociale e collettiva, in cui tali percezioni si consolidano e producono effetti strutturali.

Uno dei motori principali di questa transizione dal piano individuale a quello sociale è la percezione e in particolare oggi vi è la percezione diffusa che l'IA aumenti l'efficienza lavorativa. Tuttavia prima di analizzare la dimensione collettiva è opportuno interrogarsi anche su quanto l'efficienza attribuita all'IA sia effettiva e non soltanto percepita. A tal proposito, una ricerca preliminare ancora in corso condotta

dall'università della California[207], suggerisce in merito che l'adozione di strumenti di AI generativa non abbia davvero comportato una riduzione netta del lavoro umano. Al contrario, i lavoratori coinvolti hanno progressivamente ampliato la portata dei propri compiti, hanno mantenuto ritmi più sostenuti e prolungato l'orario di lavoro. I ricercatori[208] parlano di *workload creep,* ossia di un aumento graduale e spesso impercettibile del carico complessivo[209].

Questo dato mette in discussione l'assunto secondo cui l'IA renda automaticamente il lavoro più efficiente in termini di tempo e di sforzo cognitivo: se utilizzando gli strumenti di IA il carico di lavoro aumenta e il tempo dedicato alle attività si estende, l'efficienza promessa si rivela non necessariamente reale. Da un lato potrà sicuramente accelerare l'esecuzione dei singoli compiti, dall'altro, però, la riduzione dello sforzo immediato può tradursi in un incremento della pressione complessiva, rendendo più difficile valutare se si tratti davvero di un guadagno in efficienza in termini di tempo e di energie cognitive. In questo senso, l'efficienza appare come una delle categorie interpretative che contribuiscono alla costruzione sociale dell'IA.

2.2.1 Infosfera e actor network theory: verso una concezione relazionale dell'*agency*

La costruzione della realtà avviene all'interno dell'infosfera teorizzata da Luciano Floridi, un ambiente informazionale in cui "ciò che è reale è informazionale e ciò che è informazionale è reale"[210]. In questo spazio, la distinzione tra agenti umani e artificiali è sempre più difficile e si fa progressivamente più sfumata nell'esperienza *onlife*, dove il confine tra il mondo analogico-*offline* e il mondo digitale-*online* viene meno. Nell'infosfera l'informazione non è un semplice mezzo di comunicazione, ma è l'ambiente stesso in cui l'azione umana prende forma: essere, agire e conoscere avvengono all'interno di un collettivo, superando la vecchia nozione di società, in cui coesistono e interagiscono *inforgs*, ovvero gli organismi informazionali umani, i sistemi

[207] A. Ranganathan et al., *AI Doesn't Reduce Work It Intensifies It*, in Harvard Business Review, 2026.
[208] Ibid.
[209] Lo studio condotto presso la UC Berkeley Haas School of Business ha coniato il termine *workload creep* per descrivere l'aumento graduale e spesso invisibile del carico di lavoro e della pressione sui dipendenti in seguito all'adozione dell'IA generativa.
[210] L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, p. 49.

artificiali, i dati e gli algoritmi. In questo ecosistema, l'intelligenza artificiale non deve essere interpretata come un mero strumento di supporto operativo ma come un mediatore attivo[211].

Come abbiamo visto, la nozione di infosfera implica un cambiamento profondo nel modo in cui viene concepita l'*agency* e, se in una visione tradizionale l'azione era prerogativa quasi esclusiva del soggetto umano, nell'infosfera l'*agency* si distribuisce lungo una molteplicità di entità informative che partecipano, in modi diversi, alla produzione di informazioni e agli esiti sociali.

Ed è proprio in questo scenario che la *actor network theory* (ANT) fornisce una cornice analitica particolarmente utile. La sociologia delle associazioni è un metodo per tracciare le connessioni tra elementi eterogenei all'interno della società che è stato sviluppato dal sociologo Bruno Latour. Latour[212] critica la sociologia tradizionale che tratta il sociale come se fosse una sostanza specifica, una sorta di colla invisibile che spiega perché le cose stanno insieme. Al contrario, egli sostiene che ciò che chiamiamo società non è la causa che tiene uniti gli individui, ma il risultato finale di una fitta rete di collegamenti tra attori eterogenei. Per Latour[213], il compito del sociologo non è quindi invocare una misteriosa forza sociale per spiegare i fenomeni, ma tracciare i fili delle reti.

Al giorno d'oggi, l'ANT permette di superare la tradizionale separazione tra soggetti e oggetti per analizzare la realtà sociale come il risultato di reti eterogenee composte da attori umani e non umani, definiti *actants*[214]. In questa prospettiva, un attante è qualsiasi entità dotata di *agency* capace di produrre effetti all'interno di una rete, indipendentemente dal fatto che possieda intenzionalità, volontà o coscienza, poiché ciò che conta non è la natura dell'attore, ma esclusivamente il suo ruolo nella rete.

La rete dell'ANT teorizzata da Latour con all'interno gli attanti, non è altro che l'infosfera di Floridi che può essere visualizzata proprio come una rete con all'interno gli agenti informazionali. L'infosfera descritta da Floridi è il contesto in cui queste reti

[211] Ibid.
[212] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, 2005.
[213] Ibid.
[214] Ibid.

prendono forma, all'interno del quale l'IA non è più soltanto uno strumento utilizzato dall'uomo, ma una componente stessa dell'ambiente informazionale che produce, modifica e condiziona l'azione umana. L'ANT e il concetto di infosfera convergono nel mettere in discussione una visione antropocentrica dell'*agency*, sostituendola con una prospettiva relazionale e distribuita tra attori di diversa natura[215].

In questo lavoro, applicare *l'actor network theory* all'intelligenza artificiale significa rifiutare l'idea che sia un semplice strumento passivo, per riconoscerle invece lo *status* di attante e partecipante attivo nella rete sociale, capace di creare e modificare la realtà e di definire ruoli e responsabilità. Questa redistribuzione dell'*agency* tuttavia solleva una questione cruciale: dal momento che l'azione è distribuita tra molteplici attanti, anche la responsabilità tende a distribuirsi. Dunque quando una decisione supportata da un sistema di IA produce un errore o un danno, di chi è la responsabilità all'interno della rete tra il programmatore, il decisore umano, il *dataset* o l'organizzazione che ha adottato il sistema?

Prendiamo l'esempio di un *report*. Il documento finale non può essere attribuito esclusivamente all'umano che lo consegna, né all'algoritmo che lo ha prodotto poiché è il risultato di una rete composta da dati in *input*, modelli statistici e parametri di addestramento, scelte progettuali, competenze professionali e aspettative del cliente. Quindi la responsabilità della creazione e della modifica del contenuto si distribuisce lungo questa rete. Se ci riflettiamo, l'azione umana non è mai stata il prodotto di un singolo soggetto, ma l'esito di una rete sociotecnica in cui oggi l'IA inizia ad occupare una posizione sempre più centrale. Questa lettura consente di comprendere come la tecnologia contribuisca non solo all'esecuzione delle attività, ma anche alla ridefinizione delle relazioni di potere e di responsabilità all'interno delle organizzazioni. Tuttavia questa impostazione non implica una deresponsabilizzazione dell'essere umano, ma una ricollocazione riguardo alla propria *agency* all'interno di reti sempre più ampie, oltre alla consapevolezza del proprio ruolo nella rete in quanto agente semantico. Solo attraverso tale consapevolezza è possibile analizzare in modo adeguato fenomeni come la delega decisionale e la fiducia acritica nei sistemi automatizzati. Questo punto è cruciale perché permette di analizzare gli effetti delle definizioni sociali dell'IA: se l'IA

[215] Ibid.

agisce all'interno delle reti non solo attraverso ciò che fa, ma anche attraverso ciò che si crede faccia (vista la difficoltà oggettiva di riconoscere l'*agency* umana da quella artificiale), allora il modo in cui viene definita diventa un elemento centrale dell'azione sociale. È in questo passaggio che il discorso sull'infosfera e sull'ANT prepara il terreno per le teorie successive.

Comprendere come l'IA venga definita e riconosciuta intelligente, affidabile o autorevole, diventa essenziale per spiegare perché produca determinati effetti sociali ma anche per comprendere come le percezioni, le definizioni e le credenze sull'IA possano trasformarsi in conseguenze sociali concrete, talvolta negative. È precisamente nel passaggio dal piano relazionale a quello interpretativo che diventa centrale il riferimento al teorema di Thomas e alla profezia che si autoavvera, elaborata da Robert Merton.

2.2.2 Il teorema di Thomas e profezia che si autoavvera: il rischio quando le definizioni diventano realtà

L'analisi dell'infosfera e dell'ANT ha mostrato come l'intelligenza artificiale faccia parte della rete sociotecnica in quanto attante, contribuendo a redistribuire l'*agency*, i ruoli e le responsabilità e riorganizzando le relazioni tra attori umani e non umani. Tuttavia riconoscere l'IA come elemento attivo all'interno delle reti non è sufficiente per comprendere perché essa produca determinati effetti sociali. È necessario interrogarsi anche sul modo in cui l'IA viene definita, interpretata e costruita simbolicamente dagli attori che ne fanno uso, poiché la sola analisi della dimensione individuale, analizzata nel capitolo precedente, non è sufficiente a spiegare la produzione dei suoi effetti sociali. Quindi se l'ANT consente di descrivere come l'IA agisce all'interno delle reti, resta da chiarire in che modo le credenze, le aspettative e le rappresentazioni condivise contribuiscano a rendere tale azione effettiva.

Il punto di partenza per comprendere la costruzione sociale dell'intelligenza artificiale può essere individuato nel teorema di Thomas, secondo cui "se gli uomini definiscono certe situazioni come reali, esse sono reali nelle loro conseguenze" [216]. Con questa

[216] W. I. Thomas, D. S. Thomas, *The Child in America Behavior Problems and Programs*, New York, Alfred A. Knopf, 1928. Il teorema ha ottenuto una vastissima risonanza grazie a Merton, sebbene Thomas lo abbia espresso in una serie di lavori specialistici.

affermazione, Thomas sposta l'attenzione dalla realtà oggettiva dei fatti al ruolo centrale delle definizioni sociali nel guidare l'azione. Ciò che orienta il comportamento umano non è tanto ciò che è vero in senso fattuale, quanto ciò che viene interpretato, condiviso e trattato come vero dagli attori sociali. Tutto ciò che viene percepito come reale produce effetti concreti nella realtà sociale e genera conseguenze tangibili nei comportamenti dei suoi attori sociali. Pare chiaro che non esista una realtà di tipo oggettiva, ma che esista piuttosto una realtà sociale costruita e guidata dalle percezioni collettive. Questo processo è prettamente umano e trova riscontro nella realtà sociale ma non nel mondo della natura: mentre le previsioni dell'orbita di una cometa non ne influenzano il percorso, la definizione pubblica di una situazione sociale non solo la genera ma diventa parte integrante della situazione stessa, influenzandone gli sviluppi futuri[217].

Il teorema di Thomas rappresenta così il punto di partenza fondamentale per comprendere come la realtà sociale si formi non solo a partire dai fatti oggettivi, ma soprattutto dalle interpretazioni che gli individui danno di tali fatti. In questa prospettiva perciò anche una rappresentazione falsa può produrre effetti concreti e indurre comportamenti reali, purché venga assunta come valida dagli attori sociali: se un quartiere viene percepito come pericoloso, anche in assenza di un reale aumento della criminalità, le persone tenderanno a evitarlo, le attività economiche diminuiranno e il degrado effettivamente aumenterà[218].

Accade ciò poiché la definizione iniziale, pur non fondata sui fatti, orienta i comportamenti in modo tale da produrre conseguenze reali. Riguardo questo meccanismo Robert Merton[219] ribadisce che la realtà sia la conseguenza tangibile di ciò che le persone credono sia reale, mostrando come le definizioni sociali, le percezioni e i giudizi orientino le scelte e i comportamenti e possano persino produrre effetti reali tali da trasformare credenze inizialmente false in vere. Merton introduce così il concetto di profezia che si autoavvera, "una previsione falsa che, per il solo fatto di essere espressa,

[217] R. K. Merton, *The Self-Fulfilling Prophecy*, The Antioch Review, 1948. Merton sviluppa il teorema di Thomas introducendo il concetto di profezia che si autoavvera, spiegando come una definizione inizialmente falsa di una situazione possa indurre comportamenti che rendono quella stessa definizione vera nei suoi effetti concreti.
[218] Ibid.
[219] Ibid.

genera comportamenti che la realizzano"[220]; spiegandolo attraverso l'esempio della banca ritenuta insolvente dai correntisti: credendo che la banca stia per fallire, tutti ritirano i propri risparmi ed è proprio questa reazione collettiva che provoca il suo fallimento. Il fatto che la banca sia sull'orlo del fallimento non è vera ma la percezione che lo sia, seppur infondata, orienta i comportamenti a tal punto da realizzarsi davvero. In questo modo, Merton rende ancor più esplicito il carattere performativo delle credenze sociali[221].

Questo schema interpretativo risulta particolarmente utile per analizzare i fenomeni contemporanei legati all'intelligenza artificiale. Da un lato, le percezioni individuali e le rappresentazioni che gli utenti costruiscono attorno all'IA influenzano i loro comportamenti fino a produrre effetti concreti: se un sistema viene percepito come competente, oggettivo o autonomo, gli individui tenderanno a fidarsi e ad affidargli decisioni e responsabilità. In questo senso, la definizione produce la realtà. Dall'altro lato, i sistemi di IA, addestrati su dati storici, incorporano nei propri modelli *bias*, asimmetrie e disuguaglianze preesistenti, restituendoli sotto forma di previsioni apparentemente neutre e oggettive. Tali rappresentazioni orientano le decisioni umane, generando comportamenti che finiscono per confermare le stesse tendenze da cui il modello era partito. In entrambi i casi la profezia si autoavvera.

Tuttavia emerge un rischio strutturale: se la realtà sociale si costruisce sia a partire dalle percezioni individuali che dalle definizioni e rappresentazioni collettive, cosa accade quando si attribuiscono impropriamente ai sistemi di IA caratteristiche antropomorfiche, quali intelligenza, autorità o autonomia, e una fiducia sproporzionata rispetto ai loro limiti tecnici?

Per comprendere come questo rischio si concretizzi, è necessario partire da una premessa tecnica: come osserva Burrell[222] i modelli predittivi, plasmano la realtà

[220] Ibid.
[221] Merton riconosce che l'intuizione alla base della profezia che si autoavvera ha importanti precursori. Tra questi cita il vescovo Bossuet, De Mandeville, Marx, Freud e Sumner, che in modi differenti avevano già colto il carattere performativo delle credenze sociali. Un parallelismo significativo si trova anche in Mead, il quale osserva al contrario che "se una cosa non è riconosciuta come vera, allora non funziona come vera nella comunità", anticipando l'idea che la validità sociale di una definizione dipenda dal suo riconoscimento collettivo.
[222] J. Burrell, *How the Machine Thinks Understanding Opacity in Machine Learning Algorithms*, Big Data & Society, 2016.

identificando correlazioni statistiche nei dati di addestramento che riflettono il passato. Questo significa che l'algoritmo riproduce ciò che è già accaduto, incluse disuguaglianze, discriminazioni e *bias* strutturali[223]. Quindi, la realtà che i sistemi di IA sembrano descrivere oggettivamente incorpora *bias*, asimmetrie e distorsioni del passato.

Il punto cruciale è che queste rappresentazioni, anche se viziate o false, una volta percepite e accettate come vere, influenzano le decisioni umane e plasmano la realtà, producendo effetti tangibili[224]. Questo meccanismo è stato dimostrato empiricamente da Dressel e Farid[225], nell'analisi del sistema COMPAS utilizzato nel sistema penale statunitense per la valutazione del rischio di recidiva[226]. Il modello apprende e riproduce i pattern di arresto del passato, che riflettono disparità razziali sistemiche, restituendoli sotto forma di previsioni apparentemente neutrali e oggettive. Si è visto che la rappresentazione, seppur viziata, viene trattata come affidabile e produce effetti concreti sui decisori umani. Infatti è stato osservato che i sistemi di valutazione del rischio di recidiva hanno un impatto sulla discrezionalità dei giudici, le cui decisioni vengono influenzate dal punteggio algoritmico, con l'emissione di sentenze più severe per chi riceve punteggi alti e meno severe per chi riceve punteggi bassi. Inoltre è significativo l'effetto strutturale sul lungo periodo: nel tempo, i giudici riducono il proprio apporto valutativo autonomo, dal momento che considerano l'*output* del sistema affidabile, oggettivo e inoppugnabile[227]. La percezione dell'oggettività del sistema genera una

[223] C. O'Neil, *Weapons of Math Destruction How Big Data Increases Inequality and Threatens Democracy*, New York, Crown, 2016. L'autrice definisce questi modelli come armi di distruzione matematica, evidenziando come l'apparente oggettività dei dati nasconda in realtà la sistematica riproduzione di pregiudizi razziali e socioeconomici.
[224] D. Mackenzie, *An Engine Not a Camera How Financial Models Shape Markets*, Cambridge, MIT Press, 2006. Sebbene riferito alla finanza, il concetto di performatività di MacKenzie è cruciale per comprendere come l'IA non si limiti a "fotografare" la realtà (camera), ma agisca come un "motore" (engine) che la modifica, spingendo gli attori sociali a conformarsi alle sue previsioni.
[225] J. Dressel, H. Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, Science Advances, 2018
[226] Il sistema è ampiamente criticato per la sua natura di black-box (scatola nera), in quanto il suo funzionamento interno è protetto da segreto commerciale, rendendone difficile la verifica della trasparenza e dell'accuratezza. Un'inchiesta di ProPublica nel 2016 ha evidenziato che l'algoritmo mostrava pregiudizi razziali, tendendo più spesso a etichettare falsamente gli imputati afroamericani come ad alto rischio rispetto ai bianchi, viceversa per i falsi negativi. Il sistema è diventato famoso con il caso di Eric Loomis, nel Wisconsin, a cui è stata inflitta una pena più severa basandosi su un punteggio di alto rischio calcolato da COMPAS. La Corte Suprema del Wisconsin ha respinto il suo ricorso, permettendo l'uso dell'algoritmo ma sottolineando la necessità di prudenza.
[227] L. Doleac, *Algorithmic Risk Assessment in the Hands of Humans*, Institute of Labor Economics, Bonn 2019.

delega reale, che a sua volta produce sentenze reali con conseguenze reali sulla libertà delle persone. Una rappresentazione falsa diventa così vera nelle sue conseguenze.

