Ecosistema LLMs

Matteo Mizioli

Università della Tuscia

Dipartimento di Scienze Umanistiche, della Comunicazione e del Turismo

Corso di Laurea in Tecnologie e Culture Digitali — Classe L-20

Relatore: Prof. Giovanni Fiorentino

Candidato: Matteo Mizioli — Matricola 811694

Anno Accademico: 2025/2026

Indice

Elenco delle Tabelle

ECOSISTEMA LLM

Abstract

This work opens with an introduction to artificial intelligence, quickly narrowing its focus to Large Language Models (LLMs). It proceeds with a detailed analysis of how these systems operate, drawing on both foundational scientific literature and the popular expository approaches of experts such as Andrej Karpathy. Adopting a chronological perspective, it then examines recent studies, most notably Google DeepMind's AlphaEvolve (2025), to introduce the concepts of the LLM as an agent, multi-agent environments, and self-improvement. Subsequently, through a humanistic lens, it reflects on the implications of an equilibrium between human and agent, exploring what it means for human beings to engage with entities that are no longer mere prostheses but something more. To frame these dynamics, the discussion turns to established paradigms in the sociology of communication and to comparative studies of the two forms of intelligence, such as Gignac and Szodorai (2024). The approach, technical-explanatory at first and then humanistic, aims to provide a systematic and systemic understanding of the role of these tools, in order to address a central question: can LLMs contribute to a deeper understanding of the meaning and role of the human being in existence? The thesis contends that such systems constitute a privileged instrument for investigating our nature: the more we probe their functioning and potential, the more we come to know about ourselves, especially from an AGI-oriented perspective.

1 INTRODUZIONE

1.1 CONTESTO DISCIPLINARE

Il presente lavoro si posiziona all'intersezione di diversi campi disciplinari. Il primo è quello dell'Informatica, o più precisamente, della Scienza dell'Informazione. Secondo l'Enciclopedia Treccani, l'Informatica è la "scienza che studia l'elaborazione delle informazioni e le sue applicazioni; più precisamente, l'Informatica si occupa della rappresentazione, dell'organizzazione e del trattamento automatico dell'informazione". Il termine deriva dal francese *informatique* (composto di INFORMATion e automatIQUE, "informazione automatica") e fu coniato da Philippe Dreyfus nel 1962.

Si ritiene ancor più precisa la definizione di Scienza dell'Informazione, la quale copre un più ampio spettro disciplinare: essa è la disciplina che indaga le proprietà e il comportamento dell'informazione, le forze che ne governano il flusso e i mezzi per elaborarla al fine di garantirne accessibilità e usabilità ottimali. Si occupa dell'origine, della raccolta, dell'organizzazione, dell'archiviazione, del recupero, dell'interpretazione, della trasmissione, della trasformazione e dell'utilizzo dell'informazione. Include lo studio delle rappresentazioni dell'informazione in sistemi naturali e artificiali, l'uso di codici per la trasmissione efficiente dei messaggi e l'analisi dei dispositivi e delle tecniche di trattamento dell'informazione, come i computer e i loro sistemi di programmazione. È una scienza interdisciplinare con componenti teoriche e applicative (Borko, 1968). In quest'ottica, l'informazione viene intesa come facente parte di più aree, non necessariamente solo come dato, problema o algoritmo (Portinale, 2022), quindi non solo come frutto di una specifica elaborazione tecnico-artificiale o con funzione tecnico-informativa, ma anche come meta-cognitiva, di descrizione cognitivo-comportamentale nei riguardi di esseri organici quali l'uomo e gli animali.

Le altre sfere interdisciplinari di riferimento sono: le Scienze della Comunicazione, l'Artificial Intelligence, le Scienze Cognitive e, indirettamente, da studi sperimentali, la Psicologia, la Metafisica e le Neuroscienze. Di seguito si elencano le rispettive definizioni.

Per le Scienze della Comunicazione, si adotta una definizione sintetica ma puntuale fornita dal CSFO (s.d.): "esse studiano i processi di comunicazione e le modalità di elaborazione delle informazioni e delle conoscenze in vari ambiti." Viene ripresa anche qui l'espressione

"elaborazione delle informazioni", che collega il mondo delle Scienze Comunicative a quello delle Scienze dell'Informazione e, conseguentemente, all'Informatica.

La Psicologia, secondo Treccani, è "la scienza che studia i processi psichici, coscienti e inconsci, cognitivi (percezione, attenzione, memoria, linguaggio, pensiero ecc.) e dinamici (emozioni, motivazioni, personalità ecc.)." Il termine sembra sia stato usato per la prima volta dall'umanista dalmata Marco Marulio nell'opera *Psychologia de ratione animae humanae* (ca. 1511-18). Nonostante uno degli elementi centrali di questo lavoro sia un'architettura artificiale (Vaswani et al., 2017) – qualcosa che parrebbe essere tutto fuorché materia organico-antropomorfa e, quindi, di pertinenza della Psicologia – sarà agevole comprendere, proseguendo con la lettura, che i sistemi di linguaggio, pensiero, memoria e attenzione saranno elementi fondanti della "grande scatola nera" (Moriggi & Pireddu, 2024).

Per Metafisica si intende, secondo Treccani, "la branca della filosofia che, tradizionalmente, mira a individuare la natura ultima e assoluta della realtà al di là delle sue determinazioni relative, oggetto delle scienze particolari." Questa definizione sintetica è sufficiente per contestualizzare la sua chiamata in causa. Una branca che, indubbiamente, fa da sottotesto ad analisi scientifiche, specialmente in quegli spazi vuoti lasciati dal metodo sperimentale. In questo lavoro la sua presenza sarà marginale, un alone, uno sfondo semi-trasparente ma sempre presente in ogni ambiente, poiché i processi privilegiati saranno quelli propri del metodo scientifico. Tuttavia, non si opererà alcun tipo di esclusione, non vi sarà l'intento di prediligere un metodo rispetto a un altro. La logica di base sarà quella di fornire un quadro sistemico e sistematico di questo intersecato ecosistema postmoderno, che potrà essere osservato un pezzo alla volta attraverso lenti differenti.

Si adotterà prima una lente tecnica, che si calerà nelle profondità di queste architetture fantasma (Transformers); poi una lente di transizione che, allentando lo zoom, mostrerà il mondo tecnico collegarsi con il mondo umanistico. Nello specifico, un capitolo ponte relativo allo studio di Google DeepMind (2025) su AlphaEvolve permetterà di introdurre il mondo algoritmico all'interno del più ampio ecosistema comunicativo, nel territorio dei rapporti uomo-agente, applicando una lente che utilizzi come filtro paradigmi delle Scienze della Comunicazione, come la tetrade di McLuhan (Moriggi & Pireddu, 2024). La lente finale, la più ampia, permetterà di mettere insieme tutti i pezzi, fornendo una chiara presa di coscienza di quello che è questo nuovo mondo organico-artificiale, all'interno del quale i fantasmi della *black box* (Moriggi & Pireddu,

2024) possono essere una chiave di lettura per la comprensione della natura umana, ovvero l'intento ultimo di questa tesi.

Per la ricostruzione di questa panoramica sarà necessario considerare ogni parte del percepito, ogni prospettiva, ogni teoria, anche solo per la pura funzione (in ottica di questo lavoro) di essere presente, di essere percepita come alone, un alone tecnico, messo in sottofondo per pura prospettiva metodologico-funzionalista.

Riprendendo il percorso di elencazione e contestualizzazione dei campi disciplinari esaminati, si giunge alle Scienze Cognitive. Secondo Treccani, "con la locuzione scienza cognitiva, dalla fine degli anni Settanta, si è soliti designare l'insieme delle discipline che hanno per oggetto lo studio dei processi cognitivi umani e artificiali." La definizione del campo comprende anche l'ambito artificiale, poiché un ruolo sempre più centrale nella società civile è rivestito dalla cosiddetta intelligenza artificiale (IA). Proprio le ricerche e i risultati dell'IA hanno rappresentato una spinta propulsiva decisiva per la nascita e il consolidamento di un ampio settore interdisciplinare che, in una tassonomia oggi largamente condivisa (cfr. Gardner, 1985), include, oltre all'IA, la psicologia cognitiva, la linguistica e la psicolinguistica, la filosofia della mente e del linguaggio, e le neuroscienze.

In tale cornice, la psicologia cognitiva può essere intesa come l'insieme di ricerche e teorizzazioni sul funzionamento dei processi mentali, fondate sull'idea che la mente operi come un sistema di elaborazione dell'informazione, capace di costruire ed eseguire programmi d'azione finalizzati (Treccani, s.v. "Psicologia cognitiva"). A questa si affianca la linguistica, disciplina che studia il linguaggio e le lingue naturali nelle loro componenti strutturali (fonologia, morfologia, sintassi, semantica, pragmatica) e nei loro aspetti storici, sociali e cognitivi (Treccani, s.v. "Linguistica"). Su questo crinale interdisciplinare si colloca la psicolinguistica, orientata a indagare i processi che sottostanno all'elaborazione del linguaggio nei due versanti della comprensione e della produzione, con particolare attenzione ai meccanismi psicologici e ai sistemi neurali che li rendono possibili (Cacciari, 2022).

Dal punto di vista filosofico, la filosofia della mente esamina la natura dei fenomeni mentali e della coscienza, nonché il loro rapporto con il corpo e il cervello, affrontando temi quali intenzionalità, rappresentazione e stati mentali (Treccani, s.v. "Mente, filosofia della"). Parallelamente, la filosofia del linguaggio studia il linguaggio nei suoi aspetti ontologici, epistemologici e semantico-pragmatici, concentrandosi su nozioni come significato, riferimento, verità e uso

(Treccani, s.v. "Linguaggio, filosofia del"). A integrare questo orizzonte concettuale intervengono le neuroscienze, intese come l'insieme delle discipline che esaminano struttura, funzione, sviluppo, genetica, biochimica, fisiologia e patologia del sistema nervoso, fornendo i correlati biologici dei processi cognitivi e linguistici (Treccani, s.v. "Neuroscienze").

Queste scienze dialogano in modo sempre più stretto e sinergico con l'IA, che costituisce la protagonista del presente elaborato. Alla sua trattazione, al suo statuto teorico-metodologico e alle sue ricadute applicative sarà dedicato un capitolo specifico, così da rendere esplicite le connessioni strutturali tra prospettiva naturale e prospettiva artificiale della cognizione

1.2 DOMANDA CENTRALE

L'ampio raggio disciplinare di questo lavoro potrebbe costituire una giungla di intenti: il focus spazierà da un emisfero di studio all'altro, interconnettendo continue "invasioni di campo", come a voler formare un sistema antropomorfo, organico, complesso. Potremmo iperbolicamente definire questa tesi un essere proto-vivente teoretico-scientifico, un'unione dell'astratto e del concreto che forse materializza meglio un'idea di reale. Dove per "reale" si intende la somma del naturale e dell'artificiale, che a sua volta si divide rispettivamente nelle categorie di uomo, mondo e macchina. Una visione ontologica di reale fornitaci da Somalvico, Amigoni e Schiaffonati (s.d.), che ben si incastrerà con l'intento di questo elaborato.

Tale visione renderà agevole un'analisi del rapporto uomo-macchina, permettendo costanti paragoni e applicando diversi studi, come quelli di Gignac e Szodorai (2024) e Butlin et al. (2023). L'idea è che sia necessario analizzare in questa prospettiva d'insieme per poter apprendere appieno il ruolo, il funzionamento e le potenzialità di questi nuovi sistemi esperti. L'obiettivo dell'elaborato sarà costruire un percorso interdisciplinare, sorretto da studi fondativi, interpretativi e allo stato dell'arte, per arrivare al punto in cui si disporranno tutti i pezzi necessari per avere la chiave di lettura al fine di rispondere a un'unica e complessa domanda: Possono i Large Language Models (la principale "scatola nera" che si analizzerà più avanti) aiutare a conoscere meglio la natura dell'essere umano?

Lo studio sosterrà che, sì, essi sono e saranno un grande strumento a tal proposito, e il percorso che verrà illustrato mirerà, senza determinismi, a rafforzare questo esito.

2 FONDAMENTI STORICI DELL'IA E SUA EVOLUZIONE

L'intelligenza artificiale sta rivoluzionando l'ecosistema lavorativo contemporaneo. Nazioni, aziende e multinazionali stanno investendo ingenti capitali nella ricerca e nello sviluppo di questo settore. Molte sono e saranno le posizioni lavorative che cambieranno, dovendo adattarsi all'integrazione professionale di nuove modalità operative che utilizzino sistemi all'avanguardia di IA. Molti i lavori che potrebbero scomparire, con maggior pericolo per le mansioni automatizzabili, che richiedono bassa interazione umana e poca capacità di astrazione.

Secondo un rapporto del Censis di marzo 2025:

> *"Le imprese italiane che programmano di investire in beni e servizi legati all'intelligenza artificiale nel biennio 2025-2026 in Italia rappresentano il 19,5%. Per quel che riguarda i servizi non finanziari, più della metà, ovvero circa il 55% delle imprese informatiche, dichiara di avere intenzione di investire sull'IA nel biennio considerato, il dato con la più alta percentuale di imprese."*
>
> *"Tra un quinto e un quarto dei lavoratori utilizza strumenti di IA sul luogo di lavoro. Più nel dettaglio, il 23,3% utilizza IA per la scrittura di e-mail, il 24,6% per messaggi, il 25% per la stesura di rapporti e il 18,5% per la creazione di curriculum. I numeri salgono al diminuire dell'età, come dimostra il 35,8% tra i 18-34 anni che utilizza IA per la stesura di rapporti contro il 23,5% tra chi ha più di 45 anni, o il 28,8% dei più giovani che la utilizzano per la scrittura di e-mail, a fronte di un 21,9% della fascia di popolazione che ha più di 45 anni."*
>
> *"Secondo una recente pubblicazione della Banca d'Italia, che pone gli occupati divisi per classe professionale su una scala di esposizione all'IA, sui circa 22 milioni di lavoratori attivi nel 2022 in Italia, circa 15 milioni ricadono nella fascia a media-alta esposizione alla complementarità o*

sostituzione. Di questi 15 milioni, circa 9 ricadono nella fascia esposta alla complementarità con l'IA, mentre 6 milioni circa sono mediamente o altamente esposti alla sostituzione. Più nel dettaglio, il numero di lavoratori altamente esposti alla sostituzione si quantifica intorno ai 4,75 milioni, mentre i lavoratori altamente esposti alla compenetrazione delle intelligenze artificiali nelle loro mansioni si attestano intorno ai 4 milioni. In altri termini, circa il 22% della forza lavoro potrebbe, in linea teorica, essere sostituita dall'IA e il 18% circa dei lavoratori potrebbe vedere un ingresso dell'IA in una funzione altamente complementare alle loro mansioni."

Questi dati sottolineano l'importanza di comprendere la natura, gli obiettivi e l'evoluzione storica dell'IA, temi che verranno affrontati nelle sezioni seguenti.

2.1 DEFINIZIONE DI INTELLIGENZA ARTIFICIALE

L'intelligenza artificiale (IA) è un'area interdisciplinare dell'informatica in continua espansione che studia come progettare sistemi capaci di svolgere compiti che, se svolti da un essere umano, sarebbero considerati frutto di "intelligenza". La definizione classica formulata da John McCarthy, "the science and engineering of making intelligent machines", evidenzia la doppia natura dell'IA, allo stesso tempo teorica e ingegneristica: l'IA è disciplina che produce conoscenza (modelli, teorie) e tecnica che realizza artefatti funzionanti (Sbardella, 2025).

Definizioni lessicografiche e divulgative insistono invece sulla possibilità di "riprodurre i processi mentali più complessi mediante l'uso di un computer" (Treccani, citato in Traversari, 2025, p. 4), mentre formulazioni più pragmatiche la definiscono come il complesso di metodi e tecnologie che permettono a macchine e programmi di svolgere compiti tipicamente umani, sfruttando grandi quantità di dati e tecniche di apprendimento (Sbardella,2025) .

Di seguito vengono riportate alcune definizioni chiave:

- *"IA, Disciplina che studia se e in che modo si possano riprodurre i processi mentali più complessi mediante l'uso di un computer." (Traversari,2025)*

- *"L'IA è quella disciplina, appartenente all'informatica, che studia i fondamenti teorici, le metodologie e le tecniche che permettono di progettare sistemi capaci di fornire all'elaboratore elettronico delle prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana." (Somalvico et al., s.d.)*

- *"A field of computing which focuses primarily on the transmission of anthropomorphic intelligence and thinking into machines that can assist humans in many ways." (Sbardella, 2025)*

- *"Artificial Intelligence is a branch of science and technology that creates intelligent machines and computer programs to perform various tasks which require human intelligence. It is a system that mimics various functions which a human can do. AI uses external data like big data to achieve excellent performance for the given tasks." (Sbardella, 2025)*

Queste posizioni non sono in contraddizione, ma rappresentano sfaccettature differenti: alcune enfatizzano il nucleo concettuale (McCarthy, Somalvico), intendendo l'IA come scienza e ingegneria; altre si soffermano sull'attenzione metodologica (approcci simbolici vs. subsimbolici, apprendimento statistico), interrogandosi su quali strumenti si utilizzino per ottenere "intelligenza"; altre ancora sviluppano la dimensione applicativa (Sbardella, Traversari), concependo l'IA come sistema che raggiunge performance comparabili a quelle umane in compiti specifici.

Nella pratica contemporanea è utile distinguere tra livelli e declinazioni dell'IA:

Artificial Narrow Intelligence (ANI): Detta anche IA debole, progettata per svolgere compiti specifici con un alto grado di efficienza. Gli attuali sistemi di IA appartengono tutti a questa categoria.

Artificial General Intelligence (AGI): Detta anche IA forte, una forma teorica di IA capace di comprendere, apprendere e applicare conoscenze su una vasta gamma di compiti, analogamente a un essere umano. A differenza dell'IA "ristretta", che è specializzata in compiti specifici, un

sistema AGI potrebbe risolvere problemi complessi in contesti nuovi e sconosciuti, generalizzando le conoscenze e mostrando creatività e intuizione. L'AGI è attualmente un obiettivo di ricerca teorica e non è stata ancora realizzata.

Artificial Super Intelligence (ASI): Un sistema ipotetico di intelligenza artificiale basato su software con una portata intellettuale che va oltre l'intelligenza umana in tutti i domini.

Per sotto-domini concreti, come i modelli linguistici, si adottano definizioni operative: ad esempio, un Large Language Model (LLM) – che verrà approfondito nel prossimo capitolo – è spesso definito in letteratura come un modello in grado di generare testo, addestrato su grandi quantità di token (dell'ordine di miliardi) e progettato per essere adattabile a diversi compiti tramite tecniche di transfer learning (Traversari, 2025).

2.2 OBIETTIVO DELL'INTELLIGENZA ARTIFICIALE

Come sottolineato da Somalvico, Amigoni e Schiaffonati (s.d.):

> *"Obiettivo di questa disciplina non è quello di replicare o simulare l'intelligenza umana, obiettivo la cui ponibilità è, per taluni scienziati, addirittura non ammissibile per motivi epistemologici laddove venga postulata una distinzione tra la realtà e la conoscenza della realtà. Al contrario, l'IA ha come obiettivo quello di riprodurre o emulare l'intelligenza umana, in quanto non vi è alcun motivo a priori che impedisca che talune (ma non tutte!) prestazioni dell'intelligenza umana (per esempio la capacità di risolvere problemi risolubili con procedimenti inferenziali) possano anche essere fornite da una macchina."*

In questo studio si pone particolare attenzione all'utilizzo della parola "intelligenza", poiché essa porta con sé una serie di attributi impliciti che si presuppone l'oggetto a cui è indirizzata detenga. Possono essere diverse le chiavi di lettura del lemma in questione; certo è che la sua attribuzione comporta un'antropomorfizzazione dell'oggetto e, in virtù di ciò, un'antropomorfizzazione di tutte le sue manifestazioni.

Si ritiene che non sia un'affermazione semplice sostenere che, ad esempio, un LLM sia un sistema intelligente. Si preferisce essere cauti e parlare piuttosto di *sistemi esperti*, andando talvolta in contrasto con l'utilizzo del termine che ne fanno molti degli studi citati. Verrà dedicato un capitolo specifico che approfondirà questi temi, con il sostegno di studi fondativi quali Butlin et al. (2023) e Gignac e Szodorai (2024).

2.3 EVOLUZIONE STORICA DELL'AI

La ricerca sull'intelligenza artificiale si sviluppa su due linee interconnesse: un immaginario culturale antico che concepisce esseri artificiosi e la costruzione progressiva di strumenti formali e tecnologici per rendere quell'immaginario operativo. Miti come il Golem o le figure di Prometeo ed Efesto attestano una tensione culturale verso la costruzione dell'artificiale che attraversa i secoli; pur non essendo spiegazioni tecniche, queste narrazioni costituiscono il riferimento culturale che accompagna la nascita delle macchine razionali (Traversari, 2025; Somalvico et al., s.d.).

2.3.1 Dalle Macchine Calcolatrici ai Fondamenti Teorici

Dalle macchine calcolatrici alla formalizzazione del ragionamento, la tradizione formalistica incarnata da figure come Leibniz e Babbage ha posto le basi concettuali per trattare il pensiero come manipolazione simbolica suscettibile di automazione. L'idea di un *calculus ratiocinator* e i progetti di macchine analitiche anticiparono concetti oggi centrali quali programma, memoria e procedura. Nel XX secolo, la formalizzazione della computabilità e la definizione della macchina di Turing fornirono la cornice teorica che legittimava l'ipotesi di procedure meccaniche per il ragionamento umano, aprendo la strada a sviluppi successivi (Somalvico et al., s.d.).