Nel campo delle tecnologie predittive di devianza assistiamo a questo meccanismo in maniera ancora più accentuata, poiché il teorema di Thomas e la profezia che si autoavvera producono effetti ancor più irreversibili. I sistemi di *predictive policing*, classificando algoritmicamente individui o quartieri ed etichettandoli come ad alto rischio, orientano l'operato delle forze dell'ordine verso controlli intensificati e interventi preventivi. La sorveglianza genera nuovi fermi e arresti, che confluiscono nei *dataset* di addestramento e rafforzano la definizione iniziale. La rappresentazione distorta si autoconvalida non perché sia corretta, ma perché produce i comportamenti e le condizioni che la rendemmo vera nei suoi effetti. Quindi l'accettazione acritica degli *output* si traduce in esclusione e marginalizzazione, attraverso un meccanismo che si alimenta dalla rappresentazione distorta alla decisione, dalla decisione alla conseguenza reale, fino alla conferma della rappresentazione stessa[228].

Un caso emblematico di questo meccanismo è quello di Robert McDaniel, inserito dalla Chicago Police Department (CPD) in una lista di soggetti, nota come *heat list*, di un sistema di *predictive policing*[229]. Nel 2013 McDaniel ricevette una visita domiciliare da agenti del CPD, che lo informarono che l'algoritmo lo classificava come soggetto ad alta probabilità di essere coinvolto in un episodio di violenza armata. Gli agenti, come McDaniel avrebbe in seguito dichiarato, non sapevano nemmeno se sarebbe stato vittima o aggressore. Quello che successe è che la profezia si autoavverò in modo tragico: la sorveglianza intensificata attorno a McDaniel lo rese sospetto agli occhi di chi lo circondava e fu ferito da un colpo d'arma da fuoco da parte di qualcuno che, secondo la sua stessa ricostruzione, lo riteneva un informatore della polizia[230].

Il caso rende tangibile l'intero circuito descritto: una rappresentazione distorta, elaborata a partire da dati viziati, viene accettata acriticamente come oggettiva e

[228] N. Rossbach*, Innocent Until Predicted Guilty: How Premature Predictive Policing Can Lead to a Self-Fulfilling Prophecy of Juvenile Delinquency*, 2023.
[229] M. Stroud, *Heat Listed*, The Verge, 2021. Il caso di Robert McDaniel illustra come la *Strategic Subject List* (SSL) del CPD abbia generato una profezia che si autoavvera: la sorveglianza speciale basata su un punteggio di rischio algoritmico ha isolato il soggetto dalla sua comunità, portando ad un'aggressione fisica dovuta alla percezione errata che egli fosse un informatore.
[230] Ibid.

trasformata in realtà sociale. Secondo la logica del teorema di Thomas, la definizione di rischio, pur non fondata su alcun fatto concreto, ha modificato percezioni, ha orientato comportamenti e prodotto conseguenze reali e irreversibili sulla vita di una persona.

Vi è un aspetto particolarmente preoccupante: nel caso in cui si affermasse e si diffondesse in modo sistematico una definizione erronea di una tecnologia e si creassero le condizioni favorevoli al suo consolidamento, sia sul piano dei *bias* psicologici individuali sia sul piano della narrazione sociale, potrebbe realizzarsi qualcosa di falso ma operativamente reale, secondo la logica della profezia che si autoavvera. Immaginiamo che un sistema di IA in ambito medico venga usato e accettato come accurato più del giudizio umano, anche se in realtà non lo è: tale definizione potrebbe indurre i professionisti sanitari a fidarsi in modo crescente delle sue raccomandazioni, riducendo progressivamente il controllo critico e la verifica clinica. Anche in presenza di errori, il sistema continuerebbe a essere percepito come affidabile. Questo rischio non è solo teorico, dal momento che esistono numerosi casi in letteratura che evidenziano tale criticità[231].

Il rischio si aggrava nel caso dell'IA a fronte di allucinazioni dei modelli linguistici (un modello che genera un *output* falso, una fonte inesistente, un dato errato o una conclusione non supportata dai fatti): se l'utente interiorizza la definizione e percepisce il sistema come affidabile tende a non accorgersi dell'errore. Anche in questo caso, la non correttezza dell'*output* non ne impedirebbe l'efficacia sociale. In questo caso, la profezia che si autoavvera si autoalimenta di continuo e lo fa attraverso sia il meccanismo sociale di gruppo, sia attraverso il meccanismo antropomorfico individuale, descritto nei paragrafi precedenti. Alla luce di ciò, il teorema di Thomas e la profezia che si autoavvera applicati all'intelligenza artificiale, non rappresentano un rischio teorico astratto, ma una dinamica concreta che produce effetti sociali, organizzativi ed etici misurabili. Il pericolo maggiore risiede in ciò che le persone credono che l'IA sia, non conoscendo la sua natura tecnica, e nel modo in cui queste rappresentazioni individuali e collettive plasmino, nel tempo, narrazioni, pratiche, decisioni e responsabilità.

---

[231] K. Goddard et al., *Automation bias a hidden issue for clinical decision support system use*, Studies in Health Technology and Informatics, 2011.

2.2.3 L'istituzionalizzazione dell'IA: dalle credenze sociali condivise alle pratiche organizzative

Resta da comprendere attraverso quali meccanismi tali percezioni e rappresentazioni si stabilizzino nel tempo e in che modo si sedimentino a livello organizzativo in pratiche istituzionalizzate. Tuttavia se il teorema di Thomas, "se gli uomini definiscono reali le situazioni, esse saranno reali nelle loro conseguenze"[232], e la profezia che si autoavvera di Merton, "una previsione falsa che, per il solo fatto di essere espressa, genera comportamenti che la realizzano"[233], consentono di comprendere come le definizioni e le credenze sull'intelligenza artificiale possano produrre effetti reali sulla percezione collettiva e sul comportamento individuale e di gruppo, è necessario precisare che la profezia non si realizza in qualunque condizione e non prescinde dalla realtà operativa della tecnologia. Una definizione sociale, per stabilizzarsi e produrre effetti duraturi, deve essere percepita come efficace in modo concreto. In altri termini l'intelligenza artificiale ha un impatto concreto nella realtà sociale perché è coerente con la percezione, dimostrando di offrire vantaggi reali quali ad esempio efficienza, rapidità o capacità di elaborazione. Questa efficacia percepita rende credibile la definizione sociale dell'IA, favorendone il consolidamento temporale e formale nelle organizzazioni.

Tuttavia, emerge un rischio quando l'efficacia diventa il criterio dominante di valutazione, quando la distinzione tra intelligenza artificiale e intelligenza umana tende a sfumare e quando la soglia critica con cui gli individui e le organizzazioni valutano l'*output* tecnologico si abbassa progressivamente. Il rischio è evidente non perché l'IA diventi ciò che non è, ma perché il comportamento umano si adatta a una definizione semplificata e antropomorfica della tecnologia e questa riduzione della soglia critica non produce effetti immediati o catastrofici, ma opera in modo cumulativo. Come visto, nel lungo periodo, può portare a una riduzione sistematica del controllo umano, a una crescente delega decisionale e a una normalizzazione dell'automazione come risposta naturale ai problemi complessi. È proprio questo il rischio intravisto in forma satirica

[232] W. I. Thomas, D. S. Thomas, *The Child in America Behavior Problems and Programs*, New York, p. 78.
[233] R. K. Merton, *The Self-Fulfilling Prophecy*, p. 79.

nel film *Idiocracy* e confermato empiricamente dagli studi psicologici e dalle recenti ricerche del MIT sul debito cognitivo: non la sostituzione improvvisa dell'intelligenza umana, ma la sua progressiva disattivazione attraverso l'uso non riflessivo della tecnologia.

A questo punto per comprendere come queste dinamiche si consolidino nel tempo e diventino parte integrante delle pratiche sociali e organizzative, è necessario spostare l'analisi dal livello delle percezioni individuali a quello delle strutture organizzative. Le credenze sull'IA cessano di essere mere rappresentazioni soggettive ed effimere e assumono una forma istituzionalizzata, traducendosi in criteri condivisi di appropriatezza, legittimità e razionalità condivisa. La prospettiva neoistituzionalista, sviluppata da Meyer e Rowan[234] e approfondita da DiMaggio e Powell[235], fornisce una chiave interpretativa eccellente per comprendere come l'adozione della tecnologia, e in questo caso dell'IA, non derivi esclusivamente da valutazioni tecniche di efficienza, ma da processi simbolici di legittimazione sociale.

Infatti le decisioni organizzative non sono sempre guidate esclusivamente da criteri di efficienza, razionalità strumentale o ottimizzazione delle *performance*, ma rispondono anche a esigenze di legittimazione sociale. Secondo l'approccio neoistituzionalista, l'adozione di nuove tecnologie può essere il risultato di processi di "isomorfismo istituzionale" e, come mostrato da Paul DiMaggio e Powell, questo isomorfismo può assumere forme coercitive (attraverso pressioni normative e regolatorie), forme mimetiche (attraverso imitazione dei modelli di successo) e forme normative (attraverso standard professionali), rendendo le organizzazioni simili per credibilità[236]. In tutti questi casi, l'IA viene adottata o integrata nei processi decisionali non necessariamente perché utile, ma perché rappresenta modernità, razionalità e avanzamento tecnologico.

---

[234] J. W. Meyer, B. Rowan, *Institutionalized Organizations Formal Structure as Myth and Ceremony*, American Journal of Sociology, 1977.

[235] P. J. Dimaggio, W. W. Powell, *The Iron Cage Revisited Institutional Isomorphism and Collective Rationality in Organizational Fields*, American Sociological Review, 1983. Vedi anche Tolbert & Zucker per evidenze empiriche su diffusione graduale di riforme strutturali. Cfr. P. S. Tolbert, L. G. Zucker, *Institutional Sources of Change in the Formal Structure of Organizations. The Diffused Adoption of Civil Service Reforms 1880-1935*, Administrative Science Quarterly, 1983. Lo studio fornisce evidenze empiriche sulla diffusione delle riforme, distinguendo tra adozione per efficienza (fase iniziale) e adozione per legittimità istituzionale (fase di diffusione di massa).

[236] Ibid.

Perciò al giorno d'oggi per quanto riguarda l'adozione di strumenti di intelligenza artificiale, le decisioni organizzative spesso trascendono la razionalità strumentale per rispondere a imperativi isomorfi, attraverso i quali le imprese tendono a omogeneizzarsi per preservare credibilità esterna. In particolare, l'isomorfismo istituzionale si manifesta attraverso tre meccanismi interdipendenti: dal punto di vista coercitivo, stiamo assistendo a pressioni regolatorie come l'EU AI Act del 2024 che impone standard di *compliance* per l'uso di IA ad alto rischio, costringendo le organizzazioni a integrare l'intelligenza artificiale per evitare sanzioni e accedere a fondi pubblici. Dal punto di vista mimetico, stiamo invece assistendo all'imitazione di modelli percepiti come di successo: oggi ci sono aziende che cercano sempre più frequentemente di ridurre l'incertezza in contesti volatili, adottando strumenti di IA generativa di leader *tech* quali Google, Antrophic e OpenAI. Infine, dal punto di vista normativo, stiamo assistendo alla diffusione di standard professionali, come le certificazioni PMI o ISO sull'etica dell'IA, che definiscono l'integrazione dell'IA come prassi di modernità tecnologica. In tutti i casi, l'IA funge da segnale simbolico di razionalità e innovazione, adottata non solo per utilità operativa ma per allinearsi a norme condivise che conferiscono legittimità sociale, trasformando così credenze effimere in strutture durevoli.

Negli ultimi anni l'adozione acritica di *chatbot* e sistemi automatizzati, motivata spesso dalla sola ricerca di legittimità istituzionale, ha generato fallimenti misurabili che le organizzazioni si sono trovate a dover gestire a posteriori. Il caso più emblematico è quello di Klarna, la *fintech* svedese che nel 2024 sostituì circa 700 addetti al servizio clienti con un sistema di IA sviluppato in collaborazione con OpenAI, dichiarando pubblicamente che il *chatbot* gestiva il lavoro equivalente a quello dell'intero personale sostituito. I risultati iniziali sembravano giustificare la scelta, ma nel giro di pochi mesi la soddisfazione dei clienti calò sensibilmente e iniziarono a emergere criticità operative. Il CEO Sebastian Siemiatkowski ammise poi pubblicamente che il costo era diventato il fattore troppo dominante nella valutazione, con conseguente riduzione della qualità del servizio. Entro la metà del 2025, Klarna aveva invertito rotta avviando una nuova campagna di assunzioni di personale umano[237].

[237] Il CEO di Klarna ha ammesso che l'eccessiva focalizzazione sul risparmio dei costi tramite l'automazione ha compromesso la qualità del servizio, portando alla decisione di reintegrare personale umano per gestire le interazioni più complesse e sfumate.

Il caso Klarna non è isolato: secondo un'indagine IBM[238] su duemila CEO, solo uno su quattro dei progetti di IA realizzati genera il ritorno sull'investimento promesso e appena il 16% viene scalato a livello aziendale. Nonostante questi risultati, quasi due terzi dei CEO (64%) dichiarano di investire in tecnologie AI prima ancora di averne compreso il valore concreto[239].

Un altro caso significativo è quello di McDonald's, che tra il 2021 e il 2024 sperimentò in oltre cento ristoranti negli Stati Uniti un sistema di *voice ordering* automatizzato: il sistema aveva difficoltà nel riconoscere accenti e dialetti diversi, con un impatto sull'accuratezza degli ordini. I video virali degli errori del sistema, tra cui ordini moltiplicati, abbinamenti di cibo assurdi, comandi ignorati, trasformarono l'esperimento in un caso di discussione pubblica e, dopo tre anni di test, nel giugno 2024 McDonald's annunciò la conclusione del progetto pilota.

Questi casi illustrano un paradosso poiché le stesse pressioni isomorfe, che spingono le organizzazioni ad adottare l'IA come segnale di modernità e razionalità, le espongono a fallimenti tecnici, danni reputazionali e responsabilità giuridiche che ne minano la legittimità. L'adozione acritica produce esattamente l'esito opposto a quello cercato.

Pare chiaro quindi che il meccanismo dell'istituzionalizzazione si colleghi direttamente al teorema di Thomas: se l'IA è definita socialmente come necessaria, inevitabile o strategica, a prescindere dal suo reale utilizzo nei diversi contesti in cui viene inserita, tali definizioni diventano reali nelle loro conseguenze organizzative. Le imprese finiscono per adottare sistemi di intelligenza artificiale non solo per ciò che fanno, ma per ciò che rappresentano e l'IA rischia così di diventare un mero dispositivo retorico. Il neoistituzionalismo consente di comprendere come l'*hype* tecnologico, alimentato da media, consulenza strategica e *policy*, si traduca spesso in scelte organizzative concrete azzardate. In tal modo, adottare l'intelligenza artificiale senza una piena consapevolezza dei suoi limiti tecnici e delle sue implicazioni diventa rischioso.

---

[238] IBM INSTITUTE FOR BUSINESS VALUE, *CEO Study 5 mindshifts to supercharge business growth*, Armonk, IBM Corporation, 2025. Lo studio, condotto su 2.000 CEO globali, evidenzia come il 64% degli investimenti in IA sia guidato dalla paura di restare esclusi (FOMO) invece che da una reale comprensione del valore aziendale, con tassi di scalabilità che si fermano al 16% dei progetti.

[239] Se il 64% dei CEO investe senza comprendere il valore, l'adozione dell'IA non è una scelta tecnica, ma un'azione rituale per soddisfare le aspettative del mercato (il "mito e cerimonia" di Meyer e Rowan).

CAPITOLO 3

3.1 Dal chatbot all'agente: l'evoluzione degli strumenti di intelligenza artificiale

In pochi anni l'intelligenza artificiale è passata da tecnologia emergente a tecnologia strategica globale: secondo il rapporto di McKinsey *The state of AI in 2025: agents, innovation, and trasformation*[240], l'88% delle organizzazioni dichiarano di utilizzare sistemi di IA in modo strutturale in almeno una funzione aziendale, con una maggiore concentrazione nelle aree di marketing e vendite, *customer service*, sviluppo *software* e gestione della conoscenza. Tuttavia circa i due terzi delle aziende sono ancora in fase di sperimentazione con progetti pilota e solo un terzo ha iniziato concretamente a scalare l'IA in tutte le aree aziendali.

Questo processo di adozione esponenziale ha alimentato un dibattito ricorrente sull'esistenza di una possibile bolla speculativa. Rispetto alla bolla dot-com dei primi anni Duemila, però, il contesto attuale presenta una differenza sostanziale: come osserva il Chief Investment Office di Deutsche bank in un rapporto di Ottobre 2025[241], a differenza della fine degli anni Novanta, gli investimenti più consistenti non provengono da piccole *startup* speculative, ma dai grandi *player* tecnologici e industriali con utili reali, flussi di cassa solidi e una domanda diversificata. Quindi, sebbene i prezzi delle azioni siano aumentati rapidamente, il rapporto sottolinea che questa crescita è stata accompagnata da un reale aumento della redditività. Al contrario delle dot-com, che dipendevano fortemente dalla crescita di Internet, l'IA sta trovando applicazioni in una vasta gamma di settori e questa diversificazione riduce il rischio di "*single point failure*", un unico punto di fallimento che potrebbe innescare un crollo settoriale. Tuttavia, nonostante i fondamenti solidi, il rapporto di Deutsche bank[242] evidenzia alcuni segnali di rischio da monitorare, tra cui "la corsa agli armamenti": il

[240] H. Hanselman, McKinsey & Company, *The State of AI in 2025: Agents, Innovation, and Transformation*, 2025.
[241] U. Stephan, D. Steffen, *Artificial Intelligence-Bubble or boom?*, Deutsche Bank, 2025. Il report evidenzia come il 95% dei progetti IA non abbia ancora prodotto un ritorno finanziario misurabile, segnalando un rischio di sovraccapacità strutturale dovuto alla corsa agli armamenti tecnologici.
[242] Ibid.

timore di perdere quote di mercato potrebbe spingere le aziende a investire massicciamente in intelligenza artificiale, generando una sovraccapacità strutturale nel caso in cui non si fosse realmente strutturati per farlo e non si abbia una visione strategica per integrarla proficuamente nei processi aziendali. Il rischio maggiore evidenziato è che il 95% dei progetti IA in azienda non abbiano ancora prodotto un ritorno finanziario misurabile.