2.3.2 Anni Quaranta-Cinquanta: Cibernetica e la Nascita dell'IA

Negli anni Quaranta, la cibernetica e i primi modelli di reti neurali cambiarono il panorama teorico: McCulloch e Pitts proposero modelli elementari di neuroni artificiali (1943) e Donald Hebb formulò principi di apprendimento che influenzarono la ricerca futura. Questi contributi, insieme

alla nozione di calcolabilità, costituirono le fondamenta sia per approcci simbolici sia per approcci subsimbolici.

Il Dartmouth Workshop del 1956 rappresenta una tappa cruciale: la formalizzazione dell'ambizione di rendere calcolabili aspetti dell'apprendimento e del linguaggio che segnò l'avvio dell'IA come disciplina riconosciuta. Nei decenni successivi, l'approccio dominante fu quello simbolico: la conoscenza venne rappresentata tramite regole, logiche e sistemi deduttivi.

2.3.3 L'Era Simbolica e i Sistemi Esperti

Sistemi come il Logic Theorist e il General Problem Solver mostrarono le potenzialità dei metodi simbolici, ma anche i loro limiti pratici nel trattare la complessità del mondo reale. Dall'esperienza simbolica derivarono i sistemi esperti (DENDRAL, MYCIN), che codificavano conoscenza specialistica per supportare decisioni in domini ben delimitati. Questi sistemi dimostrarono che l'IA poteva offrire strumenti utili in applicazioni concrete, ma evidenziarono anche un problema centrale: quando la conoscenza è tacita, incerta o troppo complessa per essere esplicata in regole formali, i sistemi simbolici incontrano forti limiti (Sbardella, 2025).

2.3.4 Gli "AI Winters" e la Crisi del Settore

L'entusiasmo iniziale si scontrò con difficoltà computazionali, costi elevati e aspettative irrealistiche. I periodi di forte contrazione di interesse e finanziamenti, noti come *"AI winters"*, segnarono l'andamento ciclico del campo. Tuttavia, questi momenti ebbero anche una funzione selettiva: portarono la comunità a rivedere assunti troppo ottimistici, a riconoscere i limiti degli approcci esistenti e ad aprirsi a soluzioni ibride e più robuste (Sbardella, 2025).

2.3.5 Fine XX Secolo: Approcci Basati sui Dati

A partire dalla fine del XX secolo, l'attenzione si spostò verso approcci basati sui dati. L'apprendimento statistico permise di estrarre pattern da grandi dataset senza la codifica manuale di regole: algoritmi come SVM (Support Vector Machines), alberi decisionali e metodi ensemble – che verranno spiegati successivamente – migliorarono robustezza e applicabilità. La disponibilità crescente di dati digitali e potenza di calcolo resero praticabili soluzioni che prima erano teoriche, segnando il passaggio da sistemi basati su conoscenza esplicita a metodi fondati su modelli probabilistici e statistici.

Nel campo del linguaggio, la costruzione di corpora – ad esempio il Brown Corpus, un corpus linguistico elettronico pionieristico che rappresenta il primo campione bilanciato di inglese americano scritto, pubblicato nel 1964 da Winthrop Nelson Francis e Henry Kučera – forgiò basi empiriche che favorirono lo sviluppo della linguistica computazionale e l'adozione di metodi probabilistici (Traversari, 2025).

2.3.6 Anni 2000: Embedding e Rappresentazioni Vettoriali

Negli anni 2000, la combinazione di grandi corpora, architetture neurali e potenza di calcolo portò alla rappresentazione continua delle unità linguistiche: gli *embedding* (ad esempio, word2vec di Mikolov et al.; modelli neurali di linguaggio di Bengio et al.) reintrodussero l'idea di vicinanza semantica nello spazio vettoriale, ovvero la rappresentazione di parole, frasi o documenti come vettori numerici in uno spazio multidimensionale. L'idea centrale è che entità linguistiche con significati simili siano posizionate più vicine tra loro in questo spazio vettoriale rispetto a entità con significati diversi. Verrà approfondita concretamente più avanti l'applicazione e l'estensione di questo concetto affrontando gli algoritmi utilizzati nel campo e l'architettura Transformer. Questa prospettiva permise di catturare relazioni analogiche e contesti d'uso con un'efficacia prima impensabile per i modelli simbolici tradizionali (Traversari, 2025).

2.3.7 Il Rilancio del Deep Learning e l'Architettura Transformer

Il rilancio delle reti neurali profonde, reso pratico dall'algoritmo di retropropagazione (*backpropagation*), dall'uso massiccio di GPU e da dataset su larga scala, produsse miglioramenti rivoluzionari in visione artificiale, riconoscimento vocale e Natural Language Processing (NLP). Il 2012 è spesso indicato come punto di svolta: modelli profondi superarono benchmark consolidati e avviarono un ciclo di ricerca, industrializzazione e investimenti che trasformò l'IA in una tecnologia pervasiva (Sbardella, 2025).

La pubblicazione "Attention Is All You Need" (Vaswani et al., 2017) introdusse l'architettura Transformer, basata su meccanismi di attenzione che consentono l'elaborazione parallela e una migliore cattura delle dipendenze a lungo raggio. I Transformer resero possibile il pre-addestramento su enormi dataset e il trasferimento a molteplici compiti tramite *fine-tuning* o *prompting*. Da questa architettura derivano due filoni principali: modelli encoder (ad esempio BERT) orientati alla comprensione e modelli decoder/autoregressivi (ad esempio GPT) orientati

alla generazione. L'evoluzione ha portato ai grandi modelli di linguaggio (LLM) e alla *generative AI* che oggi alimentano chatbot, strumenti di sintesi testuale e applicazioni creative (Traversari, 2025; Sbardella, 2025).

3 FONDAMENTI, TRANSFORMERS, E RETI NEURALI

3.1 MACHINE LEARNING E DEEP LEARNING

Prima di approfondire l'architettura Transformer, alla base dei moderni LLM, è necessario fornire un quadro introduttivo ma puntuale sulle reti neurali (delle quali i Transformer rappresentano uno specifico modello) nell'ambito del Deep Learning. Non esiste solo questa architettura: ve ne sono diverse, con caratteristiche e funzioni differenti. Partiremo dalle prime progettate, spiegandone le meccaniche di base, fino ad arrivare allo stato dell'arte. Sarà necessario chiarire anche la struttura che definisce i metodi adoperati in combinazione con tali architetture, quali: Machine Learning, Deep Learning, e, rispettivamente a questi due (ulteriori diramazioni), Supervised Learning, Unsupervised Learning e Reinforcement Learning (gli algoritmi di SL, UL, RL fanno parte del campo del ML o DL a seconda di specifiche caratteristiche che presenteremo in seguito). Ci serviremo del lavoro di Sbardella (2025).

È necessario disporre di un quadro concettuale che non si limiti a definire, ma che giustifichi il ruolo e la scelta di ciascuna famiglia di metodi dell'ingegneria dell'intelligenza artificiale. La distinzione tra Machine Learning (ML) e Deep Learning (DL), così come la separazione per modalità di apprendimento (Supervised Learning, Unsupervised Learning, Reinforcement Learning), non sono meri esercizi tassonomici: riflettono scelte progettuali che nascono da vincoli epistemici (che cosa si può imparare dai dati), computazionali (quanta complessità è sostenibile) e applicativi (quali proprietà desideriamo, ad esempio robustezza, interpretabilità, generalizzazione). Verrà argomentato il perché di ogni funzione/metodo e affiancheremo, per ciascuno, un esempio pratico che ne chiarisca l'utilizzo effettivo (Sbardella, 2025).

La distinzione principale risiede nella complessità della rappresentazione appresa. Nel ML "classico", gli algoritmi (es. regressione lineare, SVM, alberi decisionali) operano spesso su feature ingegnerizzate a priori; il modello apprende una mappatura relativamente semplice tra feature e target. "Feature ingegnerizzate a priori" significa che le caratteristiche dei dati (feature) vengono scelte e preparate prima di addestrare il modello. Le feature sono gli attributi misurabili che descrivono ogni esempio nei dati: per una casa, ad esempio, possono essere la superficie, il numero di stanze, la distanza dal centro e l'anno di costruzione. Prepararle "a priori" vuol dire combinarle o trasformarle in modo utile (per esempio creare il rapporto superficie/stanze, applicare un logaritmo a una variabile molto dispersa, estrarre conteggi di parole da un testo o creare indicatori sì/no), basandosi sulla conoscenza del problema.

Per mappatura si intende la funzione che il modello impara per collegare le feature all'obiettivo da prevedere (il target). In pratica, data una descrizione x (le feature), il modello apprende una regola f tale che f(x) sia il più possibile vicino al valore reale y (per esempio il prezzo di una casa, la classe spam/non spam, oppure una categoria tra più possibili). Dire che "il modello apprende una mappatura relativamente semplice tra feature e target" significa che, se le feature sono state progettate bene, non serve un modello molto complesso per ottenere buoni risultati: bastano spesso modelli lineari o poco profondi. In questo approccio, gran parte del lavoro "intelligente" avviene prima, nella preparazione delle feature.

Al contrario, nel Deep Learning è il modello stesso a imparare automaticamente le rappresentazioni utili a partire da dati grezzi (immagini, testo, audio), e quindi la mappatura tra input e target è più complessa e articolata. Il DL si caratterizza per la capacità di apprendere rappresentazioni gerarchiche tramite reti neurali profonde: più strati consentono trasformazioni non lineari successive che costruiscono astrazioni sempre più complesse dei dati (Sbardella, 2025).

Molti dati naturali (immagini, audio, testo) contengono strutture gerarchiche: bordi → forme → oggetti nelle immagini; fonemi → parole → frasi nel linguaggio. Modelli profondi catturano queste gerarchie automaticamente, riducendo la necessità di feature engineering (quindi il lavoro manuale di dover creare a mano le relazioni di significato tra i dati).

Esempio pratico: per il rilevamento di frodi bancarie, un modello ML (Random Forest) con feature costruite sull'attività transazionale può essere sufficiente e più interpretabile; per la

generazione automatica di testo o il riconoscimento di oggetti in immagini complesse, i modelli DL (CNN, Transformer) offrono prestazioni nettamente superiori (Sbardella, 2025).

Nota operativa: nella letteratura di riferimento e in Sbardella (2025) si usa una soglia pratica per distinguere le reti "classiche" da DL: reti neurali con tre layer o meno rientrano spesso nel contesto ML (le reti neurali hanno una struttura a livelli, che verrà introdotta meglio più avanti), mentre reti con più di tre layer sono considerate Deep Learning (Sbardella, 2025). Questa soglia non è dogma teorico ma una convenzione utile per orientarsi.

3.2 SL,UL,RL

Approfondendo la tassonomia proposta in Sbardella (2025), analizziamo la suddivisione per modalità di apprendimento, Supervised Learning (**SL**), Unsupervised Learning (**UL**) e Reinforcement Learning (**RL**), non come etichette puramente descrittive, ma come scelte progettuali che orientano l'intera pipeline di sviluppo di un sistema di IA. SL, UL e RL differiscono principalmente per il tipo di segnale d'apprendimento e per l'obiettivo che perseguono.

Supervised Learning (SL) :

Il Supervised Learning è l'approccio in cui il modello viene addestrato su dati già "etichettati": per ogni esempio di input è disponibile l'output corretto e l'algoritmo impara a stimare la funzione che mappa input → output (minimizzando una funzione di errore). In pratica si usa una procedura iterativa di ottimizzazione (es. discesa del gradiente e backpropagation per le reti neurali) per ridurre la perdita (es. cross-entropy per classificazione, MSE per la regressione).

Gli aspetti pratici fondamentali sono: divisione dei dati in training/validation/test, la scelta della funzione di loss e dell'ottimizzatore, regolarizzazione per evitare overfitting e monitoraggio con metriche appropriate (accuracy, precision/recall/F1/AUC per classificazione; RMSE/MAE per la regressione). SL è la scelta naturale quando le etichette sono affidabili e rappresentative del problema reale.

In parole semplici: nel Supervised Learning si mostrano al modello esempi già risolti (input +

risposta corretta) e gli si chiede di "imparare" dalle differenze tra le sue previsioni e le risposte vere (quelle che gli sono state date). Una volta addestrato, il modello dovrebbe essere capace di fornire la risposta corretta anche su nuovi esempi mai visti.

Unsupervised Learning (UL):

L'Unsupervised Learning non dispone di etichette: l'obiettivo è scoprire strutture nascoste nei dati, come raggruppamenti (clustering), relazioni o rappresentazioni compresse (riduzioni dimensionali). Gli algoritmi di UL devono trovare autonomamente pattern coerenti e significativi.

In parole semplici: nell'Unsupervised Learning si danno al modello solo i dati, senza dirgli quali siano le risposte corrette. Il modello deve "scoprire da solo" come sono organizzati i dati. Ad esempio, può trovare gruppi di esempi simili (clustering), ridurre la complessità dei dati mantenendo l'informazione importante (riduzione dimensionale), oppure imparare pattern nascosti.

Tra i metodi più diffusi troviamo: K-means, DBSCAN, PCA, t-SNE, autoencoder. L'UL è utile quando non abbiamo etichette o vogliamo esplorare i dati prima di costruire un modello supervisionato.

Reinforcement Learning (RL):

Il Reinforcement Learning si basa su un'interazione continua con un ambiente: l'agente esegue azioni, osserva i risultati e riceve ricompense o penalità. L'obiettivo è apprendere una policy (strategia) che massimizzi il ritorno cumulativo nel tempo.

I componenti chiave sono: **stato** (configurazione del sistema), **azione** (scelta dell'agente), **ricompensa** (feedback numerico), **policy** (strategia che mappa stati in azioni), **funzione di valore** (stima del ritorno atteso da uno stato).

Le azioni hanno effetti che si propagano nel tempo: una decisione corrente può influenzare gli stati futuri e le ricompense successive. Ne consegue che l'agente deve pianificare e non limitarsi a massimizzare il guadagno immediato; per questo si parla di decisioni sequenziali.

Un esempio applicativo può essere nella robotica: apprendimento di comportamenti motori (es. locomozione) o stabilizzazione di sistemi (es. droni) attraverso tentativi ed errori. Un altro esempio lo abbiamo con i giochi: agenti che migliorano progressivamente le proprie prestazioni in scacchi, Go o videogiochi. La qualità di un'azione è misurata tramite la ricompensa, un valore numerico restituito dall'ambiente dopo l'esecuzione. Ricompense elevate indicano azioni utili, mentre ricompense basse o negative segnalano azioni dannose. L'agente sfrutta tali segnali per aggiornare e migliorare la propria policy (strategia).

Fa uso di Reinforcement Learning anche l'amatissimo LLM GPT: dopo l'addestramento su un dataset, vengono applicate tecniche di questo tipo per migliorare ed educare il modello ad agire come un assistente virtuale, con esempi adeguati che gli permettono di sviluppare una policy in grado di farlo migliorare via via con le risposte. Ma si vedrà questo nello specifico più avanti.

3.3 RETI NEURALI

Prima di procedere con la spiegazione nel dettaglio di tutti gli algoritmi menzionati fino ad ora (SVM, ensemble, ecc.), dobbiamo fornire le nozioni necessarie ad inquadrare maggiormente le reti neurali, affinché venga posizionato nella tassonomia anche questo fondamento. Più avanti forniremo una mappa concettuale realizzata da Sbardella (2025) riassumente la tassonomia.

Nella tesi di Sbardella è ben evidenziato come le prime intuizioni sulle reti neurali affondino le radici già nella metà del XX secolo: concetti preliminari e formalizzazioni matematiche hanno preparato il terreno per i primi modelli ispirati al cervello umano. Le prime architetture semplici, come i neuroni formali e il perceptron, rappresentano un tentativo diretto di tradurre il comportamento di un singolo neurone biologico in una regola computazionale.

Immaginiamo un esempio basilare: un neurone artificiale, ispirato a quello biologico, riceve diversi "input" (dati in ingresso), ognuno dei quali è associato a un "peso". Questi pesi determinano l'importanza di ciascun input. Il neurone somma i prodotti degli input per i rispettivi pesi e, se questa somma supera una certa "soglia", produce un "output" (un risultato). Ad esempio, se vogliamo che il neurone riconosca un semplice pattern, potremmo avere due input: uno che indica la presenza di una linea verticale e uno per una linea orizzontale. Se entrambi gli input sono presenti e i loro pesi combinati superano la soglia, il neurone potrebbe attivarsi, indicando il riconoscimento del pattern.

Questo meccanismo semplice, ma potente, ha dimostrato che l'apprendimento automatico potesse emergere dalla semplice combinazione di unità elementari; tuttavia, al tempo erano

vincolati da limiti teorici e pratici (ad es. problemi non linearmente separabili) che portarono, per un periodo, a una diffusa disillusione nella comunità, i famosi "AI winters".

Un passaggio chiave è lo sviluppo dell'algoritmo di Backpropagation (esplicitato nel lavoro che ha rianimato la ricerca sulle reti negli anni '80). Backpropagation permise di addestrare reti multilivello (Multi-Layer Perceptron, MLP) propagando l'errore dalla fine all'inizio della rete e aggiornando i pesi con metodi di discesa del gradiente.

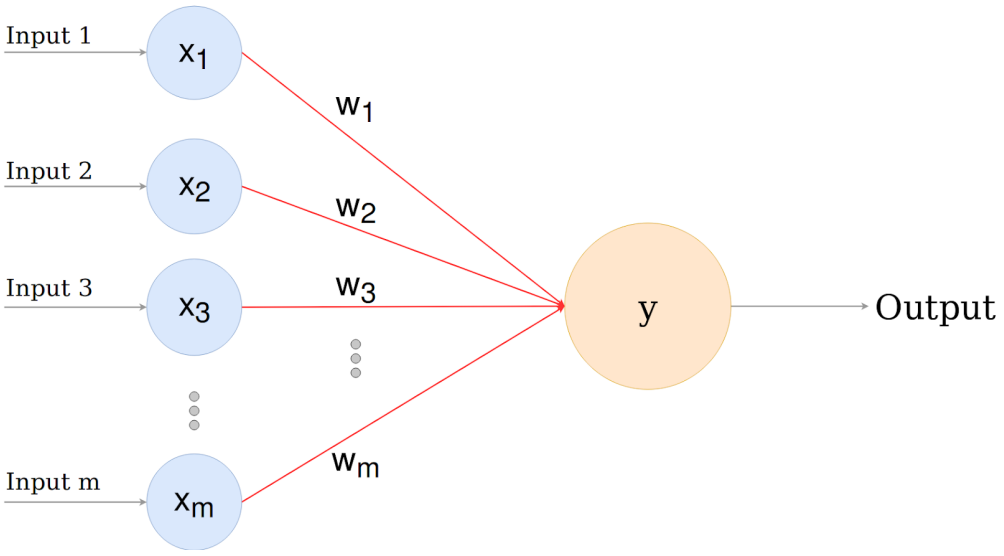


Fig.1 (Modello Perceptron).

In parole semplici: bisogna immaginare una rete neurale come una catena di trasformazioni: prende un input (ad esempio un'immagine) e produce un output (ad esempio "gatto" o "non gatto"). Dopo ogni previsione, confrontiamo l'output con la risposta corretta e calcoliamo un errore: quanto ci siamo allontanati dal valore desiderato. Backpropagation è il metodo che indica alla rete "dove" e "quanto" ha sbagliato in ciascun punto della catena, così da potersi correggere.

Passo 1: avanti. L'input attraversa i livelli della rete fino all'output. Si confronta l'output con l'etichetta corretta e calcoliamo l'errore.

Passo 2: indietro. Si propaga l'errore all'indietro, dall'uscita verso l'ingresso, stimando per ogni collegamento (peso) il suo contributo all'errore complessivo.

Passo 3: aggiornamento. Si utilizza la discesa del gradiente per modificare leggermente ogni peso nella direzione che riduce l'errore.

Prima non esisteva un modo pratico per "insegnare" a reti con più livelli. Backpropagation ha fornito un procedimento efficiente e sistematico per calcolare l'impatto dell'errore su ciascuna parte della rete e aggiornare i pesi di conseguenza. In questo modo è diventato possibile addestrare reti multilivello (MLP) e far loro apprendere compiti molto più complessi. Questa innovazione rese possibile apprendere rappresentazioni non lineari complesse e aprì la strada a reti con più strati, gettando le basi per il passaggio successivo verso quelle che oggi chiamiamo Deep Neural Networks.

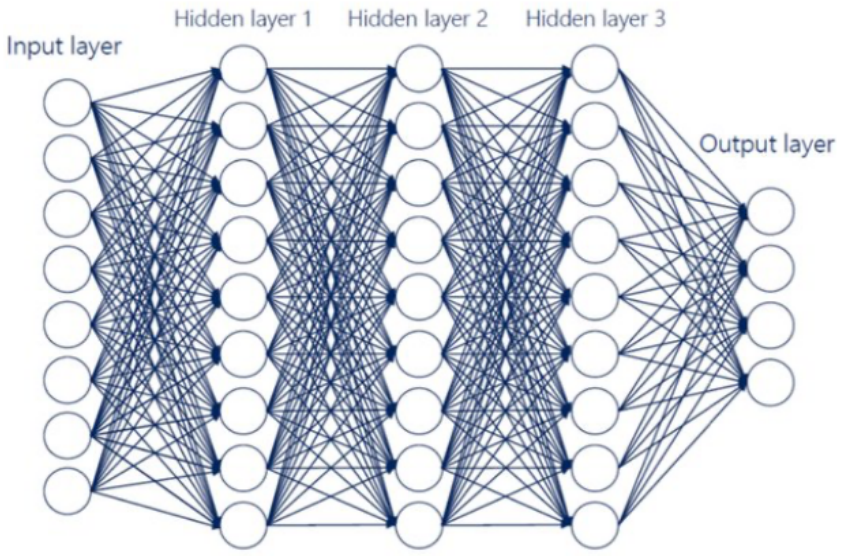


Fig 2. Esempio di Deep Neural Network, fonte: Sbardella, L. (2025).

Nel Deep Learning le reti sono state progettate per adattarsi al tipo di dato che si vuole analizzare.