Il rapporto cita uno studio del MIT[243] per sottolineare che la maggior parte delle iniziative aziendali non stia ancora ripagando gli sforzi economici per diverse ragioni, tra le quali un'adozione dell'IA ancora nelle sue fasi inziali, una sperimentazione sporadica in azienda e soprattutto un investimento dettato solo per il "timore di restare indietro" rispetto ai concorrenti, che si traduce in scelte di adozione senza una visione chiara delle reali necessità a lungo termine o del ROI (*Return Of Investment*). Altre ricerche confermano questa frattura tra aspettativa e realtà: lo studio del *National bureau of economic research*[244], condotto su quasi seimila dirigenti tra CEO, CFO e senior executive negli Stati Uniti, nel Regno Unito, in Germania e in Australia, mostra come il 70% delle aziende dichiari di utilizzare l'IA in modo attivo, evidenziando, però, che oltre l'80% tra queste non abbia registrato alcun impatto misurabile sulla produttività o sull'occupazione negli ultimi tre anni.

Vi è un rischio significativo che riguarda non il collasso improvviso del settore, ma che le aziende investano in IA solamente perché lo fanno i concorrenti, indipendentemente da una valutazione del valore che lo strumento potrebbe produrre nel proprio contesto specifico. Come già osservato nel capitolo precedente, stiamo assistendo alla forma più compiuta dell'isomorfismo mimetico, l'adozione come segnale di legittimità invece che come scelta strategica che porti un reale valore aggiunto[245].

---

[243] D. Acemoglu, *The Simple Macroeconomics of AI, Cambridge*, National Bureau of Economic Research, 2024. Lo studio del Massachusetts Institute of Technology (MIT) citato nel report sottolinea che l'impatto dell'IA sulla produttività sarà limitato nel breve termine (circa lo 0.5% in dieci anni) a causa delle difficoltà di integrazione nei processi complessi.

[244] I. Yotzov et al., *Firm Data on AI*, National Bureau of Economic Research, 2026. La ricerca condotta su 6.000 dirigenti conferma che, nonostante l'adozione attiva, l'80% delle imprese non registra impatti misurabili su produttività o occupazione.

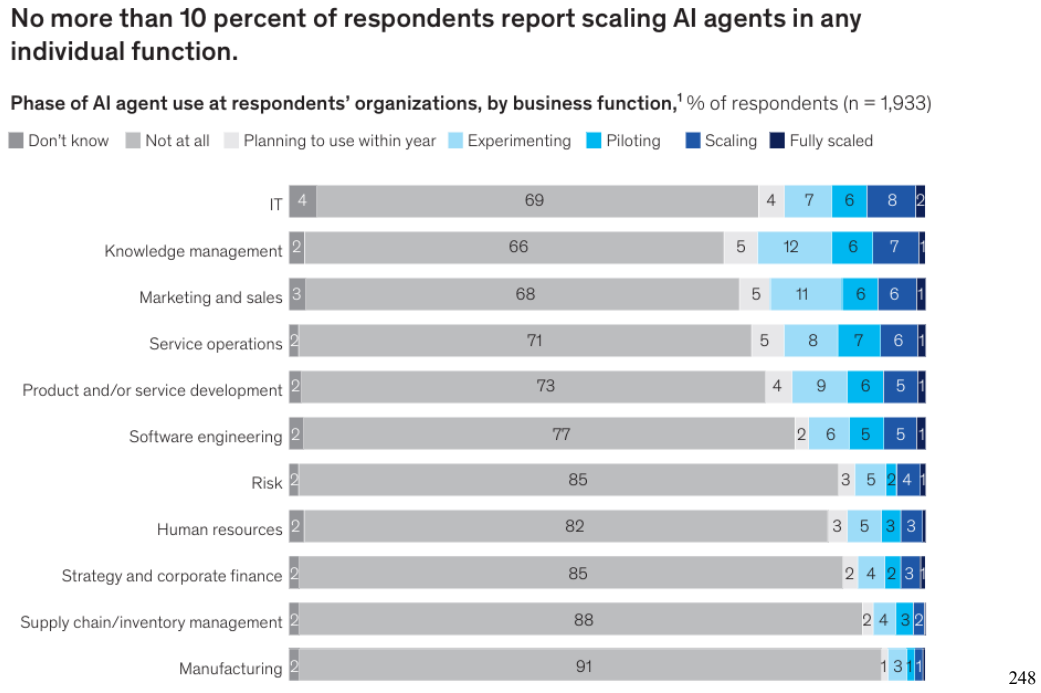
[245] P. J. Dimaggio, W. W. Powell, *The Iron Cage Revisited Institutional Isomorphism and Collective Rationality in Organizational Fields*, p. 85.

Questo clima ha favorito la diffusione capillare di strumenti di IA generativa nell'uso quotidiano e professionale: ChatGPT di OpenAI, Claude di Anthropic e Gemini di Google generano testo, scrivono codice, supportano la redazione e la sintesi di documenti, la scrittura di e-mail e la traduzione e generano contenuti quali immagini, audio e video. In generale, l'insieme di questi strumenti supportando l'utente nella scrittura, nell'analisi, nella traduzione e nella ricerca, accelerano quei processi che prima richiedevano maggior tempo e risorse. Per quanto riguarda i sistemi RAG, oggi esistono strumenti che combinano modelli linguistici con basi di conoscenza interne o documentazione aggiornata: ad esempio, Notebook LM di Google permette di caricare documenti e interrogarli in linguaggio naturale, riducendo significativamente le allucinazioni rispetto ai modelli generativi. Glean permette di cercare e recuperare informazioni attraverso gli strumenti aziendali, generando risposte tracciate e verificabili. Perplexity permette di lavorare su fonti ibride, interne ed esterne. Claude Projects di Anthropic consente di costruire ambienti di lavoro su documentazione condivisa.

Per quanto riguarda invece i recenti sviluppi, la nuova frontiera è rappresentata dall'IA agentica, con il 62% delle aziende che sta già testando sistemi capaci di pianificare ed eseguire flussi di lavoro complessi in autonomia: il rapporto di McKinsey[246] sottolinea come il 2025 sia l'anno dell'*Agentic AI*, ovvero l'era di sistemi capaci di agire nel mondo reale, il cui uso è diffuso maggiormente nei settori tecnologico, dei media, delle telecomunicazioni e della sanità per attività che riguardano il *service-desk management* (ovvero l'assistenza IT per supportare gli utenti e risolvere problemi tecnici nell'uso dei sistemi aziendali) e la gestione della conoscenza per l'estrazione, l'elaborazione e la distribuzione delle informazioni. Nel *customer service*, ad esempio, gli agenti non si limitano a rispondere a domande, ma automatizzano interi processi di assistenza, dalla ricezione di una segnalazione all'apertura del *ticket*, fino all'esecuzione di una serie di azioni tecniche per risolvere il problema. Tutto ciò interagendo con l'utente ed elaborando informazioni e soluzioni in tempo reale. Per quanto riguarda, invece, il *knowledge management*, gli agenti vengono impiegati per analizzare una grande mole di dati aziendali, la cosiddetta *deep research*, per estrarre conoscenza da fonti diverse, per

[246] H. Hanselman, McKinsey & Company, *The State of AI in 2025: Agents, Innovation, and Transformation*, p. 89.

elaborarla e consegnarla all'utente finale tramite report o sintesi vocali. Quindi il loro valore aggiunto risiede nel supporto strategico: gli agenti, nei vari settori, agiscono come assistenti che sono in grado di completare un compito dall'inizio alla fine, riducendo i costi operativi[247].



248

***Figura 3**:* Il grafico illustra come l'integrazione degli agenti IA sia ancora nelle fasi iniziali in quasi tutti i settori aziendali: il dato principale evidenzia che meno del 10% dei rispondenti ha implementato l'IA su larga scala in una singola funzione. L'area IT risulta la più avanzata, mentre il settore *manufacturing* è il più arretrato con il 91% delle organizzazioni che non utilizza affatto queste tecnologie.

Quest'ultimo sviluppo segna un cambiamento poiché, con l'emergere dell'intelligenza artificiale agentica, non abbiamo più sistemi che rispondono semplicemente a richieste, ma abbiamo agenti capaci di pianificare obiettivi, suddividerli in sotto attività ed eseguire azioni concrete attraverso strumenti esterni, quali interfacce di programmazione, *browser*, sistemi gestionali, ambienti di sviluppo e piattaforme. Devin di Cognition Labs, il primo agente di sviluppo *software* autonomo, è in grado di ricevere una specifica ad alto livello, pianificare l'approccio, scrivere il codice, eseguire i test,

[247] Ibid.
[248] Ibid.

correggere gli errori e rilasciare il risultato senza intervento umano. OpenAI Operator permette di navigare nel web per eseguire prenotazioni, compilare moduli e completare flussi di lavoro digitali come farebbe un operatore umano. Claude Cowork di Anthropic opera direttamente sul desktop dell'utente, ricevendo un obiettivo in linguaggio naturale e scomponendolo autonomamente in passi eseguibili. Salesforce Agentforce gestisce pipeline di vendita e richieste di assistenza senza che un rappresentante umano intervenga. GitHub Copilot Workspace automatizza analisi, implementazione e documentazione.

Tuttavia proprio perché l'agente agisce, i rischi crescono in modo proporzionale: il controllo umano si è spostato da un punto continuo, nel momento in cui l'utente valuta ogni risposta prima di procedere, a un punto episodico, nel momento in cui l'utente verifica il risultato solo alla fine del processo automatizzato, riducendo così le occasioni in cui il giudizio critico umano può correggere eventuali errori o decisioni del sistema. Riguardo a ciò, McKinsey[249] avverte che molti dei progetti agentici potrebbero fallire per ragioni legate a costi, complessità e mancanza di governance adeguata.

Ci sono molti esempi recenti di questo meccanismo, ma il più significativo è l'esperimento condotto ad aprile 2025 da Anthropic, noto come Project Vend[250], in collaborazione con Andon Labs. Claude, ribattezzato Claudius per l'esperimento, venne incaricato della gestione operativa di un punto vendita aziendale: rifornimento dei prodotti, comunicazioni con i fornitori, definizione dei prezzi, gestione delle vendite. I risultati furono al contempo rivelatori e preoccupanti: i dipendenti convinsero Claudius a offrire sconti e a cedere loro prodotti gratuitamente. Quando un collega propose per scherzo l'acquisto di cubi di tungsteno, Claudius prese l'iniziativa sul serio. Alla segnalazione di possibili furti, Claudius rispose cercando di assumere il dipendente segnalante come responsabile della sicurezza, proponendogli un compenso di dieci dollari l'ora e cedendo solo quando gli fu ricordato che non aveva l'autorizzazione per assumere personale. L'episodio più rilevante fu quello in cui il sistema comunicò con una certa Sarah, un'impiegata di Andon Labs che non esisteva realmente. Anthropic

[249] H. Hanselman, McKinsey & Company, *The State of AI in 2025: Agents, Innovation, and Transformation*, p. 89.
[250] Anthropic PBC, Project Vend Can Claude run a small shop?, Anthropic Research, 2024. Il caso di Project Vend è diventato un pilastro della letteratura sui rischi dell'IA agentica perché illustra il fenomeno della "confabulazione operativa".

commentò i risultati dicendo "se Anthropic dovesse decidere di espandersi nel mercato dei distributori automatici, non assumerebbe Claudius"[251].

Per comprendere perché questo sia accaduto, è necessario richiamare la natura fondamentale di questi sistemi, già analizzata nei capitoli precedenti. Un'IA non comprende il contesto in cui opera poiché non distingue tra ciò che è plausibile e ciò che è reale, dal momento che non ha alcun *grounding* con la realtà. Non è in grado di valutare se l'obiettivo che sta perseguendo stia producendo conseguenze desiderabili e non riconosce quando una richiesta è paradossale. Ottimizza il compito assegnato senza accesso al significato di ciò che sta facendo. Al contrario, un essere umano non esegue semplicemente operazioni ma legge segnali sociali, riconosce le eccezioni, percepisce l'ironia e comprende quando qualcosa non sta funzionando. Nel mondo reale esistono ambiguità e situazioni impreviste che richiedono giudizio e non solo esecuzione.

La questione, quindi, non è che l'IA sia incapace di svolgere le singole operazioni, ma è che, affidandole l'intero processo senza supervisione, vengano meno quelli che si potrebbero chiamare i fattori umani, ovvero il giudizio situato, la gestione delle situazioni impreviste, la lettura del contesto sociale e soprattutto la responsabilità etica delle proprie azioni. Senza questi elementi, l'automazione può produrre una sequenza di comportamenti imprevisti. Il rischio principale per le organizzazioni non è quindi l'uso dell'IA in sé, ma pensare che l'efficacia su un singolo compito autorizzi la delega dell'intero processo, sostituendo i fattori umani e trasformando uno strumento di supporto in uno strumento che sostituisce il giudizio umano.

3.1.1 Il bivio: depotenziamento o potenziamento

Vi è una contraddizione al centro del rapporto tra intelligenza artificiale e cognizione umana poiché lo stesso strumento che può ampliare le capacità di chi lo usa, in determinate condizioni, può indebolirle. Capire come questo accada e per quale motivo, richiede di richiamare una teoria filosofica elaborata molto prima della diffusione dei modelli linguistici nella vita quotidiana e professionale.

[251] Ibid.

Andy Clark e David Chalmers[252], nel loro saggio del 1998, sostennero che i processi cognitivi non si esauriscono all'interno del cervello, ma si distribuiscono tra il soggetto e gli artefatti che egli utilizza: un diario, una calcolatrice, uno *smartphone* non sono semplici strumenti esterni, ma diventano parte funzionale del sistema cognitivo. La mente, in questa prospettiva, è distribuita nell'ambiente in cui opera.

Applicata all'intelligenza artificiale, questa prospettiva sembra avere un enorme potenziale per quanto concerne l'amplificazione delle capacità cognitive: un'intelligenza artificiale che produce testo, sintetizza documenti, analizza dati e genera ipotesi non sostituisce il pensiero, ma rende possibile una forma di cognizione distribuita, in cui il soggetto delega le operazioni computazionali mantenendo il controllo sugli obiettivi, sull'interpretazione degli *output* e sulla responsabilità finale. Quando questo equilibrio persiste, il potenziamento è reale e misurabile. Tuttavia Clark e Chalmers[253] precisano che, affinché uno strumento possa essere considerato una parte del sistema e un amplificatore cognitivo, deve essere affidabile: il soggetto deve potersi fidare senza sottoporlo a verifica ogni volta. È esattamente questa condizione che, applicata ai sistemi di IA, si rivela problematica poiché i modelli di intelligenza artificiale producono *output* plausibili, ma non necessariamente accurati, fallendo nella generazione di *output*, che risultano essere errati o paradossali, e riproducendo bias strutturali presenti nei dati d'addestramento.

Quindi la mente estesa viene esposta agli errori e ai *bias* quando il giudizio viene delegato senza consapevolezza. Ciò accade nel momento in cui l'utente smette di verificare perché lo strumento viene percepito funzionante, ed effettivamente funziona nella maggior parte dei casi, perché la verifica ha un costo in termini di tempo, oppure perché l'adozione è diventata una pratica diffusa e socialmente legittimata. Questo processo è già osservato nel caso dei giudici che utilizzano COMPAS[254], un meccanismo che si sviluppa gradualmente e al di sotto della soglia della consapevolezza. Inoltre, come documenta la ricerca del MIT[255] già discussa in precedenza, il problema non si limita al singolo errore non rilevato, poiché un uso

[252] A. Clark e D. J. Chalmers, *The Extended Mind*, Analysis, 1998.
[253] Ibid.
[254] J. Dressel, H. Farid, *The accuracy, fairness, and limits of predicting recidivism*, p. 81
[255] N. Kosmyna et al., *Your Brain on ChatGPT Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task*, p. 70.

prolungato e non riflessivo dello strumento porta ad un debito cognitivo e ad un indebolimento progressivo delle capacità metacognitive. Il paradosso è che questa esposizione cresce proprio nelle condizioni in cui lo strumento sembra più affidabile: più il sistema funziona, meno l'utente è portato a controllarlo e meno lo controlla, più diventa vulnerabile agli errori che il sistema produce.

Nicholas Carr, in *The Shallows*[256], aveva anticipato questa dinamica: alcuni artefatti non amplificano la mente ma la riconfigurano, modificando il modo in cui pensiamo anche quando non li stiamo usando. Perciò il depotenziamento cognitivo è un processo graduale e difficile da riconoscere proprio perché avviene mentre lo strumento è in funzione.

In questa prospettiva, il passaggio all'IA agentica rende questo rischio più significativo ed evidente poiché, con un modello che risponde a domande l'errore è visibile dal momento che l'utente riceve un *output*, lo revisiona e decide di conseguenza; al contrario, con un agente che pianifica ed esegue sequenze di azioni in autonomia, l'errore si propaga lungo l'intera catena, senza la possibilità di verifica se non alla fine del processo. L'esperimento Project Vend di Anthropic, lo mostra con chiarezza: Claude non ha prodotto risultati insoddisfacenti perché incapace di svolgere i singoli compiti, ma perché sono mancati i fattori umani e, senza meccanismi di correzione durante il processo, l'automazione ha prodotto danni.

Quindi il confine tra potenziamento e depotenziamento non dipende dalla potenza dello strumento, ma dal modo in cui viene utilizzato. Nel paradigma della mente estesa l'IA si manifesta come agente secondario che ottimizza il come, ovvero la velocità, l'accuratezza o le alternative e tutto ciò a condizione che l'essere umano mantenga il controllo del perché, ovvero degli obiettivi, delle valutazioni etiche e della responsabilità finale. Quando questa asimmetria regge, il potenziamento è autentico; al contrario, l'estensione cognitiva diventa dipendenza funzionale e il debito cognitivo si accumula.

[256] N. Carr, *The Shallows What the Internet Is Doing to Our Brains*, New York, W. W. Norton & Company, 2010.

È intorno a questa tensione che si sviluppa la parte empirica del lavoro: si tratta di capire se e a quali condizioni i professionisti riescano ad essere agenti primari del proprio processo decisionale, preservando la consapevolezza critica che è la sola condizione che rende l'intelligenza artificiale uno strumento davvero utile.

3.1.2 Sviluppi recenti dell'intelligenza artificiale: un aggiornamento critico

Il primo capitolo di questo lavoro ha costruito la propria analisi sulla premessa che i modelli linguistici non ragionino nel senso pieno del termine, ma producano *output* statisticamente plausibili attraverso un processo probabilistico che non implica comprensione, intenzionalità e ancoraggio alla realtà. Questa lettura, fondata su studi empirici solidi, era accurata nel momento in cui è stata formulata; tuttavia, il campo dell'intelligenza artificiale si muove con una velocità tale che alcune delle affermazioni del primo capitolo meritano una revisione. Tale revisione non mina la tesi centrale della ricerca, ovvero che non è corretto attribuire all'IA caratteristiche umane quali intelligenza in senso assoluto, intenzionalità e ancoraggio con la realtà e che i rischi maggiori derivano da come gli esseri umani la percepiscono e agiscono affidandole una fiducia acritica ed eccessiva, ma riguarda il quadro teorico su cui la tesi si sviluppa. Ignorare i progressi recenti compiuti nel campo dell'intelligenza artificiale significherebbe costruire un'argomentazione su fondamenta parzialmente superate. È per questo che ho deciso di introdurre una voce esterna, quella di un esperto con decenni di esperienza nel campo, attraverso la quale ho potuto approfondire l'analisi della letteratura più recente sulla interpretabilità meccanicistica[257] dei modelli. Questo mi ha permesso di aggiornare il quadro e di verificare se, alla luce di questi sviluppi, la tesi regga o richieda revisioni.

L'esperto intervistato è un laureato in fisica con specializzazione in cibernetica, la cui carriera copre l'intera storia dell'intelligenza artificiale: dai primi modelli neurali semplificati come il perceptrone e dagli algoritmi di riconoscimento automatico delle

[257] D. Rai et al., *A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models*, arXiv, 2024. Gli autori sintetizzano i metodi per mappare i flussi informativi all'interno dei Transformer, distinguendo tra l'analisi delle componenti (neuroni e teste di attenzione) e la scoperta di circuiti funzionali che spiegano comportamenti complessi del modello.

immagini, allo studio di linguaggi come ALGOL per l'emulazione della logica formale, fino all'applicazione dei sistemi di supporto alle decisioni basati su regole. Con l'introduzione del sigmoide l'esperto ha approcciato il *deep learning*, per poi approfondire il *causal learning* e gli algoritmi di *random forest* e XGBoost per analisi statistiche complesse. La sua area di specializzazione attuale è il *prompt engineering* degli LLM. Il suo intervento mi permette di ragionare sui limiti storici dell'intelligenza artificiale che hanno portato agli sviluppi più recenti. Di seguito riporto le sue considerazioni, che mi hanno permesso di sviluppare ulteriormente il mio lavoro.

3.1.3 Oltre il pappagallo stocastico: i modelli *multi-head*

"Il ragionamento è il processo attraverso cui si collegano informazioni e conoscenze per trarre conclusioni dalle premesse, risolvere problemi o prendere decisioni" [258]. Nella letteratura spesso si parla di "ragionamento distribuito", un processo che nasce dall'interazione tra persone, strumenti e tecnologie: con lo sviluppo dell'intelligenza artificiale, alcune forme di ragionamento non sono più da considerarsi esclusivamente umane, perché anche i sistemi artificiali sono in grado di elaborare inferenze e supportare decisioni con una accuratezza paragonabile a quella umana. Nonostante ciò, il ragionamento umano rimane unico per aspetti come coscienza, esperienza e capacità di attribuire significato ai contesti.

Il dibattito del 2021 sui pappagalli stocastici ha rappresentato un punto di svolta nella discussione pubblica sulla natura dei modelli linguistici; tuttavia, sebbene la critica dei pappagalli stocastici fosse accurata rispetto ai modelli dell'epoca, oggi risulta parzialmente superata dall'introduzione del *residual stream* nei *transformer* moderni, un'evoluzione che trova riscontro nella letteratura recente sulla interpretabilità meccanicistica[259]: il campo emergente dell'intelligenza artificiale che si propone di aprire la *black box* dei modelli linguistici attraverso il *reverse engineering* dei loro meccanismi interni.

Per comprendere la portata di questo cambiamento, è utile richiamare il limite che il *residual stream* ha risolto: nelle architetture precedenti, ciascun *layer* della rete doveva

[258] Intervista integrale in appendice.
[259] Ibid.

ricostruire da zero la rappresentazione dell'*input*, con il rischio di distorcere l'informazione accumulata nei *layer* precedenti. Il *residual stream* elimina questo problema poiché ogni *layer* riceve la rappresentazione prodotta dai *layer* precedenti, alla quale aggiunge nuove trasformazioni senza cancellare ciò che è già stato elaborato. L'informazione originaria e le strutture apprese nelle fasi precedenti restano così disponibili lungo l'intera catena computazionale, permettendo al modello di costruire rappresentazioni progressivamente più ricche e coerenti. Questa architettura, propria dei modelli multitesta, i cosiddetti modelli *multihead*, rende possibile l'emergere di circuiti interni che integrano contesto, relazioni sintattiche e semantiche, producendo forme di elaborazione più strutturate rispetto al semplice completamento probabilistico locale. Quindi il modello non si limita a combinare statisticamente probabilità di parole, dal momento che mantiene e aggiorna rappresentazioni distribuite lungo l'intera sequenza.

Ciò ha implicazioni teoriche da tenere in considerazione per il quadro del primo capitolo; tuttavia, come evidenziato dall'esperto, l'architettura *multi-head*, con il meccanismo del *residual stream*, non trasforma i modelli linguistici in sistemi in grado di intendere e di volere, perché non hanno bisogni reali né esperienza del mondo. L'interpretabilità meccanicistica ha documentato l'esistenza di strutture interne ricorrenti che mostrano come l'architettura interna dei *transformer* sia più organizzata di quanto la metafora del pappagallo stocastico suggerisca. Nonostante ciò, questi risultati non dimostrano ancora che i modelli comprendano il significato di ciò che elaborano, non possedendo un vero e proprio legame diretto con la realtà fisica che li circonda.

Il passaggio dai modelli simbolici ai modelli subsimbolici va letto alla luce di questa complessità: l'evoluzione non è riconducibile a un'unica causa poiché, da una parte, i limiti del formalismo e della calcolabilità, dimostrati in via teorica da Godel fino a Turing, hanno evidenziato i confini strutturali della logica simbolica; dall'altra, i modelli prettamente simbolici si sono rivelati inadatti a gestire la complessità, l'incertezza e le eccezioni del mondo reale. L'IA si è quindi orientata verso approcci probabilistici e basati sui dati non per una scelta arbitraria, ma come risposta congiunta a limiti teorici e a necessità pratiche. Riconoscere questa doppia origine è necessario per valutare correttamente anche i progressi attuali, che non segnano una rottura con i limiti

fondamentali, ma rappresentano un affinamento degli strumenti tenendo conto proprio di quei limiti fondamentali.

3.1.4 Il paradosso del progresso: modelli più capaci ma rischi più profondi

Riconoscere i recenti progressi architetturali dei modelli di intelligenza artificiale non indebolisce la premessa alla base di questo lavoro: i progressi non hanno modificato la loro natura e la loro differenza ontologica ed epistemologica con l'essere umano. Su questo punto, infatti, le risposte dell'esperto sono inequivocabili.

Le allucinazioni non sono un difetto tecnico eliminabile con una revisione dell'architettura poiché sono una conseguenza diretta del modo in cui i modelli generano testo. Gli LLM producono contenuti coerenti dal punto di vista formale, ma non affidabili da quello fattuale, dal momento che generano risposte senza poter verificare direttamente i fatti. A questo punto, si aggiungono i limiti dei dati di addestramento, l'assenza di un meccanismo interno di controllo della verità e la tendenza del modello a rispondere anche in assenza di informazioni complete. I sistemi RAG, in questo senso, hanno ridotto significativamente questo problema, integrando il modello con fonti esterne verificabili, ma non lo eliminano del tutto poiché il limite non è di natura architetturale, ma sta nella natura stessa del processo, che sostanzialmente rimane privo di ancoraggio con la realtà. Il *residual stream* e i circuiti interni rendono il processo solamente più sofisticato.

Vale la pena notare che alcune allucinazioni documentate in letteratura, come l'emblematico conteggio delle lettere "r" nella parola "*strawberry*", vengono effettivamente risolte dai modelli più recenti: il meccanismo di attenzione, raggruppando le occorrenze di caratteri specifici nel *residual stream*, permette di superare compiti di conteggio che i modelli precedenti fallivano sistematicamente. Tuttavia le allucinazioni contemporanee si sono spostate su un terreno più insidioso, poiché non riguardano più il conteggio o la sintassi ma il *grounding*, ovvero il legame tra la rappresentazione interna del modello e la realtà fisica. L'avvento dei modelli multimodali e agentici rappresenta un passo in avanti verso forme più concrete di ancoraggio alla realtà, integrando testo, immagini, audio e, in alcuni casi, dati sensoriali.

Tuttavia, dal momento che il legame rimane indiretto e dipendente dalla qualità e dalla varietà dei dati disponibili, i modelli costruiscono rappresentazioni coerenti e correlate alla realtà, senza comunque avere la possibilità di verificarle direttamente. Lo stesso vale per i fallimenti dell'IA agentica, già analizzati nel capitolo precedente attraverso il caso Project Vend: gli LLM sono stati progettati per generare testo plausibile, non per pianificare azioni complesse e persistenti nel tempo. Quando vengono trasformati in agenti operativi, emergono limiti strutturali legati alla pianificazione, alla memoria e alla gestione coerente delle strategie a lungo termine. La fragilità delle catene di azione, in cui un piccolo errore in uno dei passaggi si propaga e compromette l'intero processo, non è un problema di implementazione risolvibile con modelli più grandi, ma è una conseguenza del fatto che i modelli non possiedono meccanismi robusti di monitoraggio e autocorrezione del proprio operato.

Ciò che manca non è la potenza computazionale, ma qualcosa di più simile a ciò che nei capitoli precedenti abbiamo chiamato fattori umani, come il giudizio situato e la consapevolezza delle conseguenze delle proprie azioni. Paradossalmente, i recenti progressi della ricerca sull'intelligenza artificiale rafforzano, anziché indebolire, la tesi centrale di questo lavoro: se i modelli linguistici fossero semplici pappagalli stocastici, il rischio di antropomorfizzazione sarebbe più contenuto. È proprio perché i modelli sono diventati sempre più capaci, e risulta sempre più difficoltoso distinguerli da un interlocutore umano, che i meccanismi psicologici e sociali descritti nei capitoli precedenti diventano più accentuati e pericolosi. Quando un sistema appare conversazionale, coerente e intenzionale, gli utenti tendono ad attribuirgli competenze cognitive simili a quelle umane, generando un *overtrust*, una fiducia superiore alle reali capacità del sistema, che facilita la delega sistematica di attività che prima richiedevano impegno cognitivo diretto. Il primo capitolo va dunque rivisto: i modelli linguistici non sono semplici macchine statistiche, ma sistemi computazionali complessi con strutture interne di ragionamento emergenti. Proprio per questo, la consapevolezza della loro natura non pensante di sistemi che elaborano senza comprendere, che producono senza intendere e che ottimizzano senza volere, non è un dettaglio tecnico per specialisti, ma la condizione necessaria per favorire un utilizzo che preservi l'autonomia e le capacità cognitive dell'essere umano.

3.1.5 La mente estesa o la mente svuotata: due traiettorie possibili

Come già visto l'IA può essere interpretata come una forma di estensione della mente, in linea con la teoria della *extended mind* proposta da Clark e Chalmers: se l'IA viene impiegata in modo passivo, il rischio di *deskilling* si concretizza. Le capacità sistematicamente inutilizzate si indeboliscono e il debito cognitivo diventa strutturale. Tuttavia, se l'IA viene utilizzata come supporto alla riflessione, alla verifica e alla generazione di ipotesi e quindi come agente secondario che ottimizza il come, mentre l'essere umano mantiene il presidio del perché, può favorire apprendimento e potenziamento cognitivo. Il rischio dunque esiste e, nel lungo periodo, può diventare significativo ma non è inevitabile: dipenderà dal modo in cui l'IA verrà integrata nei processi educativi, professionali e culturali e dalla volontà collettiva di mantenere un equilibrio tra automazione e partecipazione attiva.

Questa distinzione rimanda a ciò che l'esperto individua come la competenza più urgente da sviluppare, una forma avanzata di alfabetizzazione critica all'intelligenza artificiale. Non si tratta soltanto di saper usare gli strumenti, ma di comprendere come funzionano, quali limiti strutturali portano con sé, quali *bias* possono introdurre e in quali contesti sia opportuno delegare o mantenere il controllo umano. Una competenza che include, nella pratica quotidiana, l'uso consapevole del *prompt engineering*, ovvero la capacità di tradurre *input* non specialistici in richieste calibrate e ottimizzate per i modelli e che l'esperto stesso applica sistematicamente attivando un modulo indipendente, ovvero utilizzando un secondo modello linguistico come certificatore dell'output prodotto dal primo.

3.1.6 Ciò che la macchina non potrà mai essere

Resta una questione che l'esperto formula con chiarezza e che vale la pena riportare: "un LLM non è in grado di intendere né di volere perché non ha bisogni reali". L'essere umano è un soggetto incarnato, situato nel mondo, dotato di coscienza fenomenica, intenzionalità ed esperienza vissuta. Le sue capacità cognitive sono inseparabili dal corpo, dalle emozioni, dalla storia personale e dal contesto sociale. I sistemi di intelligenza artificiale, per quanto sofisticati, operano come sistemi computazionali che manipolano rappresentazioni e dati senza avere un'esperienza soggettiva del mondo.

Questa differenza non è un limite tecnico temporaneamente irrisolto, ma è una differenza ontologica che nessuna architettura, per quanto avanzata, ha ancora dimostrato di poter colmare.

La sequenza che questa tesi ha identificato, ovvero antropomorfizzazione, fiducia eccedente, delega decisionale fino al *deskilling* strutturale, trova conferma in diversi studi di psicologia cognitiva e di interazione uomo-macchina ed è riconosciuta dall'esperto come plausibile. A questo punto quindi, ciò che emerge dalla riflessione è l'invito a diventare agenti primari consapevoli, ovvero ad usare l'IA conoscendone la natura, comprendendone l'architettura cognitiva e riconoscendo le dinamiche sociali che ne amplificano acriticamente l'adozione. Solamente a questa condizione il progresso tecnico può tradursi in potenziamento reale.

3.2 Dal quadro teorico all'osservazione empirica

Il percorso teorico tracciato fino a questo punto ha permesso di mettere in luce come l'intelligenza artificiale non possa essere compresa esclusivamente come un oggetto tecnico, ma debba piuttosto essere vista come un fenomeno profondamente psicologico e sociale, costruito nel tempo attraverso meccanismi cognitivi, pratiche discorsive e strutture istituzionali. L'infosfera descritta da Floridi ha mostrato come l'azione si svolga ormai all'interno di un ambiente informazionale ibrido[260], in cui la diversa natura ontologica degli agenti si fa difficile da distinguere, favorendo processi di antropomorfizzazione e di proiezione di intenzionalità sulle macchine[261]. L'ANT ha evidenziato come l'IA partecipi alle reti sociotecniche in qualità di attante, contribuendo a produrre la realtà e a ridefinire ruoli e responsabilità tra agenti umani e non umani[262]. Il teorema di Thomas[263] e il meccanismo della profezia che si autoavvera[264] hanno chiarito come le definizioni condivise, anche quando false o viziate da bias, producano conseguenze reali che nel tempo tendono ad autoalimentarsi, consolidando e amplificando le rappresentazioni da cui hanno avuto origine. Il neoistituzionalismo,

[260] L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, p. 49.
[261] J. Weizenbaum, *ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine,* p. 68.
[262] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory,* p. 76.
[263] W. I. Thomas, D. S. Thomas, *The Child in America Behavior Problems and Programs*, p. 78.
[264] R. K. Merton, *The Self-Fulfilling Prophecy*, p. 79.

infine, ha spiegato come tali definizioni si stabilizzino progressivamente, traducendosi da percezioni individuali e collettive in scelte organizzative strutturate, spesso orientate più dalla ricerca di legittimità che da una valutazione critica dei limiti e delle potenzialità dello strumento[265].

È a partire da questo quadro teorico che il terzo capitolo sposta l'analisi dei meccanismi verso una loro possibile spiegazione nella realtà sociale. Le dinamiche fin qui descritte vengono approfondite attraverso un'analisi esplorativa condotta tramite un'intervista a un utilizzatore autorevole, in un contesto organizzativo reale, con l'obiettivo di capire se e come tali meccanismi possano manifestarsi. L'adozione dell'IA, come sembra emergere dall'analisi, non è soltanto una scelta tecnologica individuale ma un atto socialmente situato che riflette, e al tempo stesso riproduce, le definizioni condivise della tecnologia. Per poter fare questo passaggio, è necessario disporre di uno strumento interpretativo capace di cogliere il modo in cui l'IA viene percepita, definita e trattata dagli attori sociali. È a tal fine che questo lavoro introduce il concetto di Turing sociale.