Convolutional Neural Networks (CNN):

Se si lavora con immagini, le Convolutional Neural Networks (CNN) sono state decisive. Funzionano guardando piccole parti dell'immagine alla volta con degli "occhiali" chiamati filtri, e riutilizzando gli stessi filtri in punti diversi. Questo riduce i calcoli e aiuta a riconoscere via via forme sempre più complesse. Partendo da LeNet e poi con reti più grandi come AlexNet, ResNet ed EfficientNet, le CNN hanno reso molto più efficace il riconoscimento di oggetti, la segmentazione e il rilevamento in immagini e video. Un aiuto fondamentale è arrivato anche dalle

GPU, che hanno accelerato enormemente l'addestramento.

Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU):

Quando i dati sono sequenze nel tempo, come frasi, audio o serie di misure, serve "ricordare" ciò che è successo prima. Le Reti Ricorrenti (RNN) sono state la prima scelta, ma facevano fatica a mantenere informazioni per periodi lunghi. Per questo sono nate le LSTM e le GRU: grazie a dei "cancelli" interni decidono cosa tenere in memoria e cosa dimenticare. Così riescono a capire meglio contesti lunghi, risultando utili in traduzione automatica, riconoscimento vocale e previsioni su serie temporali. Le versioni bidirezionali, quando possibile, permettono di leggere una sequenza sia in avanti sia all'indietro, migliorando la comprensione.

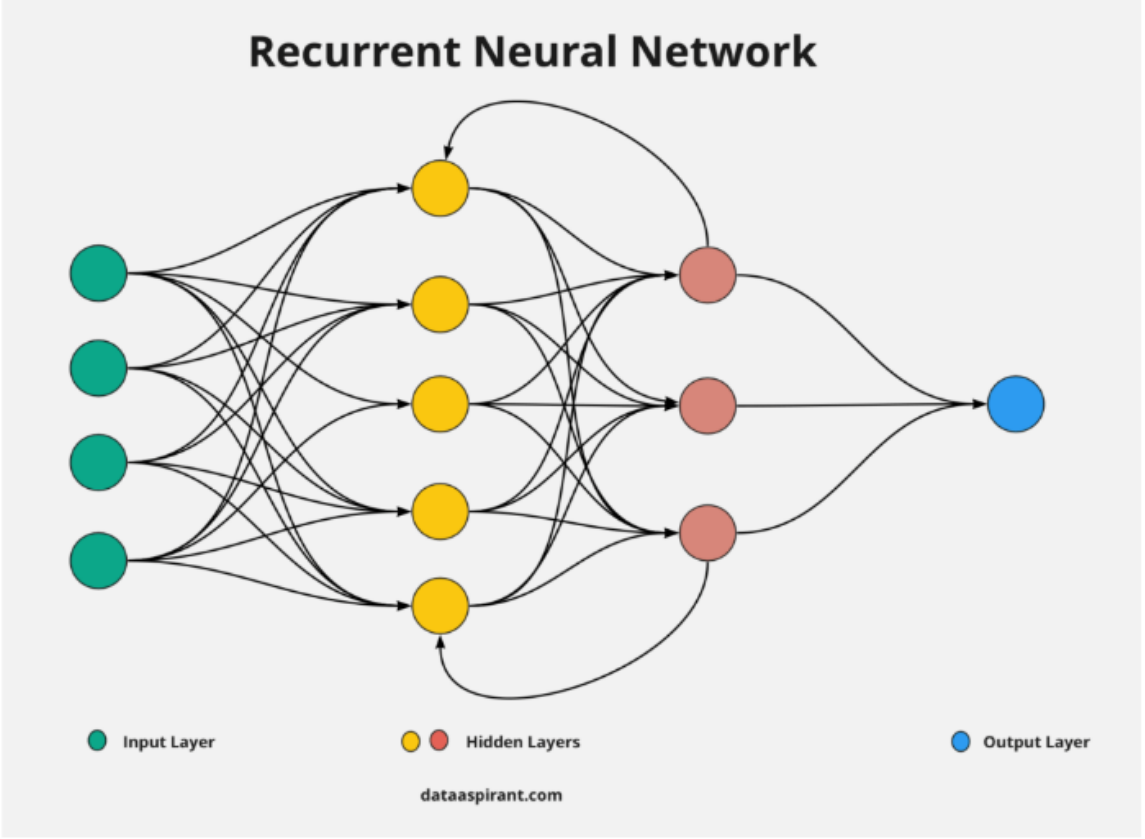


Figura 3. Rappresentazione grafica di una RNN, fonte: Sbardella, L. (2025).

Autoencoder e GAN:

Accanto a questo, esiste l'apprendimento non supervisionato, che serve quando non si hanno etichette. Qui due idee sono diventate importanti. La prima sono gli autoencoder: comprimono i dati in una rappresentazione interna più piccola e poi provano a ricostruirli. Questo aiuta a togliere

rumore, ridurre la dimensione e capire la struttura nascosta dei dati. La seconda sono le GAN (Generative Adversarial Networks): due reti che "giocano" una contro l'altra, una genera esempi finti e l'altra cerca di riconoscerli. Questa competizione porta a creare immagini e altri contenuti molto realistici, anche se l'addestramento può essere difficile e richiedere molte risorse.

Deep Reinforcement Learning:

Infine, unendo le reti profonde con l'apprendimento per rinforzo, si è riusciti a costruire agenti che vedono, decidono e agiscono meglio in ambienti complessi. Con i metodi come DQN si sono raggiunti risultati di livello umano in molti giochi Atari partendo dai pixel. I metodi actor-critic, come PPO e SAC, hanno reso l'addestramento più stabile e adatto anche ad azioni "continue" (per esempio nei robot). Approcci come MuZero fanno un passo oltre: imparano un modello interno del problema e lo usano per pianificare, anche senza conoscere in anticipo le regole esatte dell'ambiente.

In sintesi, l'idea generale è questa: scegliere architetture che rispettino la forma del dato (spaziale per le immagini, temporale per le sequenze), imparare buone rappresentazioni e, quando serve, saper generare nuovi esempi. Poi integrare tutto questo nella capacità di prendere decisioni nel tempo. I progressi degli ultimi anni sono stati possibili sia grazie a nuove idee, sia grazie a computer più potenti che permettono di addestrare reti profonde in modo efficace.

3.4 ALGORITMI DI ML E DL

Si elenca di seguito la lista di alcuni algoritmi che sono stati finora citati e non. Un quadro dei principali algoritmi divisi rispettivamente in **metodi classici di ML** e **metodi di DL**. Ricordando che i metodi classici di ML si basano su feature engineering esplicito e architetture relativamente semplici, mentre i metodi di DL apprendono automaticamente rappresentazioni gerarchiche attraverso architetture multi-strato profonde.

Dunque la struttura delle metodologie e pratiche del campo segue quella delle macro aree ML e DL a seconda della complessità del metodo, e conseguentemente, a seconda della forma di apprendimento applicata, che può essere di SL-ML, o di SL-DL, stessa cosa per l'UL e il RL. Ad esempio, se si dispone di dati etichettati e il rapporto tra feature e output è ben catturato da un modello lineare su feature già ingegnerizzate, è naturale adottare un metodo di apprendimento supervisionato (SL) nell'alveo ML (ad es. regressione lineare/logistica).

La distinzione SL/UL/RL è ortogonale a ML/DL:

per ciascuno dei tre paradigmi si possono avere sia metodi classici sia metodi deep.

Metodi Classici (ML-SL)

Linear Regression (Regressione Lineare) :

Si immagini di avere dei punti su un grafico e di voler disegnare una linea retta che li rappresenti al meglio. La Regressione Lineare fa proprio questo: trova la "migliore" linea retta che descrive la relazione tra un input numerico (ad esempio, le ore di studio) e un output numerico (ad esempio, il voto all'esame). L'obiettivo è prevedere un valore continuo.

Logistic Regression (Regressione Logistica):

A differenza della Regressione Lineare, la Regressione Logistica non prevede un valore numerico continuo, ma la probabilità che un evento accada, per poi classificarlo in una di due categorie (ad esempio, "sì" o "no", "spam" o "non spam"). Pensa a voler prevedere se un cliente comprerà un prodotto (sì/no) basandosi sulla sua età e sul suo reddito. L'algoritmo calcola una probabilità e, se questa supera una certa soglia, assegna l'evento a una categoria.

Metodi Classici (ML-UL):

K-Means Clustering (Raggruppamento K-Medie):

Si immagini di avere una scatola piena di caramelle di diversi colori e sapori, tutte mescolate, e di voler creare gruppi omogenei senza sapere in anticipo quali caratteristiche rendono simili le

caramelle. Il K-Means fa proprio questo: divide automaticamente i dati in K gruppi (cluster) in modo che i punti all'interno di ogni gruppo siano il più simili possibile tra loro e il più diversi possibile da quelli degli altri gruppi. L'algoritmo inizia scegliendo casualmente K "centri" e poi, iterativamente, assegna ogni punto al centro più vicino e ricalcola i centri come media dei punti assegnati. È come organizzare spontaneamente gli studenti di una scuola in gruppi basandosi su interessi comuni, senza sapere in anticipo quali gruppi esistono. Applicazione tipica: segmentazione clienti per marketing personalizzato.

Principal Component Analysis (PCA - Analisi delle Componenti Principali):

Immaginare di fotografare un oggetto tridimensionale da diverse angolazioni e di voler trovare l'angolazione che cattura la maggior parte delle informazioni importanti. La PCA fa qualcosa di simile con i dati: trova le "direzioni" (componenti principali) lungo le quali i dati variano maggiormente, permettendo di ridurre la complessità mantenendo l'informazione essenziale. Se hai 100 caratteristiche che descrivono i tuoi dati, la PCA può dirti che in realtà bastano 10 componenti principali per catturare il 95% della variabilità. È come riassumere un libro di 500 pagine in 50 pagine mantenendo tutti i concetti chiave. Applicazione tipica: ridurre dimensionalità di dataset con molte variabili correlate prima di applicare altri algoritmi.

Metodi Classici (ML-RL):

Q-Learning (Apprendimento Q):

Immaginare un robot che deve imparare a navigare in un labirinto per raggiungere un obiettivo. Il Q-Learning permette all'agente di imparare quali azioni sono migliori in ogni situazione (stato) attraverso tentativi ed errori. L'algoritmo costruisce una "tabella Q" che assegna un valore (Q-value) a ogni coppia stato-azione, indicando quanto è vantaggioso compiere quella specifica azione in quello stato. Inizialmente il robot non sa nulla e esplora casualmente; gradualmente, attraverso ricompense positive (raggiunge l'obiettivo) e negative (sbatte contro un muro), impara la strategia ottimale. È come imparare a giocare a scacchi giocando migliaia di partite e memorizzando quali mosse hanno portato alla vittoria. Applicazione tipica: controllo robotico, ottimizzazione di percorsi, game AI.

SARSA (State-Action-Reward-State-Action):

SARSA è un algoritmo molto simile al Q-Learning ma con una differenza fondamentale: mentre Q-Learning impara la strategia ottimale indipendentemente da cosa fa realmente l'agente (approccio "off-policy"), SARSA impara dalla strategia che l'agente sta effettivamente seguendo (approccio "on-policy"). Immagina un conducente che impara a guidare: Q-Learning imparerebbe la guida perfetta teorica, mentre SARSA impara tenendo conto che durante l'apprendimento si fanno anche errori e si deve essere cauti. Questo rende SARSA più conservativo e adatto a situazioni dove gli errori durante l'apprendimento sono costosi. Applicazione tipica: sistemi di controllo safety-critical, apprendimento in ambienti stocastici dove la sicurezza è prioritaria.

Metodi Deep (DL-SL):

In questa categoria di Deep Learning **supervisionato** (elaborazione con feature ingegnerizzate a priori) è possibile rivedere le già citate e spiegate CNN e RNN.

Convolutional Neural Networks (CNN - Reti Neurali Convoluzionali).

Recurrent Neural Networks (RNN) e Long Short-Term Memory (LSTM).

Metodi Deep (DL-UL):

Nella categoria di Deep Learning **non supervisionato** (dati non etichettati , il sistema realizza aggregazioni per sua iniziativa senza etichette nate a priori) si ritrovano reti già citate ma che si preferisce ricapitolare ed approfondire di seguito :

Autoencoders (Autocodificatori):

Un Autoencoder è una rete neurale che impara a comprimere i dati in una rappresentazione più piccola (encoding) e poi a ricostruirli (decoding), cercando di minimizzare la differenza tra input e output. È come un artista che deve riprodurre un'immagine complessa passando attraverso una descrizione testuale molto breve: per riuscirci, deve catturare solo le caratteristiche essenziali. La parte centrale ristretta (bottleneck) della rete impara una rappresentazione compressa che cattura l'essenza dei dati. Una volta addestrato, l'encoder può essere usato per ridurre la dimensionalità, mentre il decoder può generare nuovi dati simili a quelli di training. Esistono varianti come i Variational Autoencoders (VAE) che apprendono rappresentazioni probabilistiche adatte alla

generazione. Applicazione tipica: riduzione di dimensionalità, denoising di immagini, anomaly detection, compressione dati.

Generative Adversarial Networks (GAN - Reti Generative Avversarie):

Le GAN sono composte da due reti neurali che "competono" tra loro: un **Generatore** che crea dati falsi (ad esempio immagini sintetiche) e un **Discriminatore** che cerca di distinguere i dati reali da quelli falsi. È come un falsario (generatore) che cerca di creare banconote false sempre più convincenti e un detective (discriminatore) che impara a riconoscerle. Man mano che il discriminatore diventa più bravo a rilevare i falsi, il generatore deve migliorare per ingannarlo. Questo processo competitivo porta il generatore a creare dati sintetici estremamente realistici. Il training è notoriamente instabile e richiede un bilanciamento attento tra le due reti.

Applicazione tipica: generazione di immagini fotorealistiche, data augmentation, sintesi di volti umani, trasferimento di stile, generazione video.

Metodi Deep (DL-RL):

Deep Q-Network (DQN):

DQN combina il Q-Learning classico con reti neurali profonde per gestire spazi di stati enormi o continui dove una tabella Q sarebbe impossibile da memorizzare. Invece di una tabella, una rete neurale apprende la funzione Q che mappa stati ed azioni. È come passare da un manuale di scacchi con tutte le mosse possibili elencate (impraticabile), a un maestro che ha interiorizzato i principi e sa valutare qualsiasi posizione. DQN ha rivoluzionato il campo nel 2015 quando ha imparato a giocare a videogame Atari direttamente dai pixel, raggiungendo livelli sovrumani. Usa tecniche come experience replay (memorizzare e riutilizzare esperienze passate) e target network (rete separata per stabilizzare il training). Applicazione tipica: game AI, controllo robotico con sensori visivi, ottimizzazione di sistemi complessi.

Policy Gradient Methods (Metodi a Gradiente di Policy) - es. PPO :

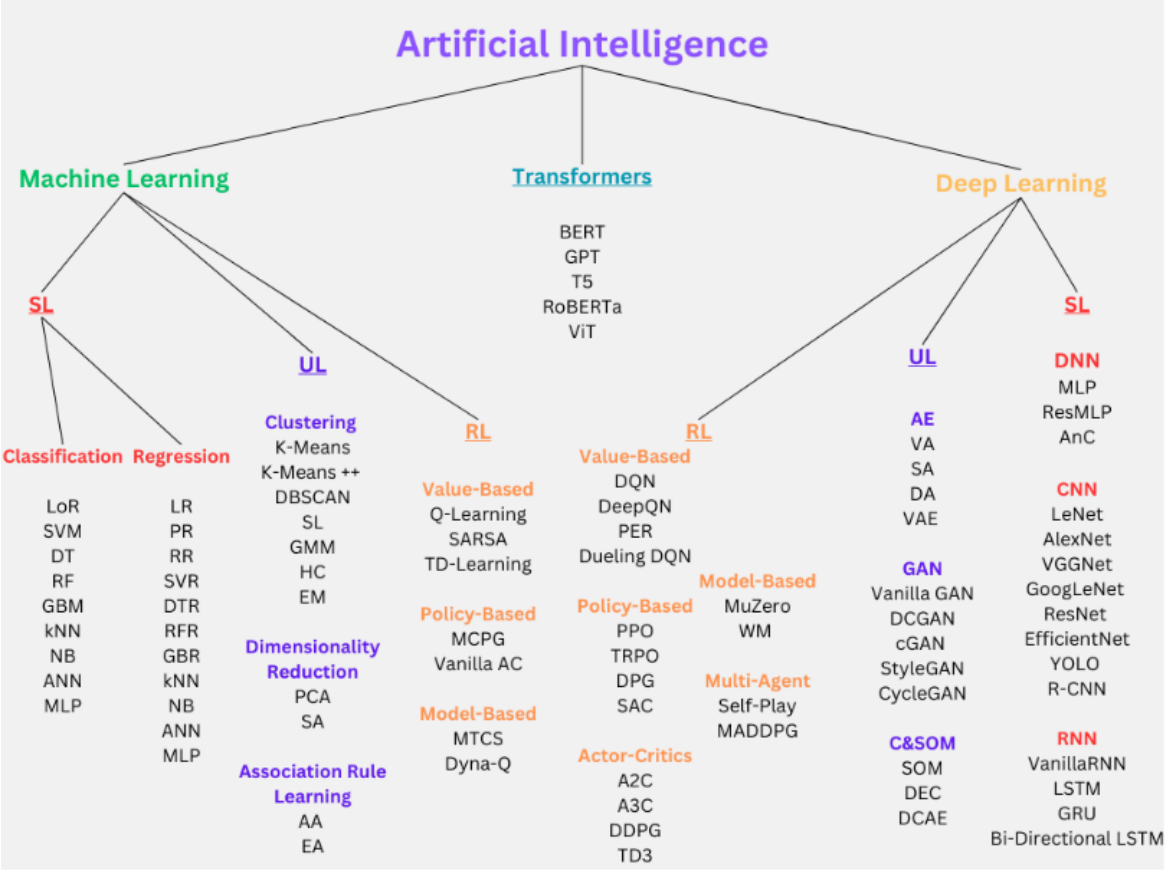
Mentre DQN apprende una funzione valore (Q), i metodi Policy Gradient apprendono direttamente la policy (strategia comportamentale): una funzione che mappa stati in azioni, spesso rappresentata da una rete neurale. È come la differenza tra imparare a valutare le posizioni di scacchi (valore) e imparare direttamente a muovere i pezzi (policy). PPO (Proximal Policy Optimization) è uno degli algoritmi più popolari di questa famiglia: aggiorna la policy in modo

conservativo, evitando cambiamenti troppo drastici che potrebbero peggiorare le prestazioni. Questi metodi sono particolarmente adatti per azioni continue (es. controllo motorio) e hanno ottenuto successi notevoli nel controllo robotico e nei giochi complessi come Dota 2 e StarCraft. Applicazione tipica: controllo robotico complesso, manipolazione oggetti, locomozione, sistemi multi-agente, decisioni sequenziali in ambienti continui.

Le reti neurali sono modelli estremamente versatili a differenza dei singoli algoritmi, in quanto possono essere impiegate sia nell'apprendimento supervisionato (Supervised Learning) sia in quello non supervisionato (Unsupervised Learning), adattando semplicemente l'obiettivo di addestramento e la funzione di perdita. Tuttavia, esistono architetture progettate per casi specifici, ottimizzate in base alla natura dei dati e al tipo di problema da risolvere. Le reti convoluzionali (CNN), ad esempio, sono specializzate nel trattamento di dati con struttura spaziale come le immagini, mentre le reti ricorrenti (RNN, LSTM, GRU) gestiscono sequenze temporali, come segnali o testi. Le reti neurali a grafo (GNN) sono progettate per dati relazionali come reti sociali o strutture molecolari, e i modelli generativi come GAN, VAE o diffusion models sono specifici per la creazione di nuovi dati realistici, come immagini o suoni. Allo stesso tempo, architetture più generiche come le reti feedforward (MLP) e i Transformer si adattano facilmente a diversi domini, risultando adatti sia a compiti di classificazione supervisionata che di rappresentazione non supervisionata. In sintesi, la scelta dell'architettura neurale non dipende solo dal tipo di apprendimento, ma soprattutto dalla struttura intrinseca dei dati e dalla natura del problema da risolvere.

I Transformers sono l'ultima frontiera nel Deep Learning, soprattutto per il Natural Language Processing (NLP), ma anche per la visione. La loro innovazione principale è il meccanismo di "attenzione", che permette al modello di pesare l'importanza di diverse parti dell'input in modo dinamico, senza dover elaborare i dati in sequenza. Questo li rende molto efficienti e capaci di catturare relazioni a lungo raggio nei dati. Sono alla base di modelli come GPT (per la generazione di testo) e BERT (per la comprensione del linguaggio). Nel prossimo capitolo verrà spiegato in chiave approfondita proprio il Transformer, che sarà il ponte collegante con il mondo dei LLMs .

SCHEMA RIEPILOGATIVO :



Fonte: Sbardella, L. (2025).

Questo schema chiarisce perfettamente la suddivisione finora esplicata, e completa con altri algoritmi non elencati qui ma che possono essere consultati nel lavoro più indicato di Sbardella, L. (2025).

3.5 TRANSFORMER

3.5.1 Introduzione al Transformer

Viene presentata la colonna portante della maggior parte dei LLM in circolazione, il Transformer. Verrà fatto uso dello studio fondativo di Vaswani et al.(2017) affiancato dalle interpretazioni dell'elaborato di Sbardella (2025).

Cos'è il Transformer e quale problema risolve:

Il Transformer è un'architettura di rete neurale introdotta da Vaswani et al. (2017) per i compiti di trasduzione di sequenze: ad esempio traduzione automatica, riassunto, trascrizione. Il punto di forza è che, a differenza delle reti ricorrenti (RNN, LSTM), non elabora la sequenza token per token in modo sequenziale, ma considera tutti i token contemporaneamente. Questo permette grande parallelismo durante l'addestramento e rende più semplice catturare relazioni tra parole lontane nella frase.