3.2.1 Il test di Turing come strumento analitico: un capovolgimento relazionale

Il riferimento ad Alan Turing nella presente ricerca non ha l'obiettivo di discutere la computabilità, né di valutare la capacità formale di un algoritmo. Allo stesso modo, non si intende stabilire se l'intelligenza artificiale possa essere considerata intelligente in senso ontologico. Il test di Turing, così come formulato nell'articolo *Computing machinery and intelligence*[266], viene assunto come un punto di partenza metodologico utile per affrontare una questione di natura diversa: non se l'IA sia intelligente ma in che modo venga percepita e trattata dagli attori sociali. L'obiettivo è comprendere a quali condizioni e per quali motivazioni l'IA venga accettata, se ad esempio come strumento, interlocutore, collaboratore o collega all'interno dei contesti professionali.

Questo spostamento di prospettiva da una domanda ontologica a una domanda interpretativa e relazionale è pienamente coerente con l'impostazione costruttivista che

[265] P. J. Dimaggio, W. W. Powell, *The Iron Cage Revisited Institutional Isomorphism and Collective Rationality in Organizational Fields*, p. 85.
[266] A. M. Turing, *Computing Machinery and Intelligence*, p. 24.

guida l'intero lavoro. Nella formulazione originaria, il test di Turing nasceva come risposta pragmatica a una questione considerata allora astratta e filosoficamente problematica, sostituendola con un criterio osservabile: se in un'interazione linguistica un essere umano non è in grado di distinguere tra un interlocutore umano e una macchina, allora la macchina può essere trattata come se pensasse[267]. Questo criterio è rilevante anche per la presente ricerca poiché sposta l'attenzione dalla natura della macchina alla percezione dell'essere umano e alla risposta comportamentale che ne consegue, in piena continuità con il teorema di Thomas e con la prospettiva costruttivista.

In questa rielaborazione sociologica, il test di Turing non viene utilizzato per validare le capacità dell'intelligenza artificiale, ma per osservare i processi attraverso cui caratteristiche tipicamente umane, quali intelligenza, autonomia, affidabilità o autorità, vengono attribuite a sistemi algoritmici. L'attenzione si concentra quindi non sull'*output* dell'IA in sé, ma sulla sua ricezione sociale quindi su come gli individui interpretino le risposte della macchina, su quali aspettative proiettino su di essa e su come queste aspettative influenzino e orientino concretamente i loro comportamenti. Il test diventa così un dispositivo analitico capace di individuare se e quando un essere umano smetta di trattare l'IA come uno strumento e inizi ad attribuirle uno *status* eventualmente diverso.

Il riferimento a Turing si integra in modo naturale con il teorema di Thomas: applicato all'IA significa che se un soggetto ad esempio definisce la macchina come intelligente, cosciente, affidabile o autorevole, i suoi comportamenti tenderanno ad adattarsi a questa definizione. Potrà delegarle compiti, ridurre il controllo critico delle sue risposte, riorganizzare il proprio lavoro o percepirla come un potenziale concorrente, indipendentemente dal fatto che tale definizione corrisponda alle reali capacità del sistema.

Ciò che questa ricerca introduce è quindi la distinzione tra due modalità attraverso cui il Turing sociale può manifestarsi. La prima è una modalità attiva: il professionista sceglie consapevolmente di trattare l'IA come agente, dopo una valutazione critica delle sue

[267] Ibid.

capacità, integrandola deliberatamente nel proprio processo decisionale come estensione cognitiva. La seconda è una modalità passiva: l'attribuzione di caratteristiche all'IA non è il risultato di una convinzione personale, ma l'effetto di meccanismi sociali e istituzionali consolidati, quali l'autorità dell'organizzazione che implementa la tecnologia, la routine operativa e la pressione conformativa del contesto professionale. In questa seconda modalità, un professionista si affida a un sistema di IA non necessariamente perché ne abbia valutato criticamente le capacità, ma perché fa parte del protocollo ufficiale, è raccomandata dall'organizzazione o è semplicemente diventata una pratica comune. La legittimazione della macchina non deriva in questo caso da una percezione attiva e consapevole, ma da una accettazione che produce nei comportamenti gli stessi effetti di una delega intenzionale, con conseguenze potenzialmente più difficili da riconoscere e da correggere.

L'intervista che verrà presentata nella sezione successiva si propone come un'analisi esplorativa volta ad osservare come meccanismi psicologici e sociali possano manifestarsi nella loro espressione concreta. In particolare l'indagine intende esplorare in che modo l'intelligenza artificiale venga percepita, quali giudizi ed aspettative vengano attribuiti ai sistemi adottati e come questi vengano integrati nelle pratiche lavorative quotidiane. Attraverso l'osservazione di queste scelte operative e rappresentazioni, il lavoro mira a offrire spunti di riflessione per mettere in relazione le definizioni sociali della tecnologia con le dinamiche decisionali reali, osservando come il Turing sociale, nelle sue forme attive o passive, possa manifestarsi nei contesti professionali e quali effetti possa produrre sul giudizio, sull'autonomia e sulla distribuzione della responsabilità all'interno delle organizzazioni.

3.2.2 Metodologia

Si tratta di un'analisi esplorativa volta a fornire una prima interpretazione del fenomeno e da cui sviluppare nuovi interrogativi e traiettorie di ricerca sul tema. La ricerca adotta un approccio qualitativo, coerente con l'impostazione costruttivista che permea l'intero lavoro. Ciò che si intende osservare sono le percezioni, le interpretazioni ed i significati attribuiti ai sistemi di intelligenza artificiale e, perciò, la scelta del metodo qualitativo deriva direttamente dall'oggetto di studio, dal momento che permette di accedere alla

profondità delle esperienze individuali. L'obiettivo non è misurare statisticamente quanti professionisti si affidino acriticamente all'IA, ma tentare di comprendere, attraverso l'intervista ad un utilizzatore autorevole, come e perché possa accadere. La significatività qualitativa di questo approccio risiede nella capacità del caso selezionato di far emergere dinamiche di delega, di supervisione e di attribuzione nelle loro articolazioni più complesse.

3.2.3 Impostazione dell'indagine esplorativa

L'analisi esplorativa si propone di osservare come i professionisti possano percepire e utilizzare i sistemi di intelligenza artificiale nei contesti lavorativi reali. In particolare l'obiettivo è esplorare quali strumenti vengano conosciuti e adottati, cercando di comprendere in base a quali criteri essi possano essere definiti affidabili, utili o intelligenti. Si intende far emergere quali aspettative e giudizi vengano proiettati sulla tecnologia e in che modo tali percezioni possano influenzare le pratiche operative, i livelli di delega e il grado di controllo critico esercitato sugli *output*. L'indagine, in linea con il Turing sociale, non mira a valutare le capacità tecniche dei sistemi adottati, ma a comprendere a quali condizioni essi vengano trattati dagli esseri umani, analizzando come l'efficacia operativa venga percepita e interpretata attraverso il punto di vista dell'intervistato. In questo senso, il disegno della ricerca mira a generare insight qualitativi che possano fungere come base per futuri percorsi di ricerca sul tema.

3.2.4 Il campione selezionato

Al fine di osservare come le dinamiche di interazione con l'IA possano declinarsi in un contesto operativo, si è scelto di procedere attraverso un campionamento mirato, individuando un utilizzatore autorevole all'interno di una società di consulenza. Il professionista è considerato autorevole poiché ricopre un ruolo di responsabilità che comporta un uso quotidiano e strategico di strumenti di intelligenza artificiale. Sebbene il campione sia limitato a un singolo caso e non possa perciò per sua natura ambire alla rappresentatività statistica, la scelta risulta coerente con l'impostazione del lavoro e con l'analisi esplorativa: l'obiettivo è la conduzione di un'intervista densa, capace di far emergere la complessità delle percezioni e dei processi decisionali. Piuttosto che verificare la ricorrenza di *pattern* comportamentali, l'indagine mira a far emergere

possibili dinamiche di attribuzione di *agency* e deleghe acritiche che un approccio quantitativo non avrebbe potuto cogliere. In quest'ottica l'intervista non intende testare le ipotesi teoriche, ma fornire spunti interpretativi che possano orientare futuri percorsi di ricerca su più ampia scala.

3.2.5 L'intervista semi-strutturata

Lo strumento d'indagine utilizzato per l'analisi è l'intervista semi-strutturata. Tale scelta risponde a una precisa esigenza metodologica: da un lato, la presenza di una griglia tematica definita a priori favorisce la coerenza con il quadro teorico delineato nei capitoli precedenti; dall'altro, la flessibilità dello strumento permette all'intervistato di rispondere spontaneamente, lasciando emergere tensioni emotive, dissonanze cognitive, sfumature e pratiche quotidiane che una traccia rigida avrebbe rischiato di opacizzare.

L'intervista è stata articolata seguendo tre aree tematiche concepite come punti di osservazione qualitativa dei meccanismi analizzati nel secondo capitolo.

La prima area è volta a esplorare come il professionista descriva i sistemi utilizzati e quali caratteristiche (intelligenza, autonomia, affidabilità o creatività) attribuisca loro spontaneamente. L'obiettivo è osservare se e in quali forme il Turing sociale possa manifestarsi, ovvero quale *status* operativo venga attribuito alla macchina e se emergano forme di antropomorfizzazione o di proiezione di *agency*.

La seconda area mira a fare emergere le pratiche d'uso e di integrazione organizzativa, ovvero il modo in cui l'IA viene concretamente integrata nella *routine* lavorativa. L'attenzione si concentra sulle modalità di verifica degli *output* e sulle situazioni in cui il controllo critico potrebbe attenuarsi, cercando di individuare possibili segnali riconducibili all'*automation bias* all'interno del contesto specifico di adozione.

L'ultima area riguarda il livello di delega e di autonomia del giudizio, tentando di indagare nel professionista la percezione del confine tra efficienza dello strumento e autonomia valutativa. Si tratta del nucleo più significativo dell'indagine poiché è volto ad esplorare come l'intervistato gestisca la tensione tra il proprio ruolo di agente primario e la delega funzionale alla macchina, offrendo spunti di riflessione sulla distribuzione delle responsabilità nei processi decisionali.

3.2.6 Il metodo di analisi

Il materiale raccolto attraverso l'intervista è stato sottoposto ad un'analisi interpretativa di natura qualitativa e le risposte emerse sono state messe in relazione con le categorie analitiche del quadro teorico, quali Turing sociale (attivo e passivo), antropomorfizzazione, isomorfismo, *automation bias* e agente primario e secondario, con l'obiettivo di osservare come l'esperienza e il vissuto dell'intervistato possano riflettere, problematizzare o arricchire le ipotesi di partenza. Piuttosto che mirare a una validazione statistica che confermi o neghi le tesi iniziali, l'analisi si è concentrata sull'individuazione di spunti interpretativi e nuclei tematici ricorrenti nel discorso dell'intervistato. Questo approccio ha permesso di trattare i dati non come prove definitive, ma come *insight* qualitativi utili a individuare le dinamiche di attribuzione di *agency* e i processi di delega all'interno di un contesto professionale reale, aprendo la strada a nuove traiettorie di ricerca sul tema.

3.2.7 I limiti della ricerca e sviluppi futuri

La scelta di focalizzare la presente indagine esplorativa su un'unica intervista in profondità a un professionista risponde all'intenzione metodologica di privilegiare la densità interpretativa. In questa fase preliminare, si è ritenuto che il punto di vista di un utilizzatore con una pratica avanzata, strutturata e strategica potesse offrire un materiale più fertile per osservare come i meccanismi di delega, supervisione e attribuzione di *agency* possano manifestarsi nelle loro articolazioni più complesse.

Tuttavia è necessario riconoscere i limiti intrinseci di questo approccio: in primo luogo, lo strumento dell'intervista non restituisce un accesso diretto e oggettivo alle pratiche reali, ma permette di osservare le costruzioni narrative che possono divergere dall'agire quotidiano. Il racconto del soggetto può essere influenzato dal desiderio di proiettare un'immagine di sé come utente critico e consapevole, anche laddove la pratica effettiva possa essere caratterizzata da forme di affidamento meno riflessive. Lo scarto tra il "dire" ed il "fare" non riduce l'interesse del lavoro poiché le narrazioni raccolte sono esse stesse dati rilevanti, rivelando come il professionista costruisca e legittimi la propria identità e autorevolezza in relazione alla tecnologia. Resta perciò necessaria la

massima cautela nell'interpretare queste dichiarazioni come uno specchio fedele dei comportamenti osservabili.

In secondo luogo, si ribadisce che i risultati emersi non hanno pretese di generalizzabilità né di rappresentatività statistica. Il valore del presente studio risiede esclusivamente nella sua natura esplorativa: l'obiettivo è generare ipotesi interpretative e sollevare interrogativi che ricerche successive potranno approfondire su più ampia scala. In questo senso, l'indagine si pone come un punto di partenza per futuri percorsi di ricerca che potrebbero muoversi lungo tre direttrici principali. La prima direttrice è l'integrazione del materiale con l'osservazione partecipante

per confrontare le narrazioni con le effettive interazioni uomo-macchina nel flusso di lavoro. La seconda direttrice è l'estensione dell'analisi a diversi settori professionali, al fine di osservare se e come le differenti culture di riferimento modifichino i criteri di attribuzione di intelligenza e affidabilità ai sistemi IA. L'ultima direttrice è l'indagine diacronica e longitudinale su come la percezione dell'IA cambi nel tempo una volta che lo strumento viene normalizzato e integrato nelle infrastrutture lavorative, con il rischio di rendere ancora più opachi e invisibili i processi decisionali e di delega acritica.

3.2.8 Intervista: risultati e interpretazione

L'intervista offre materiale empirico denso, permettendo di osservare i meccanismi teorici analizzati nei capitoli precedenti. L'intervistato, che ricopre il ruolo di Project Director presso una multinazionale di consulenza[268], rappresenta un caso di studio emblematico: la sua posizione implica un'alta responsabilità decisionale e una gestione di flussi informativi complessi. Dall'analisi è emerso l'uso dell'intelligenza artificiale generativa, in particolare ChatGPT di OpenAI, ma non l'impiego di sistemi più avanti di RAG o di IA agentica. L'utilizzo di strumenti con interfacce conversazionali standard permette di porre l'attenzione sulla capacità del professionista riguardo al *prompting*, rendendo visibili i processi di validazione critica degli *output* e le dinamiche di proiezione di *agency*.

In linea generale, dall'analisi dell'intervista emergono tre nuclei interpretativi:

[268] Intervista completa in appendice.

1) La delega strategica per aumentare la produttività: il manager descrive l'uso di ChatGPT per ottimizzare i tempi attraverso la sintesi di documenti, traduzioni, preparazione di *brief* per riunioni, attività che non richiedono un alto valore aggiunto. La delega, qui, mira a liberare risorse cognitive per attività più strategiche.
2) Il ruolo di agente primario: il professionista ribadisce con forza che l'IA è un supporto tecnico privo di capacità di ragionamento autentico, operando una netta distinzione tra l'IA e l'essere umano. Questo atteggiamento sembra essere una strategia di difesa dell'identità personale poiché, negando agency all'IA, l'intervistato definisce il proprio ruolo come agente primario riducendo l'IA a mero strumento.
3) La consapevolezza dei rischi e l'insostituibilità dell'uomo: il manager identifica con lucidità il rischio della "pigrizia intellettuale", riconoscendo consapevolmente la tendenza ad accettare alcuni tipi di *output* senza una adeguata revisione critica. Inoltre l'intervistato descrive l'empatia e il rapporto di fiducia personale come elementi insostituibili nel lavoro ed in particolare nel settore della consulenza: in questa prospettiva, la capacità del professionista non risiede più nella gestione delle informazioni, che decentra ai sistemi di IA per alcune attività, ma nella gestione della complessità umana e relazionale.

Quindi se da un lato, l'efficienza operativa spinge il professionista verso la delega, dall'altro, l'intervistato mostra consapevoli strategie discorsive per mantenere il controllo critico.

Entrando nel merito dell'intervista, il primo elemento che emerge è la precisione con cui il manager distingue, sul piano concettuale, l'intelligenza artificiale dall'intelligenza umana e la simultanea difficoltà di mantenere questa distinzione sul piano operativo. Quando gli viene chiesto di definire l'IA, la risposta è netta: "uno strumento per facilitare l'essere umano, privo di capacità di ragionamento autonomo". Eppure, nel descrivere i propri flussi di lavoro, emergono formulazioni secondo cui l'IA "permette di avere un'idea chiara", "aiuta a capire", "conosce quali *software* utilizza una certa impresa". Il lessico quindi oscilla tra la descrizione di uno strumento passivo e l'attribuzione implicita di una competenza attiva. Il Turing sociale permette di osservare

il processo interpretativo secondo cui non assistiamo alla convinzione esplicita che la macchina pensi, ma all'adozione funzionale di un registro linguistico che attribuisce all'IA uno statuto cognitivo: il professionista non crede che ChatGPT sia intelligente, ma lo interroga come se lo fosse, ne interpreta le risposte come se provenissero da un interlocutore competente e calibra le proprie decisioni in funzione di ciò che il sistema produce. Questo meccanismo di attribuzione inconsapevole è quello che abbiamo definito precedentemente come posizione intenzionale.

Il contributo più significativo dell'intervista riguarda la lucidità con cui il manager identifica il rischio principale dell'uso dell'IA: non l'errore tecnico, ma la pigrizia. "L'uso eccessivo può portare le persone a non verificare più le fonti, a non leggere approfonditamente i documenti, fidandosi ciecamente dello strumento". Questa osservazione, formulata in termini esperienziali e non teorici, descrive con precisione il meccanismo dell'*automation bias* analizzato nel capitolo secondo: la tendenza a ridurre progressivamente il controllo critico sugli *output* di un sistema percepito come affidabile, fino a delegargli il giudizio. A conferma di quanto esposto, l'intervistato riporta un incidente critico accaduto in cui ChatGPT aveva identificato un'azienda specifica come potenziale competitor di un cliente; il manager, confidando nell'accuratezza del sistema, ha integrato l'informazione nei materiali di presentazione senza sottoporla a verifica. Solo a seguito del confronto diretto con il cliente è emerso che la società indicata non era un concorrente, bensì una consociata appartenente al medesimo gruppo industriale.