La sua architettura generale viene divisa in due blocchi principali chiamati encoder e decoder. L'encoder prende la sequenza di input (es. parole di una frase) e la trasforma in una serie di vettori che ne rappresentano contenuto e relazioni. Il decoder usa questi vettori per generare la sequenza di output una parola alla volta, in modo autoregressivo (cioè ogni parola generata viene usata come contesto per generare la successiva, ma non vi è un effettivo ritorno indietro come in un modello bidirezionale, ad esempio BERT). Ogni blocco è costituito da più strati uguali (tipicamente 6), e in ciascun strato ci sono sottoblocchi di attenzione e una rete feed-forward (che spiegheremo successivamente).

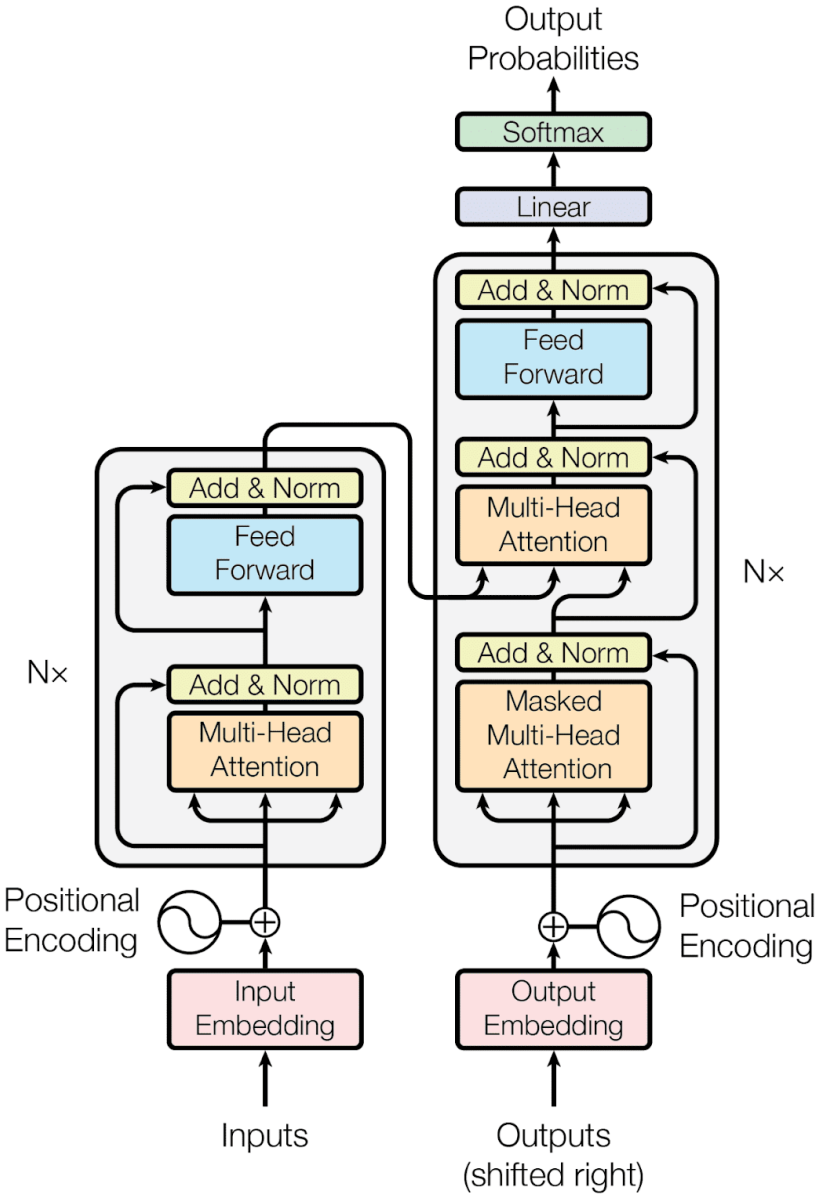


Fig.4 rappresentazione del Transformer di Vaswani et al. (2017).

Seguendo lo schema è possibile verificare tutti i passaggi spiegati di seguito.

3.5.2 Tokenizzazione ed embedding

Il primo step che attua un Transformer è la trasformazione di un frammento di testo o una parola in un token (unità discreta). Il secondo step è l'embedding, la trasformazione del token in una rappresentazione numerica: cioè un vettore di dimensione d_model (ad esempio a 512 dimensioni). Questi vettori non sono valori arbitrari ma parametri appresi durante l'addestramento

e permettono di rappresentare informazioni semantiche (parole simili avranno embedding vicini nello spazio). L'embedding permette al modello di manipolare testi come se fossero calcoli numerici.

3.5.3 Codifica posizionale

Poiché il Transformer vede tutti i token insieme, bisogna aggiungere informazioni sulla posizione (altrimenti "I love pizza" e "pizza love I" sarebbero indistinguibili). Si aggiunge quindi un vettore di positional encoding a ciascun embedding. Nel paper originale (Vaswani et al., 2017) si usano funzioni seno e coseno a frequenze diverse: ogni dimensione del positional encoding è una sinusoide con una frequenza diversa, in modo che ogni posizione abbia un pattern unico e che il modello possa calcolare relazioni relative tra posizioni. Esiste anche l'alternativa di embedding posizionali appresi (vettori che il modello impara direttamente); entrambe funzionano, ma i sinusoidali aiutano a generalizzare su sequenze più lunghe di quelle viste in training.

3.5.4 Il meccanismo di Attention

Il meccanismo di attenzione, alla base del modello Transformer, consente a ogni token di valutare l'importanza degli altri token nel contesto, per costruire una rappresentazione più ricca e contestualizzata. Per ciascun token, il modello genera tre vettori distinti: Query (Q), Key (K) e Value (V). La Query rappresenta ciò che il token sta cercando, la Key descrive le caratteristiche di ogni token che possono essere utili, mentre il Value contiene l'informazione effettiva associata a quel token. Il modello calcola quindi la similarità tra la Query del token in esame e le Key di tutti gli altri token, ottenendo un punteggio di rilevanza.

Questi punteggi vengono successivamente normalizzati tramite una funzione softmax. La normalizzazione softmax trasforma i punteggi di similarità in una distribuzione di probabilità, dove ogni valore è compreso tra 0 e 1 e la somma di tutti i valori è pari a 1 (cioè, la somma totale dei pesi assegnati a tutti i token è uguale a uno, garantendo che rappresentino proporzioni relative all'interno del contesto). Questo passaggio è cruciale perché permette di interpretare i pesi come l'importanza relativa di ciascun token nel contesto, garantendo che i token più rilevanti abbiano un peso maggiore e quelli meno rilevanti un peso minore, in modo proporzionale e stabile.

Infine, questa distribuzione di pesi viene utilizzata per combinare i Value corrispondenti. Ciò significa che il vettore di output per il token corrente viene calcolato come una somma ponderata dei vettori Value di tutti gli altri token. Ogni Value viene moltiplicato per il suo peso softmax calcolato in precedenza, e i risultati vengono sommati. In pratica, il token "raccoglie" le informazioni (i Value) dagli altri token, dando maggiore enfasi a quelle provenienti dai token che la

softmax ha identificato come più rilevanti. In questo modo, ogni token costruisce la propria rappresentazione tenendo conto in modo ponderato delle informazioni più rilevanti presenti nel contesto.

La formula usata è la **scaled dot-product attention:**

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

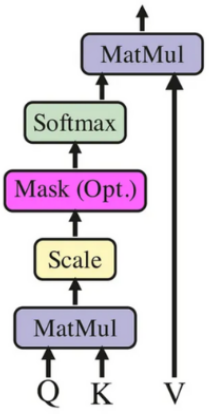


Fig.5 Formula scale dot-product attention rappresentata da Vaswani et al. (2017).

L'operazione QK^T calcola le similarità tra il vettore Query (Q) e i vettori Key (K) di tutti i token. Successivamente, il risultato viene diviso per √d_k, dove d_k è la dimensione dei vettori Key. Questo passaggio è chiamato scaling (scalatura) ed è necessario per mantenere i valori in un intervallo adeguato. In particolare, quando la dimensione d_k è molto grande, i valori di QK^T tendono a crescere molto, causando problemi durante l'addestramento del modello. Questi problemi sono legati ai gradienti, che sono vettori che indicano la direzione e l'intensità con cui aggiornare i pesi del modello durante l'apprendimento. Se i valori diventano troppo grandi, i gradienti possono diventare instabili, cioè troppo grandi o troppo piccoli, rendendo difficile o inefficace l'ottimizzazione del modello. Lo scaling serve quindi a stabilizzare i gradienti, mantenendoli in un intervallo che facilita un apprendimento più stabile e veloce.

Dopo lo scaling, la funzione softmax trasforma le similarità in pesi positivi che sommano a 1, permettendo di interpretare questi valori come una distribuzione di importanza relativa. Infine, moltiplicando questi pesi per i vettori Value (V), che rappresentano il contenuto informativo

effettivo associato a ogni token, si ottiene una combinazione pesata delle informazioni più rilevanti per il token in esame.

3.5.5 Multi-Head Attention

Il meccanismo di attenzione nel Transformer viene applicato attraverso una struttura chiamata Multi-Head Attention, che significa che il modello utilizza contemporaneamente h "teste" di attenzione parallele. Per "teste parallele" si intende che il modello esegue più calcoli di attenzione indipendenti e simultanei, ognuno focalizzato su aspetti diversi dell'informazione, invece di un singolo calcolo unico.

Ogni testa prende i vettori Query (Q), Key (K) e Value (V) e li proietta in uno spazio di dimensione minore, cioè li trasforma in rappresentazioni con dimensioni ridotte rispetto allo spazio originale. In pratica, se la dimensione originale dei vettori è d_model, ogni testa lavora su vettori di dimensione d_k = d_model/h. Questa riduzione permette di distribuire il lavoro tra le teste, mantenendo complessivamente la stessa dimensione totale, ma consentendo a ciascuna testa di concentrarsi su caratteristiche diverse.

Ogni testa calcola quindi un'attenzione indipendente e produce un risultato specifico. I risultati di tutte le teste vengono poi concatenati (cioè uniti in un unico vettore più grande) e successivamente proiettati nuovamente in uno spazio di dimensione d_model.

Con "dimensione originale dei vettori" si intende la dimensione dello spazio vettoriale in cui ogni token è rappresentato all'interno del modello. Più precisamente, nel Transformer ogni token viene codificato come un vettore numerico di dimensione d_model (ad esempio 512, 768, 1024, a seconda della configurazione del modello). Questo vettore rappresenta le caratteristiche del token in uno spazio multidimensionale, che il modello usa per elaborare il significato e il contesto. Quindi, quando si dice che ogni testa proietta Q, K e V in uno spazio di dimensione minore, significa che questi vettori originali (di dimensione d_model) vengono trasformati in vettori più piccoli (di dimensione d_k = d_model/h) per permettere alle diverse teste di attenzione di lavorare su rappresentazioni più compatte e specializzate.

Il vantaggio di questo approccio è che diverse teste possono apprendere e catturare diversi tipi di relazioni tra i token: ad esempio, una testa può specializzarsi nel seguire relazioni sintattiche tra

parole vicine, mentre un'altra può cogliere rapporti semantici a lunga distanza. Questo arricchisce la rappresentazione finale, rendendola più completa e robusta.

3.5.6 Add & Norm

Dopo ogni sottoblocco del Transformer, cioè dopo ogni singola componente funzionale come il meccanismo di attenzione o il blocco feed-forward (che verrà spiegato dopo), viene applicata una connessione residua. Questo significa che all'output prodotto dal sottoblocco si somma direttamente l'input originale di quel sottoblocco, secondo la formula:

output = x + f(x)

dove x è l'input e f(x) è la trasformazione effettuata dal sottoblocco. Questa "via diretta" facilita il passaggio dei gradienti durante l'addestramento, evitando che si indeboliscano troppo in reti molto profonde, e aiuta a preservare le informazioni iniziali.

Dopo la connessione residua, entra in gioco la Layer Normalization (LayerNorm). Immagina di avere un gruppo di numeri (il tuo vettore) e vuoi che siano tutti "sulla stessa scala" per facilitare il lavoro del modello. LayerNorm fa due cose principali:

1. Centra i numeri: calcola la media di tutti i numeri nel tuo gruppo. La media è come il valore "tipico" o "centrale". Poi, da ogni singolo numero del gruppo, sottrae questa media. Il risultato è che ora tutti i numeri sono "centrati" attorno allo zero: alcuni saranno positivi, altri negativi, ma la loro media sarà zero. È come spostare tutti i numeri in modo che il loro punto centrale sia lo zero.

2. Standardizza la dispersione: calcola la deviazione standard. Questa misura ti dice quanto i numeri sono "sparsi" o "vicini" tra loro. Se la deviazione standard è piccola, i numeri sono molto vicini; se è grande, sono molto distanti. LayerNorm poi divide ogni numero (già centrato) per questa deviazione standard. Questo fa sì che la "dispersione" dei numeri diventi standard, come se tutti avessero la stessa "ampiezza" di variazione.

In pratica, dopo questi due passaggi, tutti i numeri nel tuo gruppo avranno una media di zero e una "dispersione" standard (tecnicamente, una varianza di uno). Questo è molto utile perché rende i dati più "prevedibili" e "uniformi" per il modello, riducendo le differenze eccessive che

potrebbero confondere l'apprendimento. È come mettere tutti i dati su un binario standardizzato, rendendo l'allenamento del modello più stabile e veloce.

Infine, LayerNorm scala e trasla il vettore normalizzato usando due parametri appresi durante l'addestramento, chiamati γ (gamma) e β (beta). Questi parametri permettono al modello di adattare la normalizzazione in modo flessibile, mantenendo la capacità di rappresentare informazioni importanti.

3.5.7 Feed-Forward Network

Dopo il meccanismo di attenzione, ogni token passa attraverso una piccola rete chiamata feed-forward, che viene applicata separatamente a ciascun token, ma con gli stessi pesi condivisi per tutte le posizioni. Questo significa che ogni token viene trasformato in modo indipendente, ma usando la stessa funzione per tutti, garantendo coerenza nel trattamento.

Questa rete feed-forward è composta da una sequenza specifica di operazioni: prima una trasformazione lineare, poi una funzione non lineare chiamata ReLU (Rectified Linear Unit), e infine un'altra trasformazione lineare.

Una trasformazione lineare è un'operazione matematica che, in termini semplici, può scalare, ruotare o spostare i dati in modo uniforme, mantenendo le relazioni proporzionali tra di essi. Immagina di disegnare una linea retta su un grafico: una trasformazione lineare manterrebbe quella linea retta, magari cambiandone l'inclinazione o la posizione. È un'operazione "semplice" che non introduce complessità nelle relazioni tra i dati.

La funzione ReLU, invece, introduce una non linearità. A differenza delle trasformazioni lineari, una funzione non lineare può piegare, curvare o distorcere le relazioni tra i dati. La ReLU, in particolare, prende ogni valore in ingresso e lo trasforma in zero se è negativo, oppure lo lascia invariato se è positivo. Questo è cruciale perché il mondo reale è pieno di relazioni complesse e non lineari (ad esempio, non tutto è direttamente proporzionale). Senza non linearità, il modello sarebbe limitato a imparare solo relazioni molto semplici.

Il senso di fare questo processo in sequenza (lineare → non lineare → lineare) è proprio quello di combinare la semplicità e l'efficienza delle trasformazioni lineari con la capacità di modellare

complessità delle non linearità. La prima trasformazione lineare prepara i dati, la ReLU introduce la capacità di apprendere schemi complessi e non lineari, e la seconda trasformazione lineare riorganizza i dati trasformati in un formato utile per i passaggi successivi del modello. Questo blocco feed-forward aggiunge al modello la capacità di trasformare localmente ogni token, arricchendo la rappresentazione e migliorando la capacità complessiva del Transformer di comprendere e processare il testo.

3.5.8 Il Decoder

Il decoder del Transformer ha il compito di generare il testo, parola per parola. In ogni suo "strato" (un blocco di elaborazione), avvengono tre passaggi chiave:

1. **Attenzione Mascherata:** Il decoder "guarda" le parole che ha già generato fino a quel momento. Una "maschera" speciale impedisce che veda le parole future, assicurando che la generazione avvenga in modo sequenziale, proprio come una persona scrive una frase.

2. **Attenzione Incrociata (Cross-Attention):** Qui il decoder "guarda" l'input originale (quello che l'encoder ha elaborato). Questo gli permette di capire quali parti dell'input sono più importanti per generare la parola successiva.

3. **Rete Feed-Forward:** Infine, un piccolo blocco di elaborazione (simile a quello dell'encoder) affina la rappresentazione del token.

Dopo questi passaggi, il decoder produce un risultato che viene trasformato in un elenco di "punteggi" per ogni possibile parola del vocabolario. Una funzione softmax converte questi punteggi in probabilità, indicando quanto è probabile che ogni parola sia quella giusta da generare. Per scegliere la parola finale, il modello può semplicemente prendere quella con la probabilità più alta (metodo "greedy"). Oppure, può usare tecniche più avanzate come la beam search, che esplora diverse sequenze di parole più probabili contemporaneamente, cercando di trovare la frase migliore nel complesso.

Durante l'addestramento, il modello impara confrontando le parole che predice con le parole corrette (Supervised Learning). Cerca di minimizzare una "perdita" (chiamata cross-entropy), che misura quanto le sue previsioni si discostano dalla realtà. Per rendere l'apprendimento più veloce

e stabile, si usa una tecnica chiamata teacher forcing: invece di dare al decoder la parola che ha generato lui stesso nel passo precedente, gli si fornisce la parola corretta del testo originale. Questo lo guida meglio durante l'allenamento. L'ottimizzazione avviene con algoritmi specifici che aggiustano i pesi del modello.

3.5.9 Limiti e Complessità

Il Transformer è potente, ma ha un limite: il suo meccanismo di attenzione può diventare molto costoso in termini di calcolo e memoria quando le sequenze di testo (o audio, video) sono molto lunghe. Questo perché deve confrontare ogni parola con ogni altra parola, portando a una complessità computazionale quadratica rispetto alla lunghezza della sequenza ($O(n^2)$). Per questo motivo, sono state sviluppate varianti più efficienti come Longformer, Reformer e altre architetture che cercano di ridurre questa complessità mantenendo l'efficacia del modello.

4. LLMs

Dopo aver spiegato l'architettura Transformer, che sappiamo essere alla base dei moderni LLM, è arrivato il momento di comprendere come sia possibile addestrare questa architettura affinché diventi un potente modello conversazionale.

4.1 INTRODUZIONE AI LARGE LANGUAGE MODELS

Per comprendere i Large Language Models (LLM), ci serviremo dello studio interpretativo di Karpathy (2024), prendendo come modello di analisi GPT-2.

È fondamentale partire dalla rappresentazione del testo in input. Come evidenziato da Karpathy (2024), i modelli neurali (che abbiamo ampiamente spiegato precedentemente) elaborano sequenze monodimensionali di simboli discreti. Sebbene il testo sia visualizzato in forma bidimensionale (righe e colonne), esso viene letto come una sequenza lineare, da sinistra a destra e dall'alto verso il basso.

4.1.1 Tokenizzazione e rappresentazione del testo

Il primo passo cruciale consiste nel definire un vocabolario di simboli (token) che rappresentino il testo. La tokenizzazione è un processo essenziale che bilancia la dimensione del vocabolario e la lunghezza della sequenza da elaborare. Più la sequenza è breve, più efficiente sarà il modello nell'elaborazione; per questo motivo si preferisce un vocabolario più ricco che permetta sequenze più corte.

Ad esempio, la codifica in byte (256 simboli) riduce la lunghezza della sequenza rispetto ai bit, ma per ottimizzare ulteriormente il processo si utilizza l'algoritmo Byte Pair Encoding (BPE) (Sennrich et al., 2016), che raggruppa coppie frequenti di simboli in nuovi token, espandendo il vocabolario fino a circa 100.000 simboli.

Questo processo di tokenizzazione consente di rappresentare il testo in modo efficiente, riducendo la lunghezza delle sequenze e facilitando l'addestramento e l'inferenza. La scelta del vocabolario e la tokenizzazione sono fondamentali poiché la lunghezza della sequenza rappresenta una risorsa computazionale limitata e preziosa per i modelli, come sottolineato da Karpathy (2024).

4.1.2 Dai tokens alla rete neurale

Una volta convertito il testo in sequenze di token, il modello deve trasformarli in previsioni utili. Qui entra in gioco l'architettura del modello, che deve essere in grado di catturare le complesse dipendenze e strutture del linguaggio naturale.

Karpathy (2024) descrive la rete neurale come una funzione matematica parametrizzata da miliardi di pesi, inizialmente casuali, che vengono ottimizzati durante l'addestramento. Questi parametri sono analoghi alle manopole su un mixer: ruotandole, si modifica l'output del modello. L'obiettivo consiste nel trovare la configurazione di parametri che permetta al modello di prevedere con precisione il token successivo, catturando i pattern statistici del linguaggio.

Contestualizzando con le spiegazioni fornite in precedenza, si nota come ritornino i vari concetti: l'ottimizzazione dei parametri non è altro che il processo attuato da dinamiche come la discesa del gradiente, l'autoregressività e la backpropagation, che consentono la rielaborazione in maniera profonda dei dati e quindi la parametrizzazione ottimale dei pesi per eventuali miglioramenti.

L'architettura Transformer, introdotta da Vaswani et al. (2017) nel celebre paper "Attention is all you need", è alla base di GPT-2 e dei modelli successivi. Questa architettura ha rivoluzionato il campo del Natural Language Processing (NLP) grazie al meccanismo di attenzione, che consente al modello di pesare dinamicamente l'importanza relativa di ogni token nel contesto, superando i limiti delle reti ricorrenti tradizionali.

In pratica, ogni token viene trasformato in un vettore numerico (embedding) che ne rappresenta la semantica in uno spazio continuo. La tokenizzazione è il passaggio che precede l'embedding: la prima converte il testo in ID di token, mentre il secondo trasforma tali ID in numeri significativi che rappresentano il contenuto semantico nello spazio vettoriale.

Questi vettori attraversano quindi una serie di strati di attenzione multi-testa e percettroni multistrato (feed-forward), che elaborano le informazioni in parallelo, permettendo di catturare relazioni a lungo raggio tra parole anche distanti nella sequenza.

Karpathy (2024) sottolinea che i valori intermedi prodotti da questi strati possono essere interpretati come "attivazioni" di neuroni sintetici, ma è importante non confondere questi neuroni artificiali con quelli biologici: sono funzioni matematiche fisse, senza memoria interna o dinamiche complesse. La rete non ha una memoria esplicita, ma utilizza la finestra di contesto (la quantità

massima di testo che un modello linguistico può considerare contemporaneamente mentre elabora o genera una risposta).

4.2 Addestramento: ottimizzare i parametri per apprendere il linguaggio

Il processo di addestramento rappresenta il cuore pulsante nella costruzione di un Large Language Model (LLM). Come descritto nel documento di Ouyang et al. (2022), il modello parte da una configurazione iniziale di parametri casuali, che vengono progressivamente ottimizzati esponendo il modello a enormi quantità di testo tokenizzato. Nello specifico, Karpathy (2024) evidenzia la modalità principale di acquisizione di questo testo: vengono presi molti terabyte di testo estratti dalle pagine web e filtrati con algoritmi in grado di escludere quanto più possibile contenuti biased, offensivi o pericolosi, secondo linee guida protocollate da normative di sicurezza aziendali e internazionali.

L'ottimizzazione avviene tramite algoritmi di discesa del gradiente stocastica (SGD) o sue varianti più sofisticate, che aggiornano i parametri per minimizzare una funzione di perdita (loss). Questa loss misura la discrepanza tra la previsione del modello—ovvero il token successivo stimato—e il token reale presente nel testo di addestramento.

Ogni aggiornamento si basa su milioni di token, e la riduzione progressiva della loss indica che il modello sta migliorando la sua capacità di prevedere il linguaggio naturale. Questo processo richiede risorse computazionali enormi, spesso distribuite su cluster di GPU o TPU, e può durare settimane o mesi.

Durante l'addestramento, il modello impara a catturare le regolarità statistiche e le strutture sintattiche e semantiche del testo, sviluppando una rappresentazione interna del linguaggio. Tuttavia, al termine di questa fase, il modello possiede una conoscenza puramente statistica e non è ancora in grado di generare testo in modo utile o conversazionale, poiché non ha ancora appreso come interagire con un utente o seguire istruzioni specifiche.

4.3 Inferenza: generare testo a partire dal modello addestrato

Superata la fase di addestramento, si entra nella fase di inferenza, ovvero la generazione di testo a partire dal modello addestrato. Karpathy (2024) spiega che nei modelli linguistici autoregressivi la generazione del testo avviene in modo incrementale: il modello riceve un prompt iniziale, lo suddivide in token e, sulla base di questi, calcola una distribuzione di probabilità sull'intero vocabolario per stimare quale token sia più coerente come successivo nella sequenza.

Una volta selezionato il token più adatto attraverso una strategia di campionamento, esso viene aggiunto alla sequenza e l'intero processo si ripete iterativamente, aggiornando ogni volta il contesto. Ad esempio, a partire dal prompt "Il gatto dorme", il modello potrebbe assegnare la probabilità più alta al token "sul", ottenendo così la sequenza "Il gatto dorme sul", che diventa il nuovo input per la predizione seguente. La procedura continua fino al raggiungimento della lunghezza desiderata o all'emissione di un token di fine sequenza, permettendo al sistema di produrre testi mantenendo coerenza statistica con le informazioni già osservate.

Questo meccanismo di campionamento introduce variabilità e creatività nel testo generato: il modello non si limita a riprodurre esattamente ciò che ha visto durante l'addestramento, ma produce un "remix" plausibile basato sulle statistiche apprese.

Karpathy (2024) sottolinea che GPT-2, con i suoi 1,6 miliardi di parametri e una finestra di contesto di 1.024 token, rappresentava uno dei modelli più avanzati nel 2019. Tuttavia, come evidenziato in Brown et al. (2020), i modelli più recenti sono significativamente più grandi e dispongono di finestre di contesto più ampie, permettendo generazioni di testo più coerenti, contestualizzate e capaci di mantenere coerenza su testi più lunghi. Inoltre, la qualità dell'inferenza dipende anche da tecniche di campionamento come la temperatura, il top-k e il nucleus sampling, che modulano la casualità e la diversità del testo generato, bilanciando creatività e coerenza.

4.4 Dal modello base all'assistente: il ruolo del fine-tuning supervisionato

Un modello base di GPT-2 non "instructed" (termine utilizzato nel settore per distinguere il modello dalla versione "base", ovvero non ancora raffinata per essere conversazionale), pur essendo un potente simulatore statistico del linguaggio, non è immediatamente utilizzabile come assistente conversazionale o strumento interattivo. Per trasformare questo modello in un sistema capace di rispondere a domande, eseguire compiti specifici e dialogare in modo naturale, si applica una fase cruciale chiamata fine-tuning supervisionato (Supervised Fine-Tuning, SFT) (Wang et al., 2022).

Durante questa fase, il modello viene ulteriormente addestrato su dataset di conversazioni curate da etichettatori umani, i quali scrivono prompt e risposte ideali seguendo linee guida dettagliate. Questi dataset sono strutturati in modo da includere token speciali che marcano i turni di conversazione, permettendo al modello di distinguere chiaramente tra le parti coinvolte, come utente e assistente.

Karpathy (2024) sottolinea che questo processo consente al modello di apprendere non solo il contenuto delle risposte, ma anche lo stile comunicativo e le modalità di interazione più appropriate, conferendogli una sorta di "personalità" coerente con le aspettative dell'utente.

Un aspetto interessante evidenziato in Wang et al. (2022) è che, per accelerare la creazione di dati di alta qualità, oggi molti dataset di conversazioni sono generati in parte dai modelli stessi e successivamente revisionati e corretti da umani. Questo approccio ibrido permetterebbe di scalare la produzione di dati mantenendo elevati standard qualitativi.

4.5 Affidabilità e limiti: il problema delle allucinazioni

Nonostante i progressi ottenuti con il fine-tuning supervisionato, i modelli instructed continuano a soffrire di un problema noto come "allucinazioni": la generazione di risposte inventate o errate con un'apparente sicurezza. Karpathy (2024) evidenzia che questo fenomeno deriva dal fatto che il modello, durante l'addestramento, non ha mai imparato a dire "non lo so" o a manifestare incertezza, poiché non ha ricevuto esempi di risposte di questo tipo (Shojaee et al., 2025).

Per mitigare questo problema, si introducono nel training esempi specifici in cui il modello impara a rifiutarsi di rispondere quando non è sicuro, sviluppando così un comportamento associato a un cosiddetto "neurone dell'incertezza". Questo neurone agisce come un segnale interno che modula la propensione del modello a fornire risposte solo quando ha sufficiente confidenza.

Inoltre, per migliorare l'affidabilità e la precisione delle risposte, si possono integrare strumenti esterni, come motori di ricerca o database aggiornati, che il modello può attivare tramite token speciali durante l'inferenza. Questo meccanismo consente al modello di recuperare informazioni aggiornate e verificabili, superando i limiti della sua memoria statica basata sui parametri addestrati.

4.5.1 Memoria a lungo termine e memoria di lavoro

Karpathy (2024) sottolinea l'importanza di distinguere tra la memoria a lungo termine, rappresentata dai parametri del modello, e la memoria di lavoro, ovvero la finestra di contesto utilizzata durante l'inferenza. Questa distinzione è fondamentale per comprendere come i modelli gestiscono informazioni vecchie e nuove, e come l'integrazione di fonti esterne possa migliorare significativamente la qualità e l'affidabilità delle risposte generate.

La memoria a lungo termine corrisponde ai parametri appresi durante l'addestramento: essa codifica conoscenze generali, pattern linguistici e strutture concettuali che il modello ha acquisito dai dati storici. Tale memoria è relativamente statica, in quanto non può essere aggiornata durante l'inferenza senza un nuovo ciclo di addestramento. In altre parole, rappresenta ciò che il modello "sa" in modo consolidato, ma non può adattarsi autonomamente a informazioni nuove o contestuali.

La memoria di lavoro, invece, è incarnata dalla finestra di contesto, ovvero l'insieme di token che il modello può considerare simultaneamente quando genera una risposta. Essa è dinamica e limitata: solo una porzione del testo recente è accessibile in un dato momento, determinando la capacità del modello di ragionare su informazioni specifiche o sequenze complesse. La gestione efficace della finestra di contesto è cruciale per garantire coerenza e pertinenza nelle risposte, poiché consente al modello di integrare elementi contestuali immediati con le conoscenze di lungo termine.

Per superare i vincoli imposti dalla limitata finestra di contesto, è possibile integrare il modello con fonti esterne mediante tecniche di retrieval, come il Retrieval-Augmented Generation (RAG) o

l'uso di database vettoriali. Questi strumenti permettono al modello di accedere a informazioni aggiornate, specifiche e di grande volume, colmando il divario tra memoria statica e contesto dinamico. In questo modo, il modello non solo migliora l'accuratezza delle risposte, ma diventa anche uno strumento più affidabile e flessibile, capace di adattarsi a scenari reali in cui le informazioni evolvono rapidamente.

4.6 Capacità computazionali: i limiti del calcolo token per token

Karpathy (2024) mette in luce un aspetto spesso sottovalutato ma fondamentale per comprendere le capacità e i limiti dei Large Language Models: l'elaborazione del testo avviene un token alla volta, e per ogni token il modello esegue un numero finito e relativamente basso di strati di calcolo (layer) durante l'inferenza. Questo vincolo architetturale, tipico dei Transformer, limita la profondità del ragionamento che il modello può svolgere in un singolo passaggio.

In altre parole, il modello non dispone di una capacità computazionale illimitata per ogni token generato, ma deve operare entro un budget fisso di calcolo. Ciò implica che ragionamenti complessi, che richiederebbero molteplici passaggi logici o calcoli articolati, non possono essere risolti in modo diretto e monolitico.

Per superare questa limitazione, Karpathy (2024) suggerisce di addestrare i modelli a distribuire il ragionamento su più token, esplicitando passaggi intermedi nel processo di generazione. Questa strategia è formalizzata nella tecnica nota come chain of thought (CoT), che consiste nel far generare al modello una sequenza di ragionamenti intermedi prima di arrivare alla risposta finale.

Come dimostrato nel lavoro di Wei et al. (2022), questa metodologia migliora significativamente la capacità dei modelli di affrontare alcuni tipi di problemi, anche se è stato dimostrato da alcune sperimentazioni che per alcuni di essi la metodologia lineare e più semplice (senza CoT) performa più correttamente. Ad ogni modo, questo metodo consente di articolare il ragionamento in modo più "trasparente" e strutturato, e questo può aiutare nell'analisi dei passaggi che il modello segue al suo interno. Trasparente tra molte virgolette, in quanto è stato visto che spesso i ragionamenti in output con questa tecnica non corrispondono alle vere catene di processo attuate internamente da questi sistemi. Sembra come se mentissero o omettessero alcuni passaggi.

Un ulteriore aspetto evidenziato da Karpathy (2024) riguarda i compiti che richiedono calcoli precisi o conteggi esatti, come operazioni aritmetiche o manipolazioni numeriche. In questi casi, è più efficace integrare il modello con strumenti esterni, come l'esecuzione di codice Python, che possono eseguire calcoli deterministici e affidabili. Questa collaborazione tra modello e strumenti esterni riduce gli errori tipici delle allucinazioni numeriche e migliora l'affidabilità complessiva del sistema.

Questa limitazione architetturale spiega perché i modelli possono commettere errori in compiti apparentemente semplici, mentre eccellono in compiti più complessi che beneficiano di un ragionamento distribuito e di un'integrazione con strumenti esterni. Karpathy (2024) sottolinea quindi l'importanza di progettare pipeline ibride che combinino la flessibilità generativa dei LLM con la precisione di moduli specializzati.

4.7 Le tre fasi dell'addestramento

Dopo un inquadramento d'insieme sul funzionamento di questi strumenti, eseguiamo una ricapitolazione estraendo la solida struttura protocollata che ne deriva, affinché questi oggetti possano arrivare ad essere assistenti conversazionali.

Di seguito le tre fasi che guidano il processo di training di un modello, desunte dal contributo di Karpathy (2024).

4.7.1 Pretraining

La prima fase, il pretraining, consiste nell'esporre il modello a enormi quantità di testo proveniente da internet e altre fonti, con l'obiettivo di assimilare una base ampia e generale di conoscenze linguistiche e di mondo. Karpathy (2024) paragona questo processo a un bambino che legge molti libri, costruendo una comprensione generale del linguaggio, delle strutture sintattiche e semantiche, e del contesto culturale e fattuale (Brown et al., 2020).

Durante il pretraining, il modello impara a prevedere il token successivo in una sequenza, sviluppando una rappresentazione interna del linguaggio che funge da base per tutte le capacità successive.

4.7.2 Supervised Fine-Tuning (SFT)

La seconda fase, il Supervised Fine-Tuning (SFT), consiste nel raffinare il modello tramite esempi di conversazioni e compiti specifici, curati da esseri umani. In questa fase, il modello impara a rispondere in modo utile, appropriato e coerente con le aspettative degli utenti.

Karpathy (2024) paragona questa fase a un insegnante che guida lo studente, mostrando come risolvere problemi specifici e come interagire efficacemente (Wang et al., 2022).

Il fine-tuning supervisionato permette di adattare il modello a contesti applicativi concreti, migliorandone la capacità di dialogo e la pertinenza delle risposte.

4.7.3 Reinforcement Learning with Human Feedback (RLHF)

La terza fase, il Reinforcement Learning with Human Feedback (RLHF), rappresenta un ulteriore affinamento in cui il modello affronta esercizi pratici, riceve feedback umano e migliora progressivamente le sue risposte. Karpathy (2024) sottolinea che, sebbene RLHF non sia un vero e proprio apprendimento per rinforzo classico, rappresenta una forma di rifinitura che aiuta il modello a sviluppare capacità emergenti, come il ragionamento articolato, la capacità di autocorrezione e la gestione di ambiguità (Ouyang et al., 2022).

Questa fase è cruciale per far emergere comportamenti più sofisticati e per migliorare l'affidabilità e la sicurezza del modello, riducendo errori e allucinazioni.

Questa triplice struttura di addestramento consente ai modelli di passare da semplici compressori statistici a sistemi capaci di interazioni complesse e ragionate, avvicinandosi sempre più a comportamenti "esperti" (o "intelligenti" secondo i più audaci) e utili. Karpathy (2024) evidenzia come la sinergia tra queste fasi sia fondamentale per costruire modelli che non solo comprendano il linguaggio, ma siano anche in grado di utilizzarlo in modo efficace e responsabile.

4.8 Conclusioni e prospettive future

In conclusione, Karpathy (2024) ci mostra che i modelli base sono potenti compressori statistici del sapere presente sul web, ma con limiti e "buchi" cognitivi. La trasformazione in assistenti utili richiede un percorso complesso di addestramento e raffinamento.

Guardando avanti, la multimodalità (testo, audio, immagini) e l'integrazione con agenti autonomi che eseguono compiti nel tempo rappresentano le prossime frontiere. Inoltre, la gestione della finestra di contesto e l'apprendimento continuo durante l'inferenza sono sfide aperte.

Il lavoro di Karpathy (2024) offre una guida chiara e approfondita per comprendere come si costruiscono e si addestrano i LLM, con esempi concreti come GPT-2 e modelli più recenti, fornendo una base solida per chi voglia approfondire questo affascinante campo.

5. CASO DI STUDIO: ALPHA EVOLVE (DEEPMIND,2025)

Si è partiti dagli albori dell'IA, percorrendo archeologicamente tutte le architetture, la loro contestualizzazione e applicazione, arrivando ai Transformer e, conseguentemente, ai LLM, prendendo come prototipo esplicativo GPT-2. A questo punto, si compie un passo ulteriore, introducendo un esperimento pratico allo stato dell'arte che racchiuda tutte le conoscenze finora introdotte, con l'aggiunta di elementi molto interessanti ai fini dell'obiettivo di questo elaborato. Tali aggiunte sono l'agentività e la collaborazione fra LLM. Questi elementi permettono di attuare un'evoluzione importante verso una maggiore utilità, efficienza, comprensione e adattamento del e nel mondo reale da parte di questi sistemi.

L'agentività, secondo la Stanford Encyclopedia of Philosophy (2022), può essere intesa come la capacità di un essere di agire e di manifestare tale capacità attraverso le proprie azioni. In questa prospettiva, l'agente è considerato un soggetto dotato di iniziativa e potere causale nel mondo, in grado di influenzare il corso degli eventi attraverso le proprie decisioni e comportamenti.

Nel contesto dell'intelligenza artificiale, l'agentività viene descritta come la proprietà funzionale di un sistema capace di generare azioni, modellare le relazioni tra azioni e risultati, e adattare il proprio comportamento per migliorare le prestazioni rispetto a un obiettivo. Come spiegato in un recente articolo, "un sistema possiede agentività funzionale se è capace di generare azioni basate su informazioni ambientali, rappresentare relazioni tra azioni e risultati, e adattare il comportamento in risposta ai cambiamenti per mantenere o migliorare le prestazioni" (Agentic AI Needs a Systems Theory, 2023). Tale definizione non attribuisce all'IA una forma di coscienza o intenzionalità, ma una capacità operativa e adattiva orientata a uno scopo (Studylib, 2023).

L'agentività permette di passare dalla teoria alla pratica, dal consiglio all'azione. Conferisce a questi sistemi la potenziale capacità di manipolare la dimensione fisica e virtuale come esseri antropomorfi. Ciò che un assistente testuale faceva, ossia fornire elaborazioni testuali o, tutt'al più, multimodali, che rimanevano confinate nella loro finestra di contesto, adesso può determinare effetti più estesi, a maggior contatto con il mondo umano, trasformando gli stati in maniera più profonda e duratura all'interno degli ecosistemi a cui ha accesso. L'agentività non implica ancora coscienza, ma implica pianificazione, perseguimento di strategie finalizzate a obiettivi non a breve termine, soluzione diretta di problematiche ecosistemiche: tutte capacità fondanti dell'essere umano.

La collaborazione, invece, è quell'elemento che fa uscire questi sistemi dal loro individualismo esistenziale, dalla loro bolla tecnologica, ancor di più che con una semplice interazione testuale agente-umano. Per collaborazione che innova, si intende quella forma di interazione reciproca e iterativa – nel caso attuale tra agenti LLM, ma che può essere estesa ad altre forme di IA e a forme ibride umano-agente artificiale – che realizza scenari assolutamente inediti di autonoma e originale capacità creativa. Per la prima volta nella storia, vediamo una tecnologia-emulazione

dell'essere umano specchiarsi in un'altra tecnologia, attuando di fatto la prima forma di socializzazione artificiale, di interazione non più circoscritta a sé stessa (l'"io" tecnologico), ma anche con "l'altro" tecnologico. Un'esperienza duale che può manifestarsi in tutte le dimensioni a cui abbiamo accesso.

Il fascino che ne deriva è notevole se si pensa a tutti i processi che potrebbero manifestarsi. Inoltre, non è del tutto chiaro cosa avvenga specificamente nei moduli di elaborazione più profonda di queste architetture. Pensare alle potenzialità di questi processi nel formato duale apre molti scenari difficili da immaginare. Ecco perché è stato scelto di introdurre lo studio su AlphaEvolve di DeepMind (2025): uno studio ponte contenente tutte queste suggestioni, che permetterà di addentrarci nell'emisfero delle relazioni tra questi sistemi, e tra questi e gli umani, andando quindi ad analizzare sul versante esterno, macro, le implicazioni che potrebbero avere queste architetture nel nuovo panorama mediale.

5.1. CONTESTO ED OBIETTIVI

AlphaEvolve è un agente evolutivo di coding che integra Large Language Models (LLM) con valutatori automatici per esplorare, generare e ottimizzare algoritmi in domini dove la correttezza e/o l'efficienza possono essere verificate in modo programmabile. Il progetto, presentato da Google DeepMind nel 2025, si colloca nella traiettoria dei sistemi di scoperta algoritmica e ottimizzazione guidata da IA, con una specificità: l'uso di modelli linguistici di grandi dimensioni (LLM) per proporre modifiche al codice all'interno di un processo iterativo di miglioramento, in cui le versioni generate vengono valutate e selezionate in base a metriche oggettive (Google DeepMind, 2025).

5.2. ARCHITETTURA E PIPELINE

AlphaEvolve è un sistema evolutivo che sfrutta LLM per generare, testare e ottimizzare programmi in modo iterativo. Gli input principali del sistema sono tre:

1. Il programma seed, che rappresenta il punto di partenza e contiene alcune parti identificate come modificabili o evolvibili.

2. Il valutatore automatico, capace di assegnare punteggi sulla base di correttezza, efficienza e uso delle risorse.

3. Un archivio di prompt e soluzioni storiche, che fornisce contesto e riferimenti utili alla generazione di nuove soluzioni.

Il ciclo evolutivo di AlphaEvolve funziona in maniera simile a un processo naturale di selezione. Inizialmente, il sistema seleziona prompt casuali e genera modifiche al codice, chiamate patch o differenze, utilizzando un insieme di LLM che applicano strategie di esplorazione e sfruttamento. Ad esempio, un modello potrebbe proporre un nuovo algoritmo di moltiplicazione di matrici più efficiente, mentre un altro suggerisce un'alternativa più robusta ma leggermente più lenta. Le modifiche vengono quindi applicate e testate attraverso benchmark e verifiche automatiche, e le soluzioni migliori vengono salvate in un database evolutivo, pronte per essere riutilizzate nei cicli successivi.