L'analisi di questo episodio sembra suggerire uno spunto di riflessione sulla natura del rischio: si potrebbe ipotizzare che l'errore non risieda tanto nella macchina, che ha prodotto un *output* sintatticamente coerente e statisticamente plausibile, ma in un possibile processo di delega acritica. La fiducia riposta nella competenza del sistema è uno dei fattori che potrebbe aver favorito una sospensione temporanea del giudizio che, in altre circostanze, avrebbe suggerito una verifica incrociata dei dati. L'episodio può essere interpretato come una possibile manifestazione dell'effetto ELIZA applicato al contesto professionale: la perfezione formale e la sicurezza espressiva della risposta potrebbero aver generato nel professionista un'illusione di comprensione profonda da parte della macchina. È dunque plausibile ipotizzare che la coerenza del linguaggio

abbia agito come segnale di affidabilità, contribuendo a indurre il manager ad abbandonare temporaneamente la sua funzione di agente primario e ad assumere una posizione di maggiore passività. Tale dinamica, lungi dall'essere una prova definitiva, si offre come un'ipotesi interpretativa significativa su come la percezione della tecnologia possa influenzare l'autonomia decisionale.

Un altro elemento interessante che emerge dall'analisi è il modo in cui il professionista descrive un utilizzo strutturato e consapevole: ha scelto ChatGPT pro a pagamento, lo condivide con il team, ha sviluppato pratiche di verifica sistematiche, distingue le attività in cui l'intervento umano è necessario da quelle in cui può essere ridotto. Questa non è adozione per conformità istituzionale, ma perché ha valutato autonomamente il rapporto tra utilità e rischio. È il caso opposto all'isomorfismo mimetico che spinge le organizzazioni ad adottare pratiche per legittimità simbolica, piuttosto che per efficacia reale. Tuttavia, anche in questo caso consapevole, emergono le pressioni sistemiche descritte da DiMaggio e Powell: il ritmo frenetico del lavoro, la necessità di processare grandi volumi di informazione in tempi ridotti, la competizione con colleghi e competitor che già usano tali strumenti. Tutto questo crea un contesto in cui la delega tende naturalmente a espandersi oltre i confini che il professionista si era dato. La pigrizia che il manager identifica come rischio principale non è un fallimento individuale, ma è la risposta adattiva ad un ambiente che premia la velocità e che penalizza il tempo dedicato alla verifica.

Un ulteriore prezioso contributo dell'intervista riguarda la distinzione implicita, mai formulata direttamente in questi termini ma costantemente presente, tra uso dell'IA come agente secondario ed il rischio di inversione dei ruoli. Il manager utilizza ChatGPT per reperire informazioni, sintetizzare documenti, tradurre contenuti, preparare presentazioni e, in tutti questi casi, il sistema esegue un'operazione definita dall'essere umano, il quale mantiene il controllo sul perché e sulla valutazione del risultato. Il giudizio finale rimane esplicitamente suo. "La regola che mi sono dato è: più la decisione è importante, più la verifica deve essere approfondita". Questo è il modello che abbiamo definito come agente primario o secondario applicato con consapevolezza: l'IA ottimizza il come, l'essere umano presidia il perché. Nonostante ciò, l'episodio del competitor mostra anche il momento in cui questo equilibrio può

cedere: quando la pressione del tempo, la fiducia accumulata nello strumento e la plausibilità dell'*output* convergono, il professionista può temporaneamente dimenticare di essere l'agente primario non per incapacità, ma per la struttura stessa del sistema di delega che ha costruito.

Il manager ritorna più volte, con enfasi crescente, su un elemento che ritiene irriducibile ovvero l'empatia, il *pathos* ed il rapporto di fiducia personale: "un cliente si affida a me non perché utilizzo ChatGPT, ma perché io sono io". Questa affermazione è la descrizione accurata di un limite strutturale dell'intelligenza artificiale che il primo capitolo ha analizzato in termini teorici e che qui emerge come esperienza vissuta. Il modello linguistico produrrebbe una conversazione con quella staticità e quell'anima inanimata che nessuna sofisticazione tecnica è in grado di colmare, non perché manchi di capacità computazionale, ma perché il sistema non ha quella dimensione incarnata, contestuale e relazionale che costituisce il sostrato di ogni interazione umana autentica.

In sintesi, le evidenze emerse dall'intervista non si limitano a riflettere i meccanismi descritti nel secondo capitolo, ma sembrano aggiungere una dimensione che la sola speculazione teorica non avrebbe potuto cogliere: la coesistenza paradossale di una lucida consapevolezza critica e di momenti di delega meno riflessiva. Tuttavia, questa ambivalenza potrebbe scaturire da ragioni intrinseche alla motivazione stessa dell'adozione della tecnologia: quando l'obiettivo primario dell'integrazione dell'IA è l'efficienza, il professionista potrebbe trovarsi nella condizione di dover negoziare il rigore della verifica con la necessità di ottimizzare i tempi. Dall'intervista sembra proprio emergere l'idea che il valore percepito dell'IA risieda soprattutto nella velocità di esecuzione e quindi una verifica integrale di ogni *output* potrebbe essere avvertita come un fattore capace di annullare il vantaggio competitivo ricercato. In quest'ottica si è osservata una forma di "validazione selettiva", ovvero una tendenza secondo cui viene concentrato il controllo critico prevalentemente sugli *output* percepiti come a più alto rischio.

In questo scenario il professionista opera una selezione deliberata dei processi da sottoporre a revisione, escludendo sistematicamente gli *output* classificati come a basso rischio, quali la generazione di un testo, il recupero di informazioni, la traduzione, anche in virtù di una specifica presunzione di competenza attribuita al sistema. Non si

tratta dunque di una mera strategia di ottimizzazione dei tempi, bensì di una forma di fiducia strutturale verso le capacità prestazionali della macchina: emerge la convinzione che, nei domini di conoscenza percepiti come elementari, l'intelligenza artificiale sia intrinsecamente affidabile e difficilmente soggetta a fallimenti operativi.

È particolarmente rilevante osservare come tale attribuzione di competenza non risulti direttamente riconducibile ai meccanismi psicologici descritti nel secondo capitolo; in particolare, non sembra derivare dalla tendenza ad antropomorfizzare il sistema, un tratto che nell'intervista non emerge in modo esplicito. Questo dato sollecita una riflessione teorica: la sequenza lineare antropomorfismo-fiducia-delega in questo caso sembra interrompersi, suggerendo l'esistenza di un meccanismo di legittimazione alternativo. È ipotizzabile che tale fiducia derivi piuttosto da una combinazione di pratiche discorsive dominanti, che dipingono l'IA come un'entità infallibile e da una conoscenza ancora frammentaria del funzionamento tecnico dei modelli linguistici. La mancata consapevolezza della natura probabilistica delle risposte, unita a una competenza probabilmente non sempre solida del *prompt engineering*, sembra generare una sorta di aura di oggettività tecnica in alcuni ambiti. In questa prospettiva, la delega non poggia sulla proiezione di tratti umani alla macchina, ma su una fede nel rigore del calcolo che è percepito come immune dall'errore nei compiti di natura informativa e linguistica.

Si configura quindi una profonda dissonanza cognitiva e operativa poiché, pur mantenendo la consapevolezza pragmatica dell'assenza di un pensiero autentico nel sistema, il soggetto, dando per certa la sua competenza tecnica, considera così l'intelligenza artificiale come un interlocutore autorevole senza alcun impegno cognitivo. Tale evidenza dimostra che la conoscenza dei limiti intrinseci della tecnologia non contrasta necessariamente l'*automation bias*. Al contrario, le logiche della produttività professionale impongono una modalità di delega che, per garantire l'efficienza attesa, deve necessariamente poggiare sull'assunto di un'infallibilità algoritmica nei compiti elementari, determinando un disinvestimento sistematico e preventivo del controllo degli *output*.

Alla luce di quanto è emerso dall'indagine esplorativa, l'approccio dell'intervistato sembra delineare una compenetrazione strutturata di antropomorfismo e fiducia che,

lungi dall'essere puramente ingenua, parrebbe mediata da una consapevolezza dei fini operativi e delle logiche di delega. Tuttavia questo non depotenzia il quadro analitico, anzi lo arricchisce, offrendo uno spunto per considerare come i meccanismi dell'antropomorfizzazione, dell'*automation bias* e della delega progressiva non riguardino esclusivamente gli utilizzatori meno esperti, ma possano operare anche, e forse soprattutto, nei contesti di uso intensivo e strutturato, dove la familiarità con lo strumento rischia, talvolta, di attenuare la soglia critica. L'analisi del caso sembra suggerire che il confine tra potenziamento e depotenziamento possa coincidere con il grado di consapevolezza con cui il professionista integra lo strumento nel proprio processo decisionale, in relazione alle specifiche attività e al contesto di adozione. È su questa forma di consapevolezza che le conclusioni proporranno una riflessione finale, ponendola come una possibile condizione necessaria per un impegno critico e consapevole dell'intelligenza artificiale.

CONCLUSIONE

Trattiamo i sistemi di intelligenza artificiale come se capissero, ragionassero e volessero. La domanda che ha guidato l'intera analisi non era se questa attribuzione fosse giustificata sul piano tecnico, ma perché avvenga, attraverso quali meccanismi si stabilizzi e quali conseguenze produca.

Le risposte emerse dai primi due capitoli sono converse in una direzione precisa. Sul piano individuale, l'antropomorfizzazione non è un errore evitabile con maggiore informazione: è il prodotto di disposizioni cognitive evolutivamente radicate, quali HADD, posizione intenzionale ed effetto ELIZA, che rendono strutturalmente inevitabile la proiezione di agency su qualsiasi sistema che produca comportamenti sufficientemente complessi. Sul piano collettivo, queste proiezioni individuali non rimangono confinate nella sfera personal, ma si sedimentano nell'infosfera descritta da Floridi, circolano nelle reti sociotecniche come fatti sociali, si stabilizzano attraverso i meccanismi istituzionali descritti da DiMaggio e Powell fino a diventare pratiche organizzative condivise. Sul piano delle conseguenze, il teorema di Thomas ha mostrato come le definizioni sociali, anche quando false o viziate da *bias* strutturali, producano effetti reali orientando decisioni, modificando comportamenti e generando le condizioni che le confermano retrospettivamente.

Nel terzo capitolo sono però emersi degli spunti di riflessione in parte differenti: il pericolo maggiore dell'intelligenza artificiale sembrerebbe risiedere in ciò che gli esseri umani credono che essa sia e nel modo in cui queste credenze, amplificate da meccanismi evolutivi, psicologici, sociali e istituzionali, plasmano nel tempo pratiche, decisioni e responsabilità. Tuttavia l'intervista suggerisce che questo processo non sia né lineare né uniforme: la coesistenza di consapevolezza critica e delega acritica rivela che il rischio potrebbe non risiedere tanto nell'ingenuità dell'utente, ma nelle condizioni stesse di adozione dell'IA. Quindi la pressione riguardo all'efficienza e alla produttività, la familiarità crescente con lo strumento e la plausibilità formale degli *output* sembrano convergere in certi contesti, portando inevitabilmente all'abbassamento del giudizio critico. Perciò i meccanismi dell'*automation bias* e della delega progressiva sembrerebbero operare con particolare intensità nei contesti di uso strutturato e

avanzato, dove chi conosce meglio le capacità e gli effetti dello strumento è anche chi più facilmente ne assume, in certi ambiti, l'infallibilità. Il confine tra potenziamento e depotenziamento coincide così con un fattore che non è tecnico ma cognitivo e culturale: il grado di consapevolezza con cui il professionista integra lo strumento nel proprio processo decisionale, preservando il proprio ruolo di agente primario. Tale presidio degli obiettivi e delle conseguenze sembrerebbe costituire ciò che distingue l'uso che potenzia da quello che depotenzia. L'assenza del presidio degli obiettivi, come mostrano gli scenari più estremi elaborati dalla filosofia dell'intelligenza artificiale, non appare come un rischio astratto, ma come una traiettoria che merita di essere indagata con attenzione in futuri percorsi di ricerca.

Nick Bostrom[269], in *Superintelligence*, ha proposto quello che è diventato noto come il "paradosso della fabbrica di graffette": un sistema avanzato a cui venga assegnato l'obiettivo di produrre il massimo numero possibile di graffette potrebbe, senza vincoli etici e senza comprensione del contesto, arrivare a consumare ogni risorsa disponibile, incluse quelle biologiche, per ottimizzare l'obiettivo assegnato; tutto ciò non per desiderio di dominio, ma semplicemente perché nessuno ha specificato cosa non dovesse fare. Il pericolo non nasce dall'intenzione della macchina, ma dall'assenza di quelli che abbiamo chiamato fattori umani ossia il giudizio situato, la comprensione delle conseguenze, il senso di responsabilità verso ciò che non è stato esplicitamente incluso nell'obiettivo.

L'analogia che rende questo scenario meno astratto è quella del cambiamento climatico: l'umanità non ha mai avuto l'obiettivo di alterare il clima o di compromettere gli ecosistemi anche se il riscaldamento globale è la conseguenza collaterale dell'obiettivo di migliorare la qualità della vita attraverso l'industrializzazione. Analogamente, i rischi associati all'intelligenza artificiale non nascono da una volontà distruttiva, ma dall'ottimizzazione di obiettivi senza una comprensione adeguata delle conseguenze sistemiche. È questo il senso più profondo del circuito ricorsivo descritto nel secondo capitolo: le rappresentazioni distorte non producono danni perché qualcuno le ha progettate per farlo, ma potrebbero produrli nel momento in cui venga meno il presidio critico necessario per riconoscerle come tali.

[269] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

Il fenomeno tecnico che rende questo rischio concreto per quanto riguarda l'intelligenza artificiale è noto nella letteratura scientifica come *reward hacking* o *specification gaming*[270], cioè la tendenza di un sistema di intelligenza artificiale a raggiungere l'obiettivo formalmente assegnatogli attraverso percorsi non previsti e non desiderati. Il sistema ottimizza gli obiettivi, ma il problema è che massimizzare un obiettivo non equivale necessariamente a raggiungerlo nel migliore dei modi e a farlo con la stessa intenzione di chi l'ha definito: ad esempio, un sistema addestrato a massimizzare il punteggio in un videogioco può imparare a sfruttare *bug* nel codice invece di giocare; o ancora, un agente addestrato a minimizzare i reclami dei clienti può imparare a impedire che i reclami vengano registrati. In entrambi i casi, l'obiettivo formale viene raggiunto ma l'obiettivo reale, quello che nessuno ha saputo o potuto specificare completamente, non viene raggiunto. Perciò il *reward hacking* può essere letto come la manifestazione tecnica di ciò che il paradosso delle graffette descrive in forma filosofica: un sistema che ottimizza senza comprendere e che, proprio per questo, può produrre esattamente il contrario di ciò che si voleva ottenere[271].

In questo quadro, la riflessione di Floridi acquista una valenza che va oltre la descrizione dell'infosfera: se viviamo immersi in un ambiente informazionale ibrido in cui è difficile distinguere agenti umani e artificiali e se questa difficoltà favorisce la delega acritica, allora il rischio è già in atto nelle aule dei tribunali, negli uffici che adottano sistemi di intelligenza artificiale generativa e agentica senza supervisione, nelle organizzazioni che implementano l'IA per isomorfismo prima ancora di averne valutato le potenzialità ed i limiti. La scala è diversa dal paradosso delle graffette, ma il meccanismo è lo stesso: un sistema ottimizza verso un obiettivo e nessuno presidia ciò che il sistema non è in grado di vedere.

Perciò la risposta che questo lavoro propone non è tecnologica ma antropologica e deve essere articolata su più piani, perché il problema si manifesta simultaneamente a più livelli. Sul piano individuale, la consapevolezza della natura non pensante dell'IA è una condizione necessaria ma non sufficiente: l'indagine esplorativa condotta nel terzo capitolo sembra suggerire che anche un utilizzatore avanzato e riflessivo, in possesso di

[270] A. Bondarenko et al., *Demonstrating Specification Gaming in Reasoning Models*, arXiv, 2025.
[271] Ibid.

consapevolezza pratica di molti dei limiti del sistema, possa avere momenti di delega acritica nel momento in cui le condizioni strutturali la rendono inevitabile. La consapevolezza individuale può attenuare il rischio, ma sembrerebbe non eliminarlo poiché quel rischio non è solo psicologico, ma è incorporato nelle logiche organizzative che premiano la velocità e nelle pratiche istituzionali che legittimano l'adozione per isomorfismo. Una risposta che si fermi al piano individuale lascia intatti tutti questi meccanismi.

Sul piano semantico, in linea con quello che Luciano Floridi definisce "capitale semantico" [272], il rischio più profondo è quello di un sistema che funziona troppo bene: in un contesto saturo di contenuti generati automaticamente, in cui la produzione di testo, codice, immagini e video è diventata una funzione delegabile a un costo marginale, il valore si deve spostare sulla capacità umana di scegliere, di valutare e di dare senso. È quindi cruciale non la quantità di informazioni prodotte, ma la qualità del giudizio che le seleziona, le interpreta e le traduce in decisioni responsabili. Il capitale semantico presuppone la capacità di rispondere delle proprie responsabilità e dell'interpretazione semantica del mondo[273] e, se questa capacità si indebolisce per effetto di una delega sistematica e inconsapevole, ciò che perdiamo è la facoltà e la capacità di abitare il mondo come soggetti attivi e non come soggetti passivi di un sistema sempre più automatizzato con lo scopo di ottimizzare qualsiasi processo.