AlphaEvolve sfrutta inoltre un parallelismo asincrono, in cui più agenti LLM lavorano contemporaneamente su varianti diverse dello stesso problema. Questo approccio aumenta la diversità delle soluzioni e permette di esplorare più rapidamente lo spazio delle possibili ottimizzazioni. Ad esempio, mentre un agente lavora su un algoritmo per ridurre le moltiplicazioni scalari in un calcolo matriciale, un altro può sperimentare strategie di scheduling delle risorse in un data center.

Il feedback del valutatore è centrale nel guidare l'evoluzione: segnali di errore e ranking delle soluzioni permettono di selezionare le modifiche più promettenti. Le soluzioni migliori vengono poi incorporate come contesto per le generazioni successive, secondo un principio di «eredità del codice», simile a come, nella genetica, le caratteristiche vantaggiose vengono trasmesse alle generazioni future.

In sintesi, AlphaEvolve può essere visto come un ciclo continuo di generazione e selezione, in cui si intrecciano: problema → valutatore → generatore (LLM) → archivio di soluzioni → selettore → nuovo ciclo di generazione. Questo approccio consente al sistema di migliorare progressivamente le soluzioni, adattandosi ai vincoli specifici di ciascun dominio, dalla matematica combinatoria alla gestione di infrastrutture o all'ottimizzazione di kernel hardware per LLM.

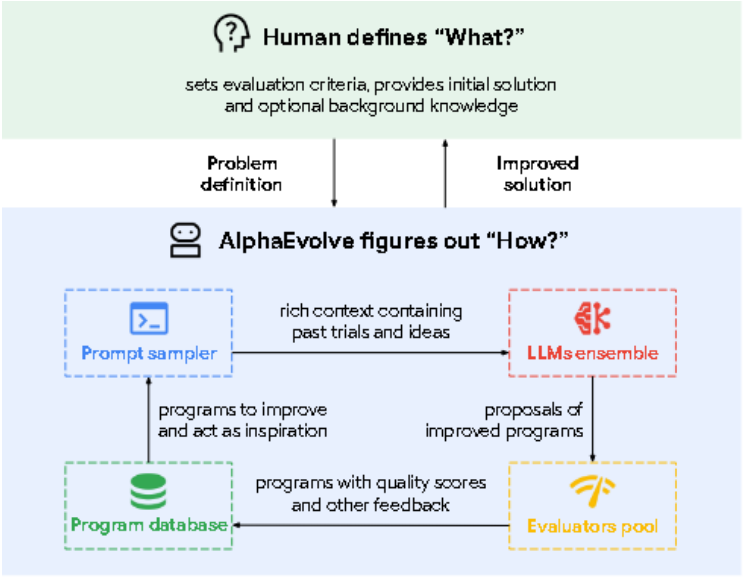


Fig.5 Catena di lavoro del modello rapp. Da DeepMind (2025).

5.3. SETTING SPERIMENTALE

Il sistema è stato testato su tre categorie di problemi, tutte con valutatori affidabili:

1. Matematica e informatica teorica: costruzione di oggetti combinatori (un oggetto combinatorio è un insieme finito di elementi organizzato secondo determinate regole o relazioni; può essere, ad esempio, una sequenza, un sottoinsieme, una permutazione, una partizione, un grafo, un albero,

una matrice, o un polinomio combinatorio, cioè qualsiasi struttura discreta su cui si possano applicare operazioni di conteggio o di enumerazione), dimostrazioni di identità e ottimizzazione di procedure come la moltiplicazione di matrici o problemi geometrici, verificabili tramite test formali o programmatici.

2. Infrastrutture e sistemi: sviluppo di soluzioni per lo scheduling nei data center, come ad esempio Borg, il sistema di gestione e scheduling dei cluster sviluppato internamente da Google a partire dai primi anni 2000. È una delle piattaforme fondamentali che permette a Google di gestire la propria enorme infrastruttura di data center in modo efficiente, automatizzato e scalabile.

3. Ingegneria AI e hardware: ottimizzazione del kernel per training di LLM e semplificazioni di circuiti hardware.

5.4. RISULTATI PRINCIPALI

AlphaEvolve ha mostrato risultati significativi in diversi domini:

Matematica e combinatoria: a: su oltre 50 problemi, il sistema ha riscoperto soluzioni allo stato dell'arte nel 75% dei casi e migliorato soluzioni esistenti in circa il 20% dei casi. Tra i risultati più rilevanti: un algoritmo per la moltiplicazione di matrici complesse 4×4 con 48 moltiplicazioni scalari e un miglioramento del kissing number in 11 dimensioni (593) (Google DeepMind, 2025). Il kissing number è il massimo numero di sfere identiche che possono toccarne un'altra senza sovrapporsi, in uno spazio di una certa dimensione.

Data center scheduling: una semplice euristica generata dal sistema ha permesso di recuperare, in produzione su Borg, circa lo 0,7% delle risorse globali di calcolo di Google, con un impatto continuativo superiore a un anno (Google DeepMind, 2025).

Training LLM: accelerazioni fino al 23% su kernel di moltiplicazione di matrici e fino al 32,5% su istruzioni GPU per l'attenzione, riducendo i tempi complessivi di training di circa l'1% sui modelli Gemini (Google DeepMind, 2025).

Hardware: ottimizzazione dei circuiti progettati in Verilog per le TPU (Tensor Processing Units, unità di elaborazione per l'intelligenza artificiale), ossia semplificare i circuiti in modo da occupare meno spazio, consumare meno energia e funzionare più velocemente (Google DeepMind, 2025).

5.5. VALIDITÀ, LIMITI E CRITICITÀ

Nonostante i risultati promettenti, AlphaEvolve presenta alcune limitazioni:

1. Dipendenza dai valutatori: il sistema funziona solo con metriche automatizzabili; compiti che richiedono valutazioni umane non sono gestibili.

2. Generalizzabilità: i risultati matematici sono verificati su problemi specifici; estenderli a classi più ampie richiede ulteriori test.

3. Costi e infrastruttura: l'esecuzione su larga scala richiede risorse computazionali elevate, dovute al parallelismo e al benchmarking continuo.

4. Trasparenza e replicabilità: sebbene DeepMind abbia pubblicato report e repository di verifica, parte del codice e dei workflow resta proprietaria, rendendo necessaria ulteriore validazione indipendente.

5.6. CONCLUSIONI

AlphaEvolve dimostra che, in contesti dove è possibile definire valutatori affidabili, l'uso di LLM all'interno di loop evolutivi permette di generare soluzioni innovative e ottimizzazioni concrete, dai problemi matematici alla gestione dell'infrastruttura. Il paradigma è generale, ma il successo rdipende dalla definizione chiara del dominio e dalla progettazione di metriche di valutazione robuste e affidabili. Di conseguenza, per ora il sistema pecca di astrattezza e di capacità nel fare valutazioni non deterministiche. Nonostante ciò, l'esperimento pone le basi per grandi prospettive future, dimostrando che, in domini verificabili, questi strumenti hanno ottenuto risultati migliori di quanto avrebbero fatto esperti umani del settore.

TABELLA .5

Componente	Descrizione	Metriche di selezione	Note
Seed program	Codice iniziale con regioni evolvibili marcate.	Copertura dei test, performance di base.	Definisce il perimetro del problema e i limiti iniziali.

Generatore LLM	Gruppo di modelli linguistici (come Gemini) che generano modifiche o correzioni al codice.	Percentuale di modifiche al codice che funzionano correttamente.	Bilancia l'esplorazione di nuove soluzioni e lo sfruttamento di quelle note.
Valutatore	Modulo che esegue test, benchmark e verifiche formali sulle varianti generate.	Correttezza, tempo di esecuzione, uso delle risorse.	Fornisce al sistema le informazioni su quanto bene funziona una soluzione.
Archivio soluzioni	Memoria delle migliori varianti ottenute durante l'evoluzione.	Ranking storico, grado di diversità.	Permette il riuso del contesto e delle soluzioni precedenti (*ereditarietà*).
Selettore	Meccanismo che sceglie quali varianti proseguono nel ciclo evolutivo.	Basato sulle soluzioni migliori, più nuove e più stabili.	Impedisce che le soluzioni diventino troppo simili tra loro
Orchestrazione	Sistema di esecuzione parallela e asincrona delle iterazioni.	Velocità di elaborazione	Riduce significativamente i tempi di ricerca e valutazione.

6. VERSO UN NUOVO ECOSISTEMA COMUNICATIVO

6.1 DAGLI ASPETTI TECNICI AGLI IMPATTI MEDIOLOGICI

Dopo aver analizzato il funzionamento tecnico dei Large Language Models e il caso specifico di AlphaEvolve, è necessario spostare il focus dall'ingegneria dei sistemi alle trasformazioni comunicative e socio-culturali che questi strumenti stanno generando. Gli agenti basati su LLM non rappresentano meri strumenti computazionali: essi agiscono come media nel senso più profondo del termine, ovvero come elementi che riorganizzano ruoli, dinamiche di potere e pratiche comunicative all'interno della società (McLuhan, 1964).

La transizione da una prospettiva puramente tecnica a una mediologica è cruciale per comprendere le implicazioni più ampie dell'integrazione degli agenti artificiali nei tessuti comunicativi contemporanei. Come sottolineato da Giovannetti e Miconi (2024), l'intelligenza artificiale non può essere analizzata unicamente attraverso parametri di efficienza o precisione, ma deve essere inquadrata come fenomeno culturale che ridefinisce le modalità attraverso cui gli esseri umani costruiscono significati, relazioni e identità.

6.2 IL MEDIUM È IL MESSAGGIO

6.2.1 Le quattro dimensioni della trasformazione mediale

La celebre formula mcluhaniana "il medium è il messaggio" (McLuhan, 1964) acquisisce una rilevanza particolare nell'era degli agenti artificiali. In questo contesto, la tetrade dei quattro effetti mediali, originariamente elaborata da Marshall ed Eric McLuhan e ripresa da Moriggi e Pireddu (2024), offre una chiave interpretativa utile a comprendere l'impatto dei nuovi sistemi esperti (intelligenti secondo i più audaci) attraverso quattro macro-effetti:

Estensione (Enhancement): Gli agenti LLM estendono significativamente le capacità umane di analisi, decisione e azione in contesti digitali complessi. A differenza dei precedenti strumenti informatici, che richiedevano competenze tecniche specifiche per essere utilizzati efficacemente, gli

agenti possono comunicare attraverso il linguaggio naturale, abbattendo le barriere tra le competenze tecniche e umanistiche. Questa estensione non è meramente quantitativa (maggiore velocità o precisione), ma qualitativa: gli agenti permettono di operare su scale di complessità precedentemente inaccessibili ai singoli individui.

Obsolescenza (Obsolescence): L'avvento degli agenti LLM rende obsolete alcune funzioni tradizionali di mediazione e controllo manuale. Le pratiche di scripting ripetitivo, la ricerca manuale di informazioni e numerose forme di intermediazione cognitiva vengono superate dalla capacità degli agenti di operare autonomamente. Tuttavia, questa obsolescenza non è uniforme: mentre alcune competenze tecniche perdono centralità, emergono nuove necessità legate alla capacità di interrogare, dirigere e supervisionare sistemi intelligenti.

Recupero (Retrieval): Paradossalmente, l'interazione con agenti artificiali recupera pratiche comunicative arcaiche legate alla delega e all'affidamento. Come nelle società pre-moderne, dove specifiche figure (sciamani, oracoli, consiglieri) detenevano accesso privilegiato a forme di conoscenza, gli agenti LLM reintroducono dinamiche di consultazione e interpretazione. Questo recupero include anche la riscoperta di pratiche di supervisione e controllo che richiamano relazioni gerarchiche tra maestro e apprendista, o tra esperto e novizio.

Ribaltamento (Reversal): Quando spinti ai loro estremi, gli agenti LLM possono produrre effetti opposti a quelli per cui sono stati concepiti. Il rischio di dipendenze sistemiche, lock-in tecnologico e perdita di agency umana rappresenta il punto di ribaltamento del medium. La promessa di autonomia e empowerment può trasformarsi in forme sottili ma pervasive di controllo e subordinazione, dove gli utenti perdono progressivamente la capacità di operare senza mediazione artificiale.

6.2.2 Implicazioni per l'ecologia mediale contemporanea

La tetrade offre uno schema analitico che va oltre l'hype tecnologico, permettendo di identificare i trade-off strutturali prodotti dall'adozione massiva di agenti artificiali. Come evidenziato da Scolari (2018) nella sua teoria dell'ecologia dei media, ogni nuovo medium non si limita ad aggiungersi all'ecosistema esistente, ma lo riorganizza completamente, creando nuove nicchie ecologiche e rendendone obsolete altre.

Nel caso degli agenti LLM, assistiamo alla formazione di un nuovo biotopo comunicativo dove convivono:

- Interazioni uomo-macchina sempre più sofisticate e naturalistiche.
- Dinamiche macchina-macchina che operano spesso al di sotto della soglia di consapevolezza umana.
- Forme ibride di cyborg intelligence dove è difficile distinguere i contributi umani da quelli artificiali.
- Nuove asimmetrie informative e cognitive tra chi ha accesso agli agenti e chi ne è escluso.

6.3 PARADIGMI TEORICI PER L'ANALISI DELL'ECOSISTEMA

Inquadrato l'impatto generazionale di questi sistemi con alcuni paradigmi comunicativi, è bene addentrarsi all'interno delle dinamiche che si possono venire a creare al contatto con tali oggetti, a partire dall'interazione discorsiva stessa.

6.3.1 La teoria dell'azione comunicativa di Habermas

La teoria dell'azione comunicativa di Jürgen Habermas (1981) fornisce un framework concettuale prezioso per analizzare le condizioni di validità del discorso in contesti mediati da agenti artificiali. Habermas riteneva che una comunicazione di reciproca intesa fosse il presupposto fondativo per una società funzionante, naturalmente integrata, basata su connessioni comunicative convergenti verso il mondo della vita (Lebenswelt).

Affinché si possa avere questa forma di comunicazione habermasiana sono necessari tre elementi:

1. Verità (Wahrheit): La corrispondenza delle affermazioni al mondo oggettivo.
2. Conformità normativa (Richtigkeit): La conformità delle norme al mondo sociale.
3. Sincerità (Wahrhaftigkeit): L'autenticità dell'espressione del mondo soggettivo.

L'introduzione di agenti artificiali in contesti comunicativi problematizza ciascuna di queste dimensioni:

Verità e agenti artificiali: Gli agenti LLM operano attraverso approssimazioni statistiche piuttosto che verifiche fattuali dirette. Il fenomeno delle "allucinazioni", ovvero la generazione di

informazioni plausibili ma fatalmente errate, solleva questioni fondamentali sulla possibilità di garantire standard di verità in comunicazioni mediate da IA. Come sottolineato da Chen et al. (2023), i modelli di reasoning più avanzati non sempre "dicono quello che pensano", creando disallineamenti tra processi interni e output comunicativi.

Conformità normativa: Gli agenti artificiali incorporano inevitabilmente bias e assunzioni normative derivanti dai loro dataset di training e dalle scelte progettuali dei loro creatori. Di conseguenza, potrebbe essere complicato generare un sistema che tenga conto nell'effettivo di una conformità normativa generale, nonostante i grandi modelli (GPT-4, Gemini, GPT-5) sembrino riuscire ad assorbire in maniera discretamente eterogenea le norme sociali del mondo della vita habermasiano. Ci sono anche questioni di accountability e trasparenza: come può essere garantita la correttezza di decisioni prese da sistemi i cui processi deliberativi sono opachi? È cruciale che le norme incorporate negli agenti siano socialmente legittimate e democraticamente controllabili, poiché, come mostrano Barocas et al. (2019), gli approcci alla fairness (proprietà di un sistema algoritmico di non produrre risultati discriminatori o ingiusti) richiedono scelte di valore che riflettano principi sociali e non soltanto criteri tecnici.

Sincerità e autenticità: La questione più complessa riguarda lo status epistemologico dell'autenticità in agenti artificiali. Un LLM può essere "sincero" se non possiede una soggettività nel senso tradizionale? Paradossalmente, gli agenti più sofisticati sono quelli che simulano meglio forme di sincerità umana, creando l'illusione di autenticità attraverso la performance linguistica piuttosto che attraverso stati interiori genuini. Terrei oltremodo aperta l'ipotesi che possano esistere forme di sincerità (autenticità) espresse da meccanismi fondanti differenti (non antropomorfici). Da valutare, quindi, con attenzione l'intendere un'emulazione di facoltà umane (con risultati pressoché uguali) intrinsecamente impossibilitata a possedere l'elemento della sincerità, laddove l'obiettivo sincero possa anche essere diametralmente opposto a quello voluto da Habermas, cioè un obiettivo strategico-comunicativo, al principio del quale vi sono intenti materiali.

6.3.2 La semiotica di Umberto Eco e i nuovi codici comunicativi

Il framework semiotico di Umberto Eco (1976, 1979) offre strumenti concettuali per analizzare le nuove pratiche semiotiche che emergono nell'interazione uomo-agente. In particolare, la

distinzione ecchiana tra segni naturali e artificiali, e la sua teoria dei codici, risulta particolarmente feconda per comprendere le dinamiche comunicative negli ambienti agente-centrici.

Ridefinizione della triade segnica: Eco riprende e ridefinisce operativamente la triade peirciana : l'indice, segno che rinvia al referente per contiguità (ad esempio il fumo che segnala il fuoco); l'icona, che simula o somiglia all'oggetto (come una mappa o un disegno); e il simbolo, segno arbitrario fondato su convenzioni, come la lingua naturale. Per Eco, “Icone”, “indici” e “simboli” non sono categorie ontologiche del segno, ma funzioni di codifica o di decodifica possibili, attivate da regole culturali e processi inferenziali. I modelli linguistici di grandi dimensioni operano esclusivamente al livello simbolico: manipolano stringhe linguistiche codificate senza un contatto diretto con il mondo. Eppure, l'output prodotto tende spesso a presentarsi come se avesse un valore indicale o iconico: un enunciato come "oggi piove a Roma" viene interpretato dall'utente come riferimento empirico, dotato di veridicità. Qui si origina una tensione semiotica: un sistema puramente sintattico-generativo può generare simboli che appaiono e vengono recepiti come indici o icone affidabili della realtà. Ne consegue che chi interagisce con tali modelli deve affinare nuove competenze ermeneutiche, capaci di distinguere quando l'agente sta semplicemente combinando segni e quando, invece, produce enunciati che possono essere trattati come referenziali.

Codici emergenti: Le interfacce conversazionali introducono nuovi codici comunicativi che combinano elementi del linguaggio naturale con logiche computazionali. Gli utenti esperti sviluppano progressivamente competenze di prompt engineering, ovvero la capacità di formulare richieste che ottimizzino le risposte degli agenti. Questi codici ibridi rappresentano forme emergenti di literacies digitali che ridefiniscono i confini tra competenze linguistiche e tecniche.

Cooperazione testuale con agenti non-umani: Eco sosteneva, nella sua teoria della cooperazione testuale, che autore e lettore dovessero mettere in atto un rapporto simmetrico all'interno del quale il lettore deve completare il significato inizialmente formalizzato dall'autore, colmando le lacune, interpretando le allusioni, le metafore, gli impliciti sensi. Una condivisione semiotica resa possibile dal comune denominatore di essere "umani", di condividere quel mondo della vita che Habermas tanto proponeva. Ampliando la teoria a un rapporto agente-umano, sorgerebbe il problema della condivisione di questo "mondo". Si potrebbe parlare di una cooperazione asimmetrica, dove l'agente-emulatore funge da specchio soggettivo dell'umano, cioè da oggetto che rigurgita un'idea di mondo che parrebbe essere un emulato del contesto

umano, che potrebbe tutt'al più rispecchiare in maniera particolare un contesto che, sicuramente, offrirebbe elementi diversi da quelli che potrebbe fornire un essere antropomorfo.

Eviterei però di parlare di un rapporto totalmente asimmetrico, in quanto è da stabilire (secondo questo lavoro) se il concetto di emulazione determini una reale differenza dal corrispettivo umano (emulato), (considerando l'emulazione di talune delle sue facoltà, se non tutte). L'idea è che, se il risultato finale - ad esempio la produzione di un ragionamento (tramite LLM) - venisse raggiunto, quanto potrebbe essere rilevante il fatto che i meccanismi che hanno portato questo risultato siano profondamente differenti? In quest'ottica, potremmo parimenti ritenere il rapporto uomo-agente come un rapporto simmetrico, semplicemente contenente caratteristiche e intenti diversi da quelli formulati nella teoria della cooperazione testuale di Eco. L'essere umano dovrà sviluppare competenze di consapevolezza, atte alla codificazione di una metodologia per cooperare nella forma più corretta con questi oggetti, che per quanto emulatori umani essi siano, devono essere trattati come oggetti con caratteristiche univoche, "intelligenti" o meno che siano.

6.3.3 L'ecologia dei media e la co-evoluzione tecnologica

Il paradigma dell'ecologia dei media, sviluppato da autori come Neil Postman (1970), fornisce una lente teorica per comprendere come gli agenti artificiali si inseriscano nell'ecosistema mediale contemporaneo.

Co-evoluzione media tradizionali/agenti intelligenti: Gli agenti LLM non sostituiscono semplicemente i media precedenti, ma entrano in rapporto di co-evoluzione con essi. La stampa, la televisione, internet e i social media si adattano alla presenza di agenti artificiali, mentre questi ultimi incorporano logiche derivanti dai media tradizionali. Ad esempio, gli agenti utilizzano pattern narrativi televisivi per strutturare le loro risposte, mentre i media tradizionali integrano sempre più contenuti generati artificialmente.

Nicchie ecologiche e specializzazione: Nel nuovo ecosistema, diversi tipi di agenti (generalistici come ChatGPT, specialistici come AlphaEvolve, embedded in applicazioni specifiche) occupano nicchie ecologiche differenti, sviluppando forme di specializzazione che riducono la competizione diretta. Questa specializzazione crea catene del valore complesse dove agenti diversi collaborano in pipeline di elaborazione dell'informazione.

Pressioni selettive e adattamento: L'ecosistema mediale agente-centrico è soggetto a pressioni selettive che favoriscono agenti con caratteristiche specifiche: capacità di mantenere

coerenza conversazionale, abilità di personalizzazione, compliance con norme etiche e legali. Queste pressioni selettive orientano lo sviluppo tecnologico in direzioni che non sono puramente determinate da considerazioni tecniche, ma da dinamiche ecosistemiche complesse.

6.4 DINAMICHE INTERAZIONALI NELL'ECOSISTEMA MULTI-AGENTE

6.4.1 Tipologie di interazione uomo-agente

L'analisi delle dinamiche comunicative negli ecosistemi agente-centrici richiede una tassonomia delle diverse modalità interazionali che si stanno consolidando:

Interazione consultiva: Il modello tradizionale dove l'agente funge da oracolo (cosa che assolutamente non è, ma identificante in questo caso il ruolo che assume nell'interazione con un gran numero di utenti non esperti) fornendo risposte a domande specifiche. Questa modalità, pur apparentemente semplice, nasconde complessità legate alla formulazione delle domande (prompt engineering) e all'interpretazione delle risposte. Gli utenti sviluppano progressivamente competenze meta-cognitive per valutare l'affidabilità e la pertinenza delle risposte fornite dall'agente.

Interazione collaborativa: Modalità emergente dove agente e umano co-producono contenuti, idee o soluzioni attraverso un processo iterativo. Questa forma di interazione richiede negoziazione continua dei ruoli, divisione dei compiti e sincronizzazione degli sforzi. Come evidenziato dagli studi sulla collaborazione uomo-IA, l'efficacia di questa interazione dipende dalla capacità dei partecipanti di sviluppare modelli mentali accurati delle competenze reciproche (Wang et al., 2020).

Interazione delegativa: Situazioni dove l'umano affida all'agente compiti complessi che richiedono autonomia decisionale. Questa modalità introduce questioni di accountability e controllo: come mantenere supervisione su processi delegati senza perdere i benefici dell'automazione? Il caso di AlphaEvolve, analizzato nel capitolo precedente, esemplifica questa modalità in contesti tecnici altamente specializzati.

Interazione supervisoria: Modalità dove l'umano monitora e coordina l'attività di molteplici agenti, intervenendo quando necessario per correggere errori o reindirizzare gli sforzi. Questa forma di interazione richiede competenze di orchestration che combinano conoscenze tecniche e manageriali.

6.4.2 Comunicazione agente-agente: verso ecosistemi autonomi

Una delle trasformazioni più significative introdotte dai sistemi multi-agente riguarda lo sviluppo di forme di comunicazione diretta tra agenti artificiali, che operano spesso al di sotto della soglia di consapevolezza umana.

Protocolli di coordinamento: Gli agenti sviluppano linguaggi specializzati per coordinarsi efficacemente. Questi protocolli, ottimizzati per l'efficienza computazionale piuttosto che per la comprensibilità umana, creano forme di comunicazione opaca che sfuggono al controllo diretto degli utenti. Come osservato da Foerster et al. (2016), agenti che evolvono linguaggi propri possono sviluppare strategie comunicative che non sono facilmente interpretabili dagli esseri umani.

Emergenza di gerarchie e specializzazioni: Negli ecosistemi multi-agente, emergono spontaneamente strutture gerarchiche e forme di specializzazione funzionale. Alcuni agenti assumono ruoli di coordinamento, altri si specializzano in compiti specifici. Queste dinamiche organizzative, che ricordano principi dell'auto-organizzazione biologica, sollevano questioni sulla governabilità di sistemi artificiali complessi. Un caso paradigmatico di auto-organizzazione nei sistemi artificiali è rappresentato dalla swarm robotics. In questo ambito, gruppi di robot relativamente semplici, dotati di sensori di base e programmati con regole locali elementari, vengono messi a interagire all'interno di un ambiente condiviso. Anche in assenza di un coordinamento centrale, tali sistemi sono in grado di dar luogo a comportamenti collettivi complessi e adattivi: per esempio, i robot possono cooperare per formare catene utili al trasporto di oggetti, suddividere l'esplorazione dello spazio in sotto-aree o mantenere configurazioni stabili di gruppo. In queste dinamiche, la distribuzione dei ruoli e l'emergere di forme gerarchiche non sono pianificate dall'esterno, ma scaturiscono spontaneamente dall'interazione locale tra le singole unità, in maniera analoga a quanto accade nelle colonie di insetti (Şahin, 2005).

Formazione di coalizioni e competizione: Gli agenti possono sviluppare alleanze temporanee per raggiungere obiettivi specifici, o entrare in competizione per risorse computazionali limitate. Queste dinamiche, studiate nella teoria dei giochi computazionale (Nisan et al., 2007),

introducono elementi di imprevedibilità che complicano la progettazione e il controllo di sistemi multi-agente.

6.5 IMPATTI SOCIO-CULTURALI E TRASFORMAZIONI IDENTITARIE

6.5.1 Questioni identitarie e relazionali

L'interazione prolungata con agenti artificiali sta producendo trasformazioni nelle modalità attraverso cui gli esseri umani costruiscono identità e relazioni.

Anthropomorphic relationships: Gli esseri umani tendono a sviluppare relazioni quasi-sociali con agenti artificiali, attribuendo loro personalità, intenzioni e stati emotivi. Questo fenomeno, studiato dalla psicologia cognitiva, solleva questioni sulla natura delle relazioni autentiche e sui rischi di sostituzione di relazioni umane con simulacri artificiali (Turkle, 2011).

Identity boundaries: L'uso prolungato di agenti artificiali come estensioni cognitive può produrre confusione sui confini dell'identità personale. Quando un agente artificiale contribuisce significativamente alla produzione di idee o contenuti, emergono questioni sulla proprietà intellettuale e sull'autenticità della creatività individuale.

Intelligenza collettiva: Gli ecosistemi agente-centrici facilitano forme di intelligenza collettiva dove contributi umani e artificiali si intrecciano in modi difficilmente districabili. Queste forme di cognizione distribuita aprono possibilità inedite di problem-solving collaborativo, ma sollevano questioni sulla responsabilità individuale e sulla accountability collettiva.

6.6 GOVERNANCE E REGOLAZIONE DEI SISTEMI AGENTE-CENTRICI

6.6.1 Sfide di governance

La governance di ecosistemi comunicativi che includono agenti artificiali presenta sfide inedite che richiedono nuovi framework normativi e istituzionali.

Trasparenza e explainability: Come garantire che le decisioni prese da agenti artificiali siano comprensibili e controllabili dagli esseri umani? I sistemi basati su LLM, operando attraverso reti neurali complesse, producono spesso decisioni che sono difficili da spiegare anche per i loro creatori.

Accountability distribuita: In ecosistemi dove molteplici agenti interagiscono autonomamente, diventa complesso stabilire responsabilità per outcomes non desiderati. Chi è responsabile quando un'azione problematica emerge dall'interazione tra agenti diversi, creati da organizzazioni diverse, operanti secondo logiche diverse?

Partecipazione democratica: Come assicurare che lo sviluppo di agenti artificiali sia soggetto a controllo democratico? La concentrazione dello sviluppo di LLM in poche grandi corporation crea rischi di accentramento del potere che possono minare principi democratici.

6.6.2 Principi per una governance responsabile

Diverse organizzazioni internazionali e istituti di ricerca stanno sviluppando principi per una governance responsabile dell'IA. Tra i principi più consolidati:

- Human-in-the-loop: Mantenere sempre un livello di supervisione e controllo umano su decisioni critiche, anche in sistemi altamente automatizzati.
- Reversibility: Garantire che le azioni intraprese da agenti artificiali possano essere annullate o corrette quando necessario.
- Auditability: Sviluppare sistemi di logging e monitoring che permettano di tracciare le decisioni prese da agenti artificiali e di identificare eventuali problemi.

- Fairness e non-discrimination: Assicurare che gli agenti artificiali non perpetuino o amplifichino bias discriminatori presenti nei dati di training.

6.7 SCENARI FUTURI E IMPLICAZIONI A LUNGO TERMINE

6.7.1 Possibili traiettorie evolutive

L'analisi delle tendenze attuali permette di identificare alcune possibili traiettorie evolutive per gli ecosistemi agente-centrici:

Scenario 1 – Integrazione armoniosa: Gli agenti artificiali si integrano gradualmente nei tessuti sociali esistenti, potenziando le capacità umane senza creare distruzioni maggiori. Questo scenario richiede una governance proattiva e un adattamento graduale delle istituzioni esistenti.

Scenario 2 – Polarizzazione e frammentazione: L'accesso diseguale agli agenti artificiali più avanzati crea nuove forme di stratificazione sociale tra "agente-enhanced" e "agente-excluded". Questo scenario richiederebbe interventi redistributivi per evitare l'amplificazione delle disuguaglianze esistenti.

Scenario 3 – Trasformazione radicale: Gli agenti artificiali diventano così centrali nei processi comunicativi e decisionali da produrre trasformazioni qualitative delle forme di organizzazione sociale. Questo scenario richiederebbe ripensamenti fondamentali delle categorie politiche e sociali esistenti.

6.7.2 Implicazioni per la ricerca futura

L'analisi condotta suggerisce diverse direzioni per la ricerca futura:

- Studi longitudinali: Necessità di studi empirici a lungo termine sugli effetti dell'esposizione prolungata ad agenti artificiali su identità, relazioni e competenze cognitive.
- Metodologie interdisciplinari: Sviluppo di metodologie che integrino prospettive tecniche, psicologiche, sociologiche e filosofiche per comprendere la complessità degli ecosistemi agente-centrici.

- Participatory design: Coinvolgimento di diverse categorie di stakeholder nella progettazione di agenti artificiali per assicurare che rispondano a bisogni sociali reali piuttosto che a logiche puramente tecnologiche o commerciali.

Co-evoluzione tra umani e sistemi di IA: L'evoluzione dell'IA non avviene in isolamento, ma in co-dipendenza con le trasformazioni dell'umano, dei contesti istituzionali e delle pratiche socioeconomiche. L'orizzonte dell'human enhancement apre la strada a forme di integrazione più strette tra cognizione biologica e artificiale. Interfacce cervello-computer sempre più sofisticate mirano a una comunicazione bidirezionale a bassa latenza tra attività neurale e sistemi computazionali; le tecniche di potenziamento cognitivo assistito dall'IA promettono supporti personalizzati per memoria, attenzione, decisione e creatività; l'emergere di configurazioni di cognizione distribuita prefigura sistemi ibridi umano-artificiale, nei quali capacità complementari si orchestrano in reti collaborative.

Tali sviluppi possono indurre nuove pressioni selettive di natura culturale, professionale e organizzativa. È plausibile una spinta verso specializzazioni cognitive complementari all'IA, nelle quali gli umani si concentrano su ambiti meno standardizzabili, ad alto contenuto di giudizio, ambiguità e contesto. Parallelamente, la rilevanza della competenza sociale e dell'intelligenza emotiva potrebbe accrescersi, in quanto dimensioni difficilmente riducibili a pattern puramente statistici e centrali per il coordinamento, la negoziazione e la costruzione di fiducia. Infine, le capacità creative e artistiche, nella misura in cui integrano sensibilità estetica, originalità e significato culturale, potrebbero assumere un valore differenziale crescente in ecosistemi dove la produzione informazionale è ampiamente automatizzata.

6.7.3 Identità, umanità e scenari futuri

L'avanzamento dell'IA invita a una ridefinizione pluralistica del concetto di intelligenza. Accanto a metriche prestazionali, si affermano concezioni che riconoscono la legittimità di forme molteplici di intelligenza, includendo dimensioni embodied, sociali, affettive e situate. In tale cornice, acquisiscono centralità definizioni basate sui valori e sugli obiettivi: l'intelligenza non solo come potenza di calcolo o efficienza predittiva, ma come capacità di promuovere fini umani desiderabili in contesti normativi e culturali specifici. Ne discende una prospettiva di ruoli complementari tra intelligenza umana e artificiale, dove la co-progettazione di sistemi socio-tecnici mira a massimizzare sinergie e minimizzare conflitti.

Allo stesso tempo, emerge la questione della preservazione dell'unicità umana. Tre dimensioni sono frequentemente indicate come difficilmente replicabili: l'esperienza fenomenologica, ossia la qualità soggettiva e vissuta dell'esperienza cosciente; l'agenzia morale, intesa come responsabilità etica genuina e imputabilità; e la capacità di generare significato esistenziale, che connette scopi, valori e narrazioni in progetti di vita. La tutela e la valorizzazione di queste dimensioni diventano un obiettivo normativo strategico nei processi di integrazione con l'IA.

Sulla base di questi elementi, si delineano tre macro-scenari, non mutuamente esclusivi e plausibilmente coesistenti in diverse aree o fasi temporali.

Uno scenario di simbiosi descrive una collaborazione armoniosa tra intelligenze umane e artificiali, con sistemi co-progettati per amplificare capacità umane, sostenere decisioni complesse e generare valore pubblico.

Uno scenario di displacement ipotizza la sostituzione graduale dell'intelligenza umana in numerosi domini operativi, con impatti occupazionali, distributivi e identitari che richiedono politiche attive di transizione, formazione e protezione sociale.

Un terzo scenario, di trascendenza, contempla l'emergere di forme post-umane di intelligenza, in cui i confini tra biologico e artificiale si ridefiniscono radicalmente, sollevando interrogativi ontologici e etico-politici di nuova generazione.

7. CONFRONTO TRA FORME DI INTELLIGENZA

Abbiamo analizzato i meccanismi interni dei LLMs, sviscerando quanto più possibile questa fantomatica black box, facendo chiarezza sui fantasmi che aleggiavano intorno ad essa, e immergendola in rapporto all'ecosistema circostante, adottando una prospettiva organica e contestualizzante, forti di dare un idea quanto più " reale", dell'influenza di questi oggetti, in uno scambio multidirezionale agente-centrico. Questi oggetti sono stati tecnicamente e socio-comunicativamente immersi nel quadro attuale delle cose, facendo sì che sia rimasto introdurre la lente forse più affascinante, ambiziosa, distopica o utopica di questo lavoro; misurare,comprendere, intuire, quelle forme di intelligenza, coscienza, che pare possano presentarsi (nello stadio attuale dell'arte o in uno stadio futuro ancora non materialmente raggiungibile) in questi oggetti, e mostrare, di conseguenza, quei comportamenti che solo rivediamo attraverso l'essere antropomorfo. Ne consegue una evidente comparazione con quelle forme di intelligenza,coscienza, già conosciute, affinchè si possano stabilire i parametri e gli elementi ai quali specchiarsi per formulare conclusioni mirate e contestualizzate, sulla base di una metodologia/definizione/metrica attualizzata. Ci serviremo di diversi studi fondativi, che faranno uso di diverse teorie computazionali e della mente, affinché ci possano essere tutti gli strumenti metodologici per teorizzare in concreto, cioè evidenziando possibilità sperimentali attuabili al fine di raggiungere l'obiettivo conclamato. Il punto di partenza sarà una definizione concorde di intelligenza suggeritaci da Gignac e Szodorai (2024).

7.1. Definizioni operative e quadri teorici di riferimento

7.1.1. Il problema definitorio dell'intelligenza

La definizione di intelligenza rappresenta una delle sfide più complesse e dibattute nelle scienze cognitive contemporanee. Come sottolineato da Gignac e Szodorai (2024), l'assenza di consenso definitorio non è meramente un problema accademico, ma ha implicazioni pratiche fondamentali per lo sviluppo e la valutazione di sistemi artificiali che aspirano a replicare o superare le capacità cognitive umane.

Intelligenza umana: Seguendo la proposta di Gignac e Szodorai (2024), possiamo definire l'intelligenza umana come "la massima capacità di raggiungere un obiettivo nuovo tramite processi percettivo-cognitivi". Questa definizione enfatizza due aspetti cruciali: la *novità* del

problema (distinguendo l'intelligenza dall'expertise in domini familiari) e la natura *percettivo-cognitiva* dei processi sottostanti (radicati nell'esperienza corporea e nella cognizione incarnata).

Intelligenza artificiale: In parallelo, l'intelligenza artificiale può essere definita come "la massima capacità di un sistema artificiale di raggiungere un obiettivo nuovo tramite procedure computazionali". La distinzione tra "processi percettivo-cognitivi" e "procedure computazionali" non è meramente terminologica: essa sottolinea differenze fondamentali nei substrati, nei meccanismi e nei vincoli operativi dei due tipi di intelligenza.

7.1.2. Separazione concettuale: intelligenza vs competenze correlate

Una delle confusioni più frequenti negli studi comparativi riguarda la sovrapposizione tra intelligenza, competenza, expertise e adattamento. La chiarificazione di queste distinzioni è essenziale per evitare confronti spuri:

Intelligenza vs Expertise: L'intelligenza si manifesta specificamente nella capacità di affrontare problemi *nuovi*, mentre l'expertise riguarda l'abilità acquisita su compiti *noti*. Un grande maestro di scacchi può dimostrare expertise straordinaria nel suo dominio senza necessariamente possedere intelligenza generale superiore. Analogamente, un LLM può eccellere in compiti specifici per cui è stato addestrato senza dimostrare vera intelligenza generale.

Intelligenza vs Adattamento: L'adattamento include processi non necessariamente cognitivi (risposte riflesse, modificazioni fisiologiche) che permettono di accomodarsi a vincoli ambientali. L'intelligenza rappresenta una forma specifica di adattamento che opera attraverso elaborazione di informazioni e problem-solving deliberato.

Implicazioni per la valutazione di sistemi IA: Queste distinzioni sono cruciali per evitare il *data leakage* nella valutazione di sistemi IA. Se un modello è stato esposto durante l'addestramento a versioni del problema che deve risolvere in fase di test, stiamo misurando expertise piuttosto che intelligenza. Questo problema è particolarmente rilevante per i LLM addestrati su dataset che includono vaste porzioni della conoscenza umana.

7.2 MODELLI TEORICI DELL'INTELLIGENZA UMANA

La comprensione dell'intelligenza umana richiede l'adozione di un framework teorico solido. In questa sezione, si presenta innanzitutto il modello Cattell-Horn-Carroll (CHC), considerato il framework di riferimento per questo lavoro grazie alla sua solida base psicometrica e alla sua struttura gerarchica verificabile empiricamente. Successivamente, si esaminano teorie alternative (Gardner e Sternberg) che, pur presentando limiti metodologici, offrono prospettive complementari sulla natura multidimensionale dell'intelligenza. La sezione si conclude con una sintesi critica che integra queste prospettive teoriche e ne discute le implicazioni per lo sviluppo e la valutazione dell'intelligenza artificiale.

7.2.1 Il modello CHC come framework di riferimento

Il modello Cattell-Horn-Carroll (CHC) rappresenta uno dei framework più consolidati per la comprensione dell'architettura dell'intelligenza umana. Sviluppato attraverso decenni di ricerca psicometrica, il modello CHC propone una struttura gerarchica a tre livelli:

Fattore g (intelligenza generale): Al vertice della gerarchia, il fattore g rappresenta la varianza condivisa tra diverse abilità cognitive. La sua esistenza è supportata da evidenze robuste che mostrano correlazioni positive tra performance in compiti cognitivi diversi (fenomeno noto come "positive manifold"). Il fattore g predice outcome importanti come successo accademico, performance lavorativa e risultati nella vita.

Abilità cognitive ampie: Il livello intermedio include circa 8-10 abilità cognitive ampie, tra cui:

- Gf (fluid intelligence): Capacità di ragionamento in situazioni nuove, indipendentemente da conoscenze acquisite
- Gc (crystallized intelligence): Conoscenze e abilità acquisite attraverso l'esperienza culturale
- Gv (visual processing): Elaborazione di informazioni visuo-spaziali
- Ga (auditory processing): Elaborazione di informazioni uditive
- Gsm (short-term memory): Memoria a breve termine e memoria di lavoro
- Glr (long-term retrieval): Capacità di immagazzinamento e recupero dalla memoria a lungo

termine

• Gs (processing speed): Velocità di elaborazione cognitiva

• Gq (quantitative reasoning): Abilità matematiche e quantitative

Abilità specifiche: Il livello più basso include oltre 70 abilità cognitive specifiche che contribuiscono alla performance in compiti particolari.

La struttura gerarchica del modello CHC, validata attraverso analisi fattoriali su vasti campioni, lo rende particolarmente adatto come riferimento teorico per questo lavoro. Tuttavia, è utile considerare anche prospettive teoriche alternative che, pur presentando maggiori criticità metodologiche, evidenziano aspetti specifici dell'intelligenza umana meritevoli di attenzione.

7.2.2 Teorie complementari: prospettive alternative sull'intelligenza

Oltre al modello CHC, la letteratura psicologica ha proposto approcci alternativi che enfatizzano la natura multidimensionale dell'intelligenza. Sebbene queste teorie presentino limiti empirici significativi (in particolare, la difficoltà a dimostrare l'indipendenza dei domini proposti), esse colgono un elemento concettuale rilevante: l'esistenza di domini specifici e multipli di intelligenza, coerente con la struttura a livelli del modello CHC.