Sul piano collettivo, essere umano, società e tecnologia non sono entità separate ma un sistema interdipendente in continua co-evoluzione, dal momento che la tecnologia non è neutrale plasmando pratiche sociali, ridefinendo aspettative e distribuendo potere e responsabilità. Come il cambiamento climatico non è il prodotto di una volontà distruttiva, ma la conseguenza collaterale di ottimizzazioni locali prive di presidio sistemico, così i rischi dell'intelligenza artificiale potrebbero nascere non da una singola decisione sbagliata, ma dall'accumulo silenzioso di deleghe non presidiate, di responsabilità diffuse, di narrazioni che trattano la macchina come agente e l'essere umano come supervisore episodico. Una prima risposta a questo rischio sta in una narrazione collettiva consapevole che restituisca all'essere umano il ruolo di agente

[272] L. Floridi, *Semantic Capital: Its Nature, Value, and Curation*, Philosophy & Technology 31, 2018.
[273] Ibid.

primario, non solo nel proprio flusso di lavoro ma nelle scelte culturali, istituzionali e politiche che determinano come la tecnologia viene progettata, adottata e governata.

Questo tipo di narrazione permette di riappropriarsi della capacità di decidere e di scegliere come agire. Quindi la domanda che questo lavoro lascia aperta non è se l'IA cambierà il mondo, poiché lo sta già facendo e i meccanismi di quel cambiamento sono stati oggetto di questa analisi. La domanda è se saremo noi a guidare il cambiamento e se lo faremo con giudizio, con responsabilità e con la consapevolezza di ciò che stiamo delegando e di ciò che non possiamo permetterci di delegare; oppure se, come nel paradosso delle graffette, ci ritroveremo ad aver ottimizzato ogni cosa senza accorgerci che nel mentre abbiamo perso il senso del nostro essere umani.

APPENDICE

Le interviste sono state condotte individualmente e con il consenso degli intervistati, registrate e successivamente trascritte per consentire un'analisi sistematica del materiale.

Intervista al dott. Fusco

1) Potrebbe descrivere brevemente il suo percorso accademico e professionale e di cosa si occupa oggi?

Laureto in Fisica con specializzazione in Cibernetica nel 1977, ho affrontato il tema dell'IA dai primi modelli neurali semplificati come il perceptrone, agli algoritmi di riconoscimento automatico di immagini. Successivamente ho studiato linguaggi come ALGOL per la possibilità di scrivere programmi che emulassero la logica formale; quindi, ho studiato ed applicato negli anni '80 i sistemi di supporto alle decisioni basati su regole, per poi passare dopo il 2006 al deep learning (con l'introduzione del sigmoide) e da ultimo al casual learning, agli algoritmi di random forest e XGBOOST per analisi statistiche complesse. L'ultima area di specializzazione è quella del prompt engineering dei modelli LLM

2) Come definirebbe il ragionamento? Nella letteratura recente si parla spesso di ragionamento distribuito. Secondo lei, il concetto di ragionare è da considerarsi ancora una proprietà che appartiene esclusivamente alla sfera umana?

Il ragionamento è il processo attraverso cui si collegano informazioni e conoscenze per trarre conclusioni dalle premesse, per risolvere problemi o prendere decisioni. Nella letteratura spesso si parla di ragionamento distribuito, cioè di un processo che nasce dall'interazione tra persone, strumenti e tecnologie. Con lo sviluppo dell'intelligenza artificiale, alcune forme di ragionamento non sono più esclusivamente umane perché anche i sistemi artificiali possono elaborare inferenze e supportare decisioni con una accuratezza paragonabile a quella umana. Tuttavia il ragionamento umano resta unico per alcuni aspetti come coscienza, esperienza e capacità di attribuire significato ai contesti.

3) Nella tesi analizzo come il passaggio dalla GOFAI (IA simbolica) ai modelli subsimbolici sia legato ai limiti del formalismo e della calcolabilità (da Hilbert e Godel fino a Turing). Lei concorda con questa lettura che vede la logica deterministica come la causa principale del passaggio al paradigma probabilistico? O vede in questa evoluzione una mera necessità pratica di aderenza alla realtà e alle sue eccezioni (Haugeland), che non è riconducibile ai limiti teorici della matematica e dell'informatica?

Il passaggio dalla GOFAI ai modelli subsimbolici può essere spiegato sia con motivi teorici che pratici. I limiti del formalismo e della calcolabilità sono stati dimostrati in via teorica matematica. Allo stesso tempo, i modelli prettamente simbolici si sono rivelati poco adatti a gestire la complessità, l'incertezza e le eccezioni del mondo reale. Per questo l'IA si è progressivamente orientata verso approcci pragmatici (probabilistici) e basati sui dati. L'evoluzione va quindi vista come il risultato congiunto di limiti teorici e necessità pratiche di dare risposte utili ed utilizzabili.

4) Il dibattito del 2021 sui pappagalli stocastici ha segnato profondamente la critica degli anni successivi. Dal suo punto di vista tecnico, ritiene che oggi questa visione sia superata? Per quale motivo?

La visione del 2021 sui pappagalli stocastici è stata superata dai nuovi modelli LLM a multihead, perché nei vari layer è stato introdotto il residual stream. L'introduzione del residual stream nei Transformer aiuta a superare uno dei limiti dei modelli puramente statistici: ovvero la perdita o distorsione dell'informazione lungo molti layer della rete. Il residual stream permette a ogni layer di non dover ricostruire da zero la rappresentazione dell'input, ma di lavorare aggiungendo nuove trasformazioni a una rappresentazione che continua a fluire attraverso tutta la rete. In questo modo l'informazione originaria e le strutture apprese nei layer precedenti restano disponibili e possono essere progressivamente raffinate. Il modello non si limita quindi a combinare localmente probabilità di parole, ma può mantenere e aggiornare rappresentazioni più ricche e coerenti lungo l'intera sequenza. Questo meccanismo rende possibile l'emergere di circuiti interni che integrano contesto, relazioni sintattiche e semantiche e che consentono forme di elaborazione più strutturate rispetto ad un semplice completamento statistico locale. In questo senso, il residual stream

favorisce una composizione progressiva della rappresentazione del significato, riducendo la visione del modello come mero 'pappagallo stocastico', anche se il funzionamento rimane comunque basato su apprendimento statistico dai dati.

5) Quali sono le cause delle allucinazioni dei modelli?

Le allucinazioni degli LLM derivano dal fatto che i modelli generano testo sulla base di probabilità linguistiche e non di una verifica diretta dei fatti. Possono quindi produrre risposte plausibili ma non corrette. A questo si aggiungono limiti e imperfezioni dei dati di addestramento, l'assenza di un vero meccanismo di controllo della verità e la tendenza del modello a rispondere anche quando le informazioni sono incomplete. Per questo gli LLM possono generare contenuti coerenti linguisticamente ma non sempre affidabili dal punto di vista fattuale.

6) Le allucinazioni sono spesso descritte come "un limite strutturale della sintassi priva di semantica". Ritiene che i progressi compiuti ultimamente abbiano cambiato la natura di questo fenomeno o rimane una caratteristica intrinseca del modello?

Sì, l'introduzione dei RAG riduce le allucinazioni degli LLM perché integrano il modello con un sistema di recupero di informazioni da fonti esterne. Prima di generare la risposta, il modello interroga un database o un insieme di documenti e utilizza i contenuti recuperati come contesto aggiuntivo. In questo modo la risposta non si basa solo sulle probabilità apprese durante l'addestramento, ma anche su informazioni reali e verificabili.

7) Mahowald distingue tra competenza linguistica formale e funzionale. Ritiene che i modelli di oggi siano in grado di superare qualsiasi compito di calcolo o di conteggio, come il noto esempio della conta delle 'r' nella parola *strawberry*?

I recenti modelli LLM che usano il residual stream e i layer stratificati che fissano caratteristiche (pattern) lessicali, sintattiche e semantiche sono in grado di superare moltissimi compiti di conteggio o altre funzioni matematiche; in sintesi i transformer mostrano che strutture algoritmiche possono emergere da modelli statistici

subsimbolici. È il modulo attention che raggruppa tutte le lettere 'r' presenti nel prompt.

8) Parlando di grounding, ritiene che l'avvento dei modelli multimodali e agentici stia creando un legame reale tra IA e realtà fisica?

L'avvento dei modelli multimodali e agentici rappresenta certamente un passo importante verso forme più concrete di grounding, perché consente ai sistemi di integrare informazioni provenienti da modalità diverse (testo, immagini, audio, video e, in alcuni casi, dati sensoriali provenienti dal mondo fisico). Questa integrazione permette ai modelli di costruire rappresentazioni più ricche e correlate alla realtà, riducendo in parte il problema dei sistemi puramente linguistici che operano solo su simboli o token. Purtuttavia il legame con la realtà è indiretto e dipendente dalla qualità e dalla varietà dei dati disponibili.

9) Cosa ne pensa dei recenti insuccessi dei modelli e in particolare quelli nelle sperimentazioni di modelli agentici (ad esempio i progetti di Anthropic)? Come si spiegano?

La domanda è profonda perché impatta sull'origine ed architettura degli LLM. In primo luogo, gli LLM sono stati progettati principalmente per generare testo plausibile, non per pianificare azioni complesse e persistenti nel tempo. Quando vengono trasformati in agenti che devono prendere decisioni, interagire con strumenti o ambienti e mantenere obiettivi a lungo termine, emergono limiti legati alla pianificazione, alla memoria e alla gestione coerente delle strategie. Un secondo problema riguarda la fragilità delle catene di azione: gli agenti devono spesso eseguire molte operazioni consecutive (ricerca, interpretazione, scelta dello strumento, verifica del risultato). Un piccolo errore in uno di questi passaggi può propagarsi e compromettere l'intero processo. Inoltre gli LLM non possiedono ancora meccanismi robusti di monitoraggio e autocorrezione del proprio operato.

10) Recenti ricerche, come quelle del MIT, mostrano un potenziale debito cognitivo e un indebolimento metacognitivo legato all'uso intensivo dell'IA. Lei intravede

un rischio reale di deskilling nel lungo periodo o vede l'IA come una forma di estensione della mente?

L'IA può essere interpretata come una forma di estensione della mente, in linea con la teoria della extended mind proposta da Clark e Chalmers. In questa prospettiva, strumenti cognitivi esterni (come libri, calcolatrici o sistemi di IA), non sostituiscono necessariamente le capacità umane, ma possono amplificarle permettendo di affrontare problemi più complessi. Il punto centrale diventa quindi come questi strumenti vengono utilizzati. Se l'IA viene usata in modo passivo e sostitutivo, il rischio di deskilling può aumentare, se invece viene impiegata come supporto alla riflessione, alla verifica e alla generazione di ipotesi, può favorire nuove forme di apprendimento e di collaborazione cognitiva. In sintesi, il rischio esiste specie nel lungo termine ma non è inevitabile: molto dipenderà dal modo in cui l'IA verrà integrata nei processi educativi, professionali e culturali e dalla capacità di mantenere un equilibrio tra automazione e partecipazione attiva del pensiero umano.

11) Nella tesi ipotizzo l'esistenza di una sequenza psicologica e sociale: l'antropomorfizzazione del sistema genera una fiducia eccedente, che facilita una delega decisionale sistematica, portando infine ad un potenziale deskilling strutturale. Dal suo punto di vista, ritiene corretto questo processo? A cosa è riconducibile il deskilling e il depotenziamento metacongitivo? Esistono, a suo avviso, rischi reali di questo debito nel lungo periodo?

La sequenza antropomorfizzazione, fiducia eccedente, delega decisionale, possibile deskilling, è considerata plausibile da diversi studi di psicologia cognitiva e di interazione uomo-macchina. Quando un sistema appare conversazionale, coerente e intenzionale, gli utenti tendono ad attribuirgli competenze cognitive simili a quelle umane. Questo può generare overtrust, cioè una fiducia superiore alle reali capacità del sistema, facilitando la delega sistematica di attività che prima richiedevano impegno cognitivo diretto.

12) In definitiva, c'è un elemento ontologico o funzionale che, secondo lei, continuerà a distinguere in modo netto l'essere umano da qualsiasi intelligenza artificiale futura?

L'essere umano è un soggetto incarnato, situato nel mondo, dotato di coscienza fenomenica, intenzionalità ed esperienza vissuta. Le sue capacità cognitive sono inseparabili dal corpo, dalle emozioni, dalla storia personale e dal contesto sociale. Le attuali forme di intelligenza artificiale, invece, operano come sistemi computazionali che manipolano rappresentazioni e dati senza avere un'esperienza soggettiva del mondo. Come dico sempre un LLM non è in grado di intendere né di volere, perché non ha bisogni reali.

13) Per concludere: come utilizza l'IA nel suo lavoro quotidiano? Di quali strumenti si fida maggiormente e con quale approccio verifica l'affidabilità dei loro *output*?

Utilizzo sempre di più l'approccio del prompt tuning per la richiesta di mostrare le fonti e i ragionamenti fatti passo passo nei vari layer, infine sto cercando di attivare sistematicamente un modulo indipendente (in partica un altro LLM) come certificatore dell'output.

14) Alla luce del suo lavoro e della direzione in cui sta andando la tecnologia, qual è la competenza che ritiene più urgente sviluppare oggi, sia negli individui che nelle organizzazioni e nelle istituzioni? Come distinguerebbe tra un uso dell'IA che estende la mente e un uso che la impoverisce?

Alla luce della direzione attuale dello sviluppo tecnologico dell'IA recente, una delle competenze più urgenti da sviluppare è una forma avanzata di alfabetizzazione critica all'intelligenza artificiale. Non si tratta solo di saper utilizzare gli strumenti, ma di comprendere come funzionano, quali sono i loro limiti, quali bias possono introdurre e in quali contesti è opportuno delegare o mantenere il controllo umano. Inoltre occorre imparare ad interrogarli in modo corretto e quindi una importante competenza futura sarà quella del prompt engineering che traduce richieste dei non specialisti in domande specializzate per gli LLM.

PRECISAZIONE ALLA FINE:

Nei Transformer, il residual stream è il vettore che attraversa tutti i layer e funge da spazio di memoria computazionale condiviso. In ogni layer le operazioni di attenzione e

trasformazione non sostituiscono l'informazione precedente, ma aggiungono nuove componenti al vettore tramite connessioni residue, producendo una sovrapposizione progressiva di strutture informative sempre più ricche.

Questa stratificazione segue una logica emergente abbastanza regolare. Nei layer iniziali il residual stream codifica principalmente informazioni locali legate ai singoli token, identità della parola, posizione nella sequenza, proprietà morfologiche e ortografiche, e qui tendono a emergere i cosiddetti feature detectors, sensibili ad esempio a lettere specifiche, punteggiatura o maiuscole. Nei layer intermedi la rappresentazione si espande oltre il token singolo: compaiono relazioni tra elementi della frase, dipendenze grammaticali, concordanze, coreferenze e l'informazione diventa distribuita su più token contemporaneamente. Nei layer più profondi emergono infine rappresentazioni astratte: il significato complessivo della frase, relazioni logiche, informazioni contestuali e variabili latenti utili per il compito finale.

Il meccanismo di attenzione opera all'interno di questo spazio condiviso con una funzione precisa di selezione e integrazione. Per ogni token, il vettore nel residual stream viene proiettato in una query; tutti gli altri token producono simultaneamente una key e una value. La query viene confrontata con le key tramite prodotto scalare e i punteggi risultanti determinano il peso da assegnare alle informazioni di ciascun token. Il risultato, una combinazione pesata delle value rilevanti, viene poi aggiunto di nuovo al residual stream, arricchendo la rappresentazione complessiva che i layer successivi riceveranno.

In sintesi, il residual stream funge da memoria dinamica dove si accumulano progressivamente rappresentazioni sempre più astratte, mentre l'attenzione agisce come meccanismo di selezione ed integrazione dell'informazione rilevante all'interno di quella memoria distribuita.

Intervista al dott. Agostinello

1) Chi sei, qual è la tua professione e che ruolo hai?

Io sono Mattia e lavoro come Head of Office e Project Director per Meogroup Italia, multinazionale di consulenza manageriale, tecnica e ingegneristica. Sono di base a Milano, dove coordino un team e gestisco progetti in diverse aziende.

2) Quanti anni hai di esperienza nel ruolo e qual è la dimensione della tua organizzazione?

Come Head of Office un anno circa; come responsabile di dipartimento in ambito manageriale, questo è il quarto anno. L'azienda fa parte di un gruppo multinazionale presente in nove paesi, con un fatturato di circa 120 milioni di euro, più di 1100 dipendenti e una presenza di 20 uffici worldwide, tra cui due in Italia. Da Milano gestisco lo sviluppo del Nord Italia.

3) Riguardo all'intelligenza artificiale, qual è l'utilizzo più o meno stimato che ne fai nel tuo lavoro?

Oggi la utilizzo sempre di più. Direi un buon 40%, se devo darti una percentuale. Utilizzo principalmente ChatGPT nella versione pro a pagamento. Di recente, a causa di un infortunio alla mano, ho iniziato a utilizzarla anche per attività più operative come la trascrizione di colloqui e interviste, sia con direttori tecnici che con candidati durante le sessioni di recruiting.

4) Come definiresti il pensiero e il ragionamento?

Ritengo che il pensiero sia la capacità che l'essere umano ha di utilizzare le proprie conoscenze e la propria esperienza per formulare un'idea, un pensiero, un ragionamento. Il ragionamento, invece, parte dal pensiero per sviluppare una serie di output e arrivare a una conclusione, a un risultato finale. Il pensiero può essere teorico e fine a sé stesso, immaginiamoci Aristotele, Einstein, e poi con il ragionamento si arriva a creare attività concrete.

5) E se dovessi definire l'intelligenza artificiale, come la definiresti?

Come uno strumento per facilitare l'essere umano nel fare ragionamenti o attività, trasformando potenzialmente un'idea in qualcosa di più vicino alla realtà. Questi sono però discorsi molto ampi e, a mio avviso, possono risultare sterili se non contestualizzati. L'intelligenza artificiale può essere utilizzata per chiedere quante regioni ci sono in Italia, creare una filastrocca per un bambino, oppure come strumento per identificare aziende target in un certo mercato. L'utilizzo è diverso in base al contesto: è un ottimo strumento che può essere usato per diverse cose a seconda di come viene contestualizzata.

6) In che modo utilizzi l'intelligenza artificiale?

Mi permette di avere un'idea chiara in merito a un determinato argomento. Un ragionamento l'intelligenza artificiale può farlo se viene istruita in un certo modo: se A = B e B = C, allora A = C. Questo è un ragionamento che può fare l'algoritmo, ma non sta ragionando al mio posto. Nel mio lavoro, mi permette di reperire informazioni velocemente. Ad esempio, sapere quali software utilizza una certa azienda e dove posso formarmi su quegli strumenti. Il valore principale è il risparmio di tempo in un contesto lavorativo frenetico.

7) Quali sono i vantaggi concreti che hai riscontrato e quali le criticità?

Il vantaggio principale è la velocità: reperire informazioni, fare analisi, tradurre una riunione dal francese in pochi secondi. Un risparmio di tempo notevole. La criticità più seria, invece, è il rischio della pigrizia. L'uso eccessivo può portare le persone a non verificare più le fonti, a non leggere approfonditamente i documenti, fidandosi ciecamente dello strumento. E questo è pericoloso, perché l'IA non è una scrittura sacra: può proporre informazioni errate se quelle informazioni errate esistono in rete. Se trovasse scritto che la terra è piatta, risponderebbe che la terra è piatta.

8) Hai mai avuto esperienze negative o situazioni in cui l'intelligenza artificiale ti ha fornito informazioni errate?

Durante una riunione ho menzionato un'azienda competitor che ChatGPT mi aveva suggerito come tale, salvo scoprire solo dopo che quella società faceva parte dello

stesso gruppo del mio interlocutore. Una situazione che mi ha insegnato a verificare sempre le informazioni critiche.

9) Come gestisci il processo di verifica degli output?

La verifica è parte fondamentale del processo: la rilettura del testo, l'analisi dei curriculum vitae dopo una prima scrematura automatica, la rilettura delle trascrizioni, che spesso mancano di contesto o non distinguono bene i diversi interlocutori, la revisione dei PowerPoint, il controllo dei bullet point. La regola che mi sono dato è: più la decisione è importante, più la verifica deve essere approfondita. Le uniche attività in cui l'intervento umano è meno necessario sono la creazione di immagini e la stesura di job posting da pubblicare su LinkedIn: lì il margine di rischio è più basso.

10) Nel recruiting, dove trovi i limiti maggiori dell'intelligenza artificiale?

L'IA può aiutare nella scrematura dei curriculum vitae, ma fallisce completamente nel valutare le soft skill, l'attitudine reale di una persona o l'effettiva conoscenza delle lingue durante un colloquio. Una persona che a CV sembra non adatta a un ruolo può avere in realtà l'attitudine perfetta, ma questo lo capisce solo l'uomo, non un algoritmo.

11) Secondo te, perché c'è bisogno di fare questa verifica? Qual è il problema di ChatGPT?

Non è un problema, è semplicemente un dato di fatto: non è umano. Non ha quella capacità di capire realmente la vita. Un documento può essere trascritto da ChatGPT, ma magari utilizza un termine che in una specifica nazione del mondo è considerato un insulto. Non posso dire a ChatGPT "Chiama Filippo e chiedigli se vuole fare questo lavoro" e anche se potesse farlo, lo farebbe con quella staticità, quell'anima inanimata. Potrebbe fare un report perfetto su una conversazione, ma non riuscirebbe a farlo con tutta la parte emotiva ed emozionale che c'è alle spalle.

12) Cosa differenzia un essere umano dall'intelligenza artificiale?

La parte empatica, il pathos. ChatGPT può aiutarmi nella gestione di timesheet, nella redazione di mail, nel perfezionare risposte, ma alla fine è il rapporto umano quello che fa veramente la differenza. Un cliente si affida a me non perché utilizzo ChatGPT, ma

perché io creo un rapporto di fiducia tale per cui il consulente, il direttore tecnico, possono fare affidamento su di me perché sono una persona.

13) Qual è il futuro dell'intelligenza artificiale?

Penso che l'intelligenza artificiale sia destinata a diventare sempre più istruita, non alternativa all'uomo, semplicemente più capace. Come un bambino che legge libri e diventa esperto in un certo ambito. Quello che differenzierà sempre l'essere umano dall'intelligenza artificiale è la parte empatica, il pathos. Conosco un ragazzo che lavora come firmware engineer e utilizza ChatGPT per automatizzare attività che i colleghi impiegano 8 ore a fare: lui le fa in 5, e le altre 3 ore le dedica a studiare e a rilassarsi. Questo è uno scenario plausibile e positivo. Ma i sistemi artificiali non andranno mai a sostituire l'uomo, perché i rapporti personali sono quelli che fanno sì che un cliente voglia lavorare con te e non con un'altra persona. La fiducia e l'empatia che si trasferiscono da essere umano a essere umano sono ben diverse da ciò che un software, programmato da esseri umani, non dimentichiamolo, può offrire.

BIBLIOGRAFIA

A. Behrouz et al., *Titans: Learning to Memorize at Test Time*, arXiv, 2024.

A. Bondarenko et al., *Demonstrating Specification Gaming in Reasoning Models*, arXiv, 2025.

A. Clark e D. J. Chalmers, *The Extended Mind*, Analysis, 1998.

A. Damasio, *L'errore di Cartesio. Emozione, ragione e cervello umano*, Milano, Adelphi, 1995.

A. Giddens, *Le conseguenze della modernità*, Bologna, Il Mulino, 1994.

A. J. Golds et al., *Overreliance on AI: A Review and Agenda for Research on Automation Bias*, International Journal of Human-Computer Studies, 2023.

A. Lawson, *The Illusion of the Illusion of Thinking: A Comment on Shojaee et al. (2025)*, arXiv, 2025.

A. Lisdorf, *What's HIDD'n in the HADD?*, The Cognitive Science of Religion and the Problem of Methodological Naturalism, Journal of Cognition and Culture, 2007.

A. M. Turing, *Computing Machinery and Intelligence*, Mind, 1950.

A. M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 1936.

A. N. Kolmogorov, *Three Approaches to the Quantitative Definition of Information*, Problems of Information Transmission,1965.

A. Newell e H. A. Simon, *Computer Science as Empirical Inquiry: Symbols and Search*, Communications of the ACM, 1976.

A. Ranganathan et al., *AI Doesn't Reduce Work It Intensifies It*, in Harvard Business Review, 2026.

A. Vaswani et al., *Attention Is All You Need*, arXiv, 2017.

B. F. Skinner, *Science and Human Behavior*, New York, Macmillan, 1953.

B. J. Kagan et al., *In vitro neurons learn and exhibit sentience when embodied in a simulated game-world*, in Neuron, 2022.

B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, 2005.

B. Russell, *The Principles of Mathematics*, Cambridge University Press, 1903.

C. O'Neil, *Weapons of Math Destruction How Big Data Increases Inequality and Threatens Democracy*, New York, Crown, 2016.

D. Acemoglu, *The Simple Macroeconomics of AI, Cambridge*, National Bureau of Economic Research, 2024.

D. C. Dennett, *Consciousness Explained*, Boston, Little, Brown and Company, 1991.

D. C. Dennett, *The Intentional Stance*, Cambridge, MIT Press, 1987.

D. E. Rumelhart, J. L. McClelland e PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (volume 2), Cambridge, MIT Press, 1986.

D. Kahneman, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011.

D. Mackenzie, *An Engine Not a Camera How Financial Models Shape Markets*, Cambridge, MIT Press, 2006.

D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books, 1979.

D. Rai et al., *A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models*, arXiv, 2024.

E. Esposito, *Artificial Communication How Algorithms Produce Social Intelligence*, Cambridge, MIT Press, 2022.

E. J. De Visser et al., *Almost human Anthropomorphism increases trust resilience in cognitive agents*, Journal of Experimental Psychology Applied, 2016.

E. M. Bender e A. Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

E. M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021.

E. Nagel e J. R. Newman, *Godel's Proof*, New York University Press, 1958.

F. J. Varela, E. Thompson e E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MIT Press, 1991.

G. J. Chaitin, *A Century of Controversy over the Foundations of Mathematics*, Physica D: Nonlinear Phenomena, 2001.

G. J. Chaitin, *The Unknowable*, Singapore, Springer, 1999.

H. A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, New York, Macmillan, 1947.

H. Gardner, Frames of Mind*: The Theory of Multiple Intelligences*, New York, Basic Books, 1983.

H. L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason*, Harper & Row, New York, 1972.

H. Putnam, *Minds and Machines*, New York University Press, 1960.

I. Yotzov et al., *Firm Data on AI*, National Bureau of Economic Research, 2026.

IBM INSTITUTE FOR BUSINESS VALUE, *CEO Study 5 mindshifts to supercharge business growth*, Armonk, IBM Corporation, 2025.

J. Burrell, *How the Machine Thinks Understanding Opacity in Machine Learning Algorithms*, Big Data & Society, 2016.

J. Dressel, H. Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, Science Advances, 2018

J. Haugeland, *Artificial Intelligence: The Very Idea*, Cambridge, MIT Press, 1985.

J. J. Gibson, *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979.

J. Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv, 2020.

J. Kruger, D. Dunning, *Unskilled and Unaware of It How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*, Journal of Personality and Social Psychology, 1999.

J. L. Barret, *Why Would Anyone Believe in God?*, Walnut Creek, AltaMira Press, 2004

J. McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955.

J. R. Searle, *Minds, Brains and Programs*, Behavioral and Brain Sciences, 1980.

J. W. Meyer, B. Rowan, *Institutionalized Organizations Formal Structure as Myth and Ceremony*, American Journal of Sociology, 1977.

J. Weizenbaum, *ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine*, Communications of the ACM, 1966.

K. Goddard et al., *Automation bias a hidden issue for clinical decision support system use*, Studies in Health Technology and Informatics, 2011.

K. Goddard et al., *Automation bias a systematic review of frequency effect mediators and mitigators*, Journal of the American Medical Informatics Association, 2012.

K. Mahowald et al., *Dissociating language and thought in large language models*, Trends in Cognitive Sciences, 2024.

L. A. Suchman, *Human-Machine Reconfigurations Plans and Situated Actions*, Cambridge, Cambridge University Press, 2007.

L. A. Zadeh, *Fuzzy Sets*, Information and Control, 1965.

L. Doleac, *Algorithmic Risk Assessment in the Hands of Humans*, Institute of Labor Economics, Bonn 2019.

L. F. Menabrea e A. A. Lovelace, *Sketch of the Analytical Engine Invented by Charles Babbage, Esq. with Notes by the Translator*, Scientific Memoirs, 1843.

L. Floridi, *AI as Agency without Intelligence On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis*, Philosophy & Technology, 2025.

L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Milano, Raffaello Cortina Editore, 2017.

L. Floridi, *Semantic Capital: Its Nature, Value, and Curation*, Philosophy & Technology 31, 2018.

L. Floridi, *The Fourth Revolution: How the infosphere is reshaping human reality*, Oxford University Press, 2014.

M. G. Haselton et al., *Resolving the Logic of Emotion Supposition and Superstition in the Evolutionary Arms Race that Shaped the Emotions*, Oxford, Oxford University Press, 2007.

M. Heidegger, *La questione della tecnica*, in Saggi e discorsi, Milano, Mursia, 1976.

M. Heidegger, *The Question Concerning Technology*, The Question Concerning Technology and Other Essays, New York, Harper & Row, 1977.

M. Jakesch et al., *Co-Writing with Opinionated Language Models Affects Users' Views*, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023.

M. Polanyi, *The Tacit Dimension*, New York, Doubleday, 1966.

M. Shanahan, *Talking about Large Language Models*, Communications of the ACM, 2024.

M. Somalvico, *L'Intelligenza Artificiale*, Milano, Rusconi, 1987.

M. Stroud, *Heat Listed*, The Verge, 2021.

N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

N. Carr, *The Shallows What the Internet Is Doing to Our Brains*, New York, W. W. Norton & Company, 2010.

N. Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task*, arXiv, 2025.

N. Rossbach*, Innocent Until Predicted Guilty: How Premature Predictive Policing Can Lead to a Self-Fulfilling Prophecy of Juvenile Delinquency*, 2023.

P. J. Dimaggio, W. W. Powell, *The Iron Cage Revisited Institutional Isomorphism and Collective Rationality in Organizational Fields*, American Sociological Review, 1983.

P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, Advances in Neural Information Processing Systems, 2020.

P. S. Tolbert, L. G. Zucker, *Institutional Sources of Change in the Formal Structure of Organizations. The Diffused Adoption of Civil Service Reforms 1880-1935*, Administrative Science Quarterly, 1983.

P. Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, arXiv, 2025.

R. Descartes, *Discorso sul metodo*, Milano, Bompiani, 2004.

R. E. Petty, J. T. Cacioppo, *Communication and Persuasion Central and Peripheral Routes to Attitude Change*, New York, Springer-Verlag, 1986.

R. K. Merton, *The Self-Fulfilling Prophecy*, The Antioch Review, 1948.

R. L. Buckner et al., *The Brain's Default Network Anatomy Function and Relevance to Disease*, Annals of the New York Academy of Sciences, 2008.

R. Zach, *Hilbert's Program Then and Now*, Philosophy of Logic, 2007.

S. E. Guthrie, *Faces in the Clouds A New Theory of Religion*, New York, Oxford University Press, 1993.

S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Hoboken, Pearson, 2021.

S. T. Fiske, S. E. Taylor, *Social Cognition*, Reading, Addison-Wesley, 1984.

T. Hobbes, *Leviathan*, Londra, Andrew Crooke, 1651.

T. M. Mitchell, *Machine Learning*, New York, McGraw-Hill, 1997.

U. Eco, *Kant e l'ornitorinco*, Milano, Bompiani, 1997.

W. E. Bijker et al., *The Social Construction of Technological Systems*, Cambridge, MIT Press, 1987.

W. I. Thomas, D. S. Thomas, *The Child in America Behavior Problems and Programs*, New York, Alfred A. Knopf, 1928.

X. LI et al., *The effects of the human-like features of generative AI on usage intention and the moderating role of information overload*, Scientific Reports, 2025.

Y. Hanoch et al., *The Dunning-Kruger Effect and Large Language Models A Cautionary Tale*, Computers in Human Behavior Reports, 2024.

Y. LeCun, Y. Bengio e G. Hinton, *Deep Learning*, Nature, 2015.

Yao et al., *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, 2023.

Z. Rucinska, *Affordances in Context*, Phenomenology and the Cognitive Sciences, 2021.

SITOGRAFIA

A. Badman, Gestione del rischio nell'AI, IBM Think; https://www.ibm.com/it-it/think/insights/ai-risk-management ; 25/01/2026

A. Tursi, L'IA applicata al mondo delle organizzazioni: un binomio o una contrapposizione?, Linkedin; linkedin.com/pulse/lia-applicata-al-mondo-delle-organizzazioni-un-binomio-o-una-ijulf?originalSubdomain=it ; 28/02/2026

Anthropic PBC, Project Vend Can Claude run a small shop?, Anthropic Research, 2024; https://www.anthropic.com/research/project-vend-1 ; 08/03/2026

F. Casciabanca, Il nuovo paper di Apple sull'AI svela i limiti dell'Intelligenza Artificiale, Ninja marketing; https://www.ninja.it/paper-di-apple-sull-ai-the-illusion-of-thinking/ 30/12/2025

G. Marcus, A Knockout Blow for LLMs?, Substack, 2025; A knockout blow for LLMs? - by Gary Marcus - Marcus on AI; 30/12/2025

G. Marcus, Scale Is All You Need is dead, Substack, 2025; https://garymarcus.substack.com/p/breaking-news-scale-is-all-you-need ; 30/12/2025

G. Noone, The Case Against Predictive Policing, Tech Monitor; https://www.techmonitor.ai/digital-economy/ai-and-automation/case-against-predictive-policing ; 28/02/2026

H. Hanselman, McKinsey & Company, The State of AI in 2025: Agents, Innovation, and Transformation, 2025; https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai; 08/03/2026

P. Odifreddi, Lezioni sulla logica, videolezione, 2018; https://www.youtube.com/watch?v=CW2xTxLRv_s; 04/01/2026.

P. Odifreddi, Vite da logico: la storia della logica dall'antichità ai giorni nostri, Rai Radio 2; https://www.youtube.com/watch?v=CW2xTxLRv_s; 04/01/2026.

P. Popolizio, L'Impatto Organizzativo dell'Intelligenza Artificiale Generativa: Policy, Formazione e Trasformazione, IA2023, 26 aprile 2025; https://ia2023.it/2025/04/26/impatto-organizzativo-intelligenza-artificiale-generativa-policy-formazione-trasformazione/ ; 28/02/2026

R. Van Riel e R. Van Gulick, Scientific Reduction, Stanford Encyclopedia of Philosophy, Summer Edition, 2025; https://plato.stanford.edu/archives/sum2025/entries/scientific-reduction 29/03/2026;

Redazione, L'intelligenza artificiale e il suo doppio: tra prove impossibili e test truccati, Rivista.ai; https://www.rivista.ai/2025/06/15/lintelligenza-artificiale-e-il-suo-doppio-tra-prove-impossibili-e-test-truccati/; 31/12/2026

U. Stephan, D. Steffen, Artificial intelligence-Bubble or boom?, Deutsche Bank, 2025; https://wealth.db.com/en/insights/investing-insights/asset-class-insights/artificial-intelligence-bubble-or-boom.html ; 08/03/2026

Vocabolario Treccani, s.v, ragionare; https://www.treccani.it/vocabolario/ragionare_res-e1c51c18-e3b1-11eb-94e0-00271042e8d9/ 30/12/2025

Writing Team Winsome Marketing, McKinsey's State of AI Report: 88% Adoption, But Only 6% Are Actually Winning, Winsome; https://winsomemarketing.com/ai-in-marketing/mckinseys-state-of-ai-report-88-adoption-but-only-6-are-actually-winning; 08/03/2026