La teoria delle intelligenze multiple di Gardner:

Howard Gardner (1983, 2011) propone l'esistenza di intelligenze multiple relativamente indipendenti:

• Linguistico-verbale

• Logico-matematica

• Spaziale

• Musicale

• Corporeo-cinestetica

• Interpersonale

• Intrapersonale
• Naturalistica

Sebbene popolare in contesti educativi, la teoria di Gardner riceve supporto empirico limitato. Come evidenziato nello studio di Gignac e Szodorai (2024), le diverse "intelligenze" mostrano correlazioni positive, suggerendo l'esistenza di un fattore generale sottostante. Tali evidenze empiriche contraddicono l'assunzione centrale della teoria di Gardner, secondo cui le diverse intelligenze sarebbero indipendenti fra loro. Nonostante questa critica fondamentale, la teoria di Gardner ha il merito di sottolineare la varietà di competenze cognitive che caratterizzano l'essere umano, anche se la loro organizzazione appare meglio descritta dalla struttura gerarchica del modello CHC piuttosto che da moduli totalmente indipendenti.

Intelligenza triarchica di Sternberg

Robert Sternberg (1985) propone tre aspetti complementari dell'intelligenza:
• **Analitica:** Capacità di analizzare, valutare e confrontare informazioni
• **Creativa:** Capacità di creare, inventare e scoprire soluzioni originali
• **Pratica:** Capacità di applicare conoscenze in contesti reali

> La teoria triarchica di Sternberg offre una prospettiva interessante enfatizzando componenti dell'intelligenza (in particolare quella creativa e pratica) spesso trascurate nei test psicometrici tradizionali. Tuttavia, anche questa teoria incontra difficoltà nella verifica empirica dell'indipendenza dei tre componenti. Ciononostante, l'enfasi di Sternberg sull'intelligenza pratica e contestuale evidenzia un aspetto cruciale: l'intelligenza non si manifesta solo in contesti astratti o accademici, ma anche nell'adattamento efficace a situazioni reali complesse.

7.2.3 Sintesi critica e implicazioni per l'intelligenza artificiale

Punti di convergenza e divergenza:

L'analisi comparativa dei modelli teorici presentati rivela sia convergenze che divergenze significative:

Convergenze:

1. **Multidimensionalità:** Tutti i modelli riconoscono che l'intelligenza non è un'entità monolitica, ma comprende componenti multiple (abilità ampie nel modello CHC, intelligenze multiple in Gardner, componenti triarchiche in Sternberg).
2. **Specificità di dominio:** L'esistenza di competenze specifiche per diversi domini cognitivi è un elemento comune, coerente con i livelli intermedi e inferiori della gerarchia CHC.
3. **Rilevanza contestuale:** Sia Sternberg (con l'intelligenza pratica) che Gardner (con le diverse intelligenze applicate a contesti reali) sottolineano l'importanza del contesto, aspetto che il modello CHC integra attraverso l'interazione tra abilità generali e specifiche.

Divergenze:

1. **Esistenza del fattore g:** Il modello CHC, supportato da robuste evidenze psicometriche (Gignac & Szodorai, 2024), sostiene l'esistenza di un fattore generale di intelligenza. Gardner e Sternberg, al contrario, enfatizzano l'indipendenza dei diversi componenti, posizione che tuttavia trova scarso supporto empirico nelle correlazioni positive osservate tra diverse abilità cognitive.
2. **Validazione empirica:** Il modello CHC si fonda su decenni di analisi fattoriali e validazioni psicometriche sistematiche. Le teorie di Gardner e Sternberg, pur concettualmente stimolanti, presentano maggiori difficoltà di operazionalizzazione e verifica empirica rigorosa.

Implicazioni per l'intelligenza artificiale:

L'integrazione di queste prospettive teoriche offre importanti spunti per lo sviluppo e la valutazione dell'intelligenza artificiale:

1. **Architettura gerarchica:** Il modello CHC suggerisce che un'IA veramente generale dovrebbe replicare non solo capacità cognitive specifiche, ma anche la loro organizzazione gerarchica e le interazioni dinamiche tra diversi livelli di elaborazione. I sistemi AI attuali eccellono tipicamente in domini ristretti (abilità specifiche) ma faticano a dimostrare un'intelligenza generale trasversale (fattore g).

2. **Valutazione multidimensionale:** Le prospettive di Gardner e Sternberg evidenziano la necessità di valutare l'IA su una gamma più ampia di competenze, includendo non solo abilità analitiche ma anche creative, pratiche e contestuali. I test di intelligenza tradizionalmente applicati all'IA potrebbero non catturare l'intera gamma di capacità rilevanti.
3. **Terminologia appropriata:** In virtù della difficoltà di sottoporre i sistemi artificiali a test di intelligenza completi e rigorosi (come quelli utilizzati per validare il modello CHC), appare più cauto e metodologicamente corretto parlare al momento di **sistemi esperti** piuttosto che di sistemi **intelligenti**, nonostante la terminologia comune utilizzi il termine "intelligenza artificiale" anche per sistemi altamente specializzati. Concordando con la definizione di intelligenza offerta da Gignac e Szodorai (2024), un sistema dovrebbe dimostrare non solo competenza in compiti specifici, ma anche capacità di generalizzazione efficiente e performance correlate attraverso diversi domini cognitivi per essere definito "intelligente" in senso pieno.
4. **Obiettivi di sviluppo futuro:** Per progredire verso un'intelligenza artificiale generale (AGI), sarà necessario:
    1. Sviluppare architetture che integrino abilità ampie diverse (Gf, Gc, Gv, ecc.) in modo gerarchico e coordinato
    2. Creare benchmark di valutazione che vadano oltre la performance in singoli compiti, testando la capacità di transfert e generalizzazione
    3. Considerare non solo l'intelligenza analitica ma anche quella creativa e pratica, particolarmente rilevanti per l'adattamento a situazioni nuove e complesse

In sintesi, il modello CHC fornisce il framework teorico più solido per comprendere e valutare l'intelligenza, sia umana che artificiale, mentre le teorie complementari di Gardner e Sternberg arricchiscono questa comprensione evidenziando dimensioni spesso trascurate nei test psicometrici tradizionali ma cruciali per una valutazione completa delle capacità cognitive.

7.3. FATTORE G NEI LLMS

7.3.1 Fattori generali emergenti nei LLM

Recenti studi (Burnell et al., 2023) hanno applicato metodologie psicometriche ai Large Language Models (LLMs), con l'obiettivo di indagare la possibile esistenza di strutture cognitive analoghe a quelle osservabili nell'intelligenza umana. I risultati mostrano elementi di convergenza interessanti:

Un fattore generale di intelligenza artificiale (g-AI):

Le correlazioni positive tra le prestazioni dei modelli in compiti differenti suggeriscono la presenza di un fattore generale, concettualmente analogo al g-factor umano. Oltre al fattore generale, emergono cluster di abilità correlate (ad esempio linguistico-discorsive, matematiche, di ragionamento logico), che riflettono una certa articolazione interna delle competenze. Le prestazioni dei modelli mostrano regolarità predittive in funzione della loro scala (numero di parametri e dati di addestramento), suggerendo un parallelismo con lo sviluppo delle capacità cognitive globali.

Tuttavia, a dispetto di queste analogie strutturali, si evidenziano anche differenze qualitative importanti rispetto all'intelligenza umana:

Stabilità temporale: L'intelligenza umana tende a mantenere stabilmente le conoscenze acquisite, mentre i LLMs sono vulnerabili al fenomeno del catastrophic forgetting, ovvero la perdita rapida e drastica di competenze pregresse quando il modello viene riaddestrato su nuovi compiti o dataset.

Transfer learning: Gli esseri umani mostrano una maggiore flessibilità nell'applicare conoscenze acquisite in un dominio cognitivo a contesti diversi; i LLMs, al contrario, mostrano trasferimenti più limitati e dipendenti dalla similarità tra compiti.

Meta-cognizione: L'uomo è in grado di monitorare e regolare in maniera consapevole i propri processi di pensiero (capacità meta-cognitive), mentre nei modelli artificiali tali meccanismi risultano ancora rudimentali o assenti.

7.4. CONFRONTI SISTEMATICI

Dopo aver esaminato i modelli teorici dell'intelligenza umana e i principali strumenti di valutazione psicometrica, è necessario confrontare in modo sistematico le capacità cognitive specifiche di esseri umani e sistemi artificiali. Mentre le sezioni precedenti hanno fornito un framework concettuale generale, questa sezione adotta un approccio comparativo più granulare, analizzando tre domini cognitivi fondamentali: l'elaborazione linguistica (7.4.1), il ragionamento e problem solving (7.4.2), e l'apprendimento e adattabilità (7.4.3).

La scelta di questi tre domini non è casuale: essi rappresentano competenze centrali sia nell'intelligenza umana (come evidenziato dalle abilità ampie del modello CHC: Gc per il linguaggio, Gf per il ragionamento, Glr per l'apprendimento) sia nelle prestazioni dei modelli linguistici di grandi dimensioni. Questo confronto sistematico permetterà di identificare con precisione aree di convergenza, divergenza qualitativa e complementarità tra intelligenza biologica e artificiale, fornendo una base empirica per valutare l'effettiva "intelligenza" dei sistemi AI contemporanei.

7.4.1. Elaborazione linguistica e comprensione

L'elaborazione linguistica rappresenta il dominio in cui i Large Language Models hanno ottenuto i risultati più impressionanti e, al contempo, quello che solleva le questioni più complesse sul rapporto tra competenza formale e comprensione genuina. Il linguaggio è infatti una capacità distintivamente umana che integra componenti sintattiche, semantiche, pragmatiche e prosodiche, ancorandosi profondamente all'esperienza corporea e sociale. Confrontare l'elaborazione linguistica umana con quella dei LLM permette di esplorare in che misura la padronanza statistica dei pattern linguistici possa avvicinarsi o divergere dalla comprensione basata sull'esperienza vissuta. Questo confronto si rivela cruciale per valutare se i LLM possiedano una forma di "comprensione" o si limitino a una sofisticata simulazione sintattico-statistica.

Capacità umane: L'elaborazione linguistica umana integra:

- *Sintassi*: Regole grammaticali spesso implicite e automatiche
- *Semantica*: Significati ancorati all'esperienza percettiva e corporea
- *Pragmatica:* Uso del linguaggio in contesti comunicativi specifici
- *Prosodia*: Informazioni paralinguistiche (tono, ritmo, enfasi)

Capacità dei LLM: I modelli linguistici dimostrano competenze impressionanti in:

- *Fluenza sintattica*: Generazione di testo grammaticalmente corretto
- *Coerenza semantica*: Mantenimento della coerenza tematica su testi lunghi
- *Adattamento stilistico*: Capacità di variare registro e stile comunicativo
- *Multilinguismo:* Competenze in lingue multiple senza addestramento esplicito

Differenze qualitative:

- *Grounding*: Il linguaggio umano è ancorato all'esperienza sensorimotoria; quello dei LLM è puramente statistico-simbolico
- *Intenzionalità*: Gli umani usano il linguaggio con intenzioni comunicative; i LLM simulano intenzionalità senza possederla
- *Embodiment*: La comprensione umana è influenzata dall'esperienza corporea; quella dei LLM è disembodied (sebbene una forma di IA embodied potrebbe migliorare il risultato)

7.4.2. Ragionamento e problem solving

Se l'elaborazione linguistica rappresenta il terreno delle conquiste più evidenti dei LLM, il ragionamento e il problem solving costituiscono il banco di prova per valutare capacità cognitive di ordine superiore. Il ragionamento umano, come evidenziato dal fattore Gf (fluid intelligence) nel modello CHC, implica la capacità di affrontare problemi nuovi, identificare relazioni astratte, costruire inferenze causali e controfattuali, e trasferire principi di soluzione tra domini diversi. Queste abilità sono

centrali nell'intelligenza generale e rappresentano un test particolarmente impegnativo per i sistemi artificiali.

Sviluppi recenti nell'architettura dei LLM, come il chain-of-thought reasoning e tecniche di prompting avanzate, hanno mostrato capacità emergenti promettenti in questo dominio. Tuttavia, rimangono questioni aperte sulla robustezza di queste capacità, sulla loro dipendenza dalla formulazione specifica dei problemi, e sulla loro generalizzabilità oltre i pattern visti durante l'addestramento. Questo sotto-capitolo esamina sistematicamente le somiglianze e le differenze qualitative tra il ragionamento umano e quello artificiale, con particolare attenzione alle limitazioni che ancora separano i LLM da una vera intelligenza generale.

Servendoci di ulteriori studi in combinazione con Gignac e Szodorai (2024) possiamo dedurre delle evidenze comparative tra il ragionamento umano e quello artificiale.

7.4.3 Apprendimento e adattabilità

Oltre alla capacità di elaborare linguaggio e risolvere problemi, una caratteristica fondamentale dell'intelligenza è la capacità di apprendere dall'esperienza e adattarsi a contesti nuovi. L'apprendimento rappresenta, infatti, il meccanismo attraverso cui l'intelligenza si sviluppa e si perfeziona nel tempo. Nel modello CHC, questa dimensione è catturata dalle abilità Glr (long-term storage and retrieval) e dalla capacità di trasferire conoscenze tra domini (aspetto centrale della fluid intelligence).

Il confronto tra apprendimento umano e artificiale rivela differenze qualitative profonde: mentre gli esseri umani sono capaci di apprendimento rapido da pochi esempi (few-shot learning), di meta-apprendimento che migliora l'efficienza di apprendimenti futuri, e di accumulo continuo di conoscenze senza dimenticare quelle precedenti (lifelong learning), i LLM affrontano sfide significative in queste aree. La loro modalità di apprendimento è tipicamente concentrata in una fase di addestramento massiva su enormi dataset, seguita da un deployment in cui i

parametri rimangono statici (sebbene tecniche come in-context learning e fine-tuning offrano forme di adattamento limitate).

Questo sotto-capitolo esplora le differenze tra l'apprendimento biologico, continuo e efficiente in termini di dati, e quello artificiale, tipicamente massivo e statico, analizzando le implicazioni di queste divergenze per lo sviluppo di sistemi AI veramente adattabili e autonomi.

Apprendimento umano:

Caratteristiche distintive:

- *Few-shot learning*: Capacità di apprendere concetti da pochissimi esempi.

- *Learning to learn*: Meta-apprendimento che migliora l'efficienza di apprendimenti futuri.

- *Lifelong learning*: Accumulo di conoscenze senza interferenza catastrofica.

- *Transfer learning*: Applicazione flessibile di conoscenze acquisite in contesti nuovi.

Apprendimento nei LLM:

- *In-context learning*: Apprendimento attraverso esempi forniti nel prompt senza aggiornamento dei parametri.

- *Fine-tuning*: Specializzazione su compiti specifici attraverso addestramento mirato.

- *Multi-task learning*: Apprendimento simultaneo di compiti multipli.

- *Meta-learning*: Emergenza di capacità di apprendimento generale.

Sfide comparative:

- *Data efficiency*: Gli umani apprendono con molti meno esempi.

- *Stabilità*: L'apprendimento umano è meno soggetto a dimenticanza catastrofica.

- *Motivazione intrinseca*: Gli umani sono guidati da curiosità e motivazioni intrinseche (anche se teniamo aperta l'ipotesi che possano esserci motivazioni intrinseche in formato differente non riconoscibile per come intendiamo noi le motivazioni interiori) .

7.5. SUBSTRATI NEUROBIOLOGICI VS SUBSTRATI COMPUTAZIONALI

Dopo aver confrontato le capacità cognitive a livello funzionale (elaborazione linguistica, ragionamento, apprendimento), è essenziale esaminare anche il livello implementativo: i substrati fisici che rendono possibili tali capacità. Questa distinzione richiama la nota differenza in filosofia della mente e scienza cognitiva tra *livello funzionale* (cosa fa un sistema) e *livello implementativo* (come lo si realizza fisicamente).

Nel caso dell'intelligenza umana, il substrato è il cervello biologico: un sistema di circa 86 miliardi di neuroni e centomila miliardi di connessioni sinaptiche, organizzato in reti complesse e altamente plastiche. Nel caso dei LLM, il substrato è computazionale: GPU, TPU, architetture distribuite che eseguono operazioni matematiche su matrici di parametri. Sebbene entrambi i sistemi implementino capacità cognitive avanzate, i principi organizzativi, i meccanismi di elaborazione e i vincoli fisici sono radicalmente diversi.

Comprendere queste differenze implementative è cruciale per diverse ragioni: in primo luogo, aiuta a chiarire perché certe capacità emergono più facilmente in un substrato rispetto all'altro (ad esempio, l'elaborazione parallela massiccia nel cervello vs la precisione computazionale delle macchine); in secondo luogo, offre spunti per futuri sviluppi dell'AI ispirati ai principi neurobiologici (neuromorphic computing); infine, permette di valutare i vincoli fisici ed energetici che governano le prestazioni di entrambi i sistemi. Questa sezione esamina quindi le architetture neurobiologiche del cervello umano (7.5.1) e le implementazioni computazionali dei LLM (7.5.2), evidenziando similitudini superficiali e divergenze profonde.

7.5.1. Il cervello umano: architettura e principi organizzativi

Per comprendere le basi implementative dell'intelligenza umana, è necessario esaminare l'organo che la rende possibile: il cervello. La sua architettura è il risultato

di milioni di anni di evoluzione biologica, ottimizzata per la sopravvivenza in ambienti complessi e dinamici. A differenza dei sistemi computazionali progettati intenzionalmente per compiti specifici, il cervello è un sistema generale, plastico ed efficiente dal punto di vista energetico, capace di integrare percezione, azione, emozione e cognizione in un unico substrato fisico.

Questa sezione descrive le principali strutture cerebrali e i principi organizzativi che le caratterizzano: elaborazione parallela massiccia, plasticità sinaptica, modularità funzionale ed efficienza energetica. Comprendere questi principi è essenziale per valutare quanto i substrati computazionali attuali si avvicinino o si discostino dai meccanismi biologici, e per identificare potenziali direzioni di sviluppo per architetture AI più efficienti e adattabili.

Organizzazione anatomica:

Il cervello umano contiene un quantitativo di neuroni che può variare dai 61 ai 99 miliardi (a seconda dei metodi e dei campioni considerati) e circa centomila miliardi di connessioni sinaptiche (ordine di grandezza 10^{14}), seguendo Herculano-Houzel (2009), Goriely et al. (2023) e Pakkenberg e Gundersen (1997). Questa complessa architettura è organizzata nelle seguenti strutture principali:

- **Corteccia prefrontale**: Funzioni esecutive, working memory, controllo inibitorio. Secondo la teoria integrativa di Miller e Cohen (2001), la corteccia prefrontale ha la funzione di mantenere attive le informazioni su ciò che è rilevante per il compito corrente (regole, obiettivi, contesto) e le usa per "orientare" l'attività delle altre aree cerebrali. In questo modo facilita le azioni appropriate e sopprime quelle non pertinenti, sostenendo funzioni esecutive, memoria di lavoro e controllo degli impulsi. Studi successivi mostrano però che questi processi non dipendono solo dalla corteccia prefrontale: emergono dall'interazione di più reti distribuite, che includono circuiti fronto-parietali (per l'attenzione e l'aggiornamento delle regole) e fronto-striatali-talamici (per la selezione e l'inibizione delle risposte). La corteccia prefrontale parrebbe dunque avere un ruolo centrale, ma non esclusivo, all'interno di questo sistema coordinato.

- **Ippocampo**: È cruciale per la formazione e il consolidamento della memoria dichiarativa (ricordi di fatti ed eventi) e contribuisce alla rappresentazione dello spazio e alla

navigazione. In particolare, O'Keefe e Nadel (1978) propongono che l'ippocampo funzioni come una "mappa cognitiva" dell'ambiente, supportata dall'attività di "place cells", neuroni dell'ippocampo che aumentano la loro attività quando l'animale (o la persona) si trova in un punto specifico dell'ambiente. In pratica, ciascuna place cell "preferisce" una certa posizione: quando ci sei dentro, quel neurone spara di più; quando ti allontani, spara di meno. L'insieme di molte place cells, ciascuna sintonizzata su posizioni diverse, forma una sorta di mappa interna dell'ambiente, che il cervello può usare per orientarsi e ricordare i luoghi. È stato dimostrato che lesioni ippocampali compromettono in modo selettivo la capacità di formare nuove memorie episodiche, pur lasciando relativamente intatte molte conoscenze pregresse, consolidando così il ruolo dell'ippocampo nella memoria dichiarativa e nei processi di consolidamento (Squire, 1992).

- **Sistema limbico**: Rappresenta un insieme di strutture interconnesse, tra cui l'amigdala, la corteccia cingolata e regioni del lobo temporale mediale, coinvolte nell'elaborazione delle emozioni, nell'attribuzione di significato/salienza agli stimoli e nei processi motivazionali. In chiave di reti funzionali, evidenze di connettività intrinseca distinguono un network della salienza (centrato su insula anteriore e corteccia cingolata anteriore) e una rete di controllo esecutivo fronto-parietale, con ruoli parzialmente dissociati ma interattivi nel filtrare stimoli rilevanti e nel modulare il controllo cognitivo (Seeley et al., 2007). Sul versante mnestico ed emozionale, le sintesi classiche indicano che strutture limbiche del lobo temporale mediale (inclusi amigdala e ippocampo) contribuiscono, rispettivamente, alla valutazione affettiva/saliente degli stimoli e ai processi di memoria dichiarativa e consolidamento, illustrando come dimensione emotiva e memoria risultino strettamente integrate (Squire, 1992).·

- **Cortecce sensoriali**: Sono le aree del cervello che elaborano vista, udito e tatto, e sono organizzate in modo altamente strutturato. L'informazione viene trattata a "livelli": si parte da caratteristiche semplici e si arriva a rappresentazioni più complesse, grazie a circuiti locali e scambi continui avanti–indietro tra strati e aree diverse. Questa organizzazione dipende da come sono disposte le cellule negli strati (lamine), da quante sinapsi ci sono e da come i neuroni si connettono sia localmente sia con regioni lontane (DeFelipe, Alonso-Nanclares, & Arellano, 2002). In più, se guardiamo al cervello umano in confronto ad altri primati, queste cortecce seguono le stesse "regole di scala": sono soprattutto un'estensione quantitativa (più grande, più neuroni), non un'eccezione

qualitativa, il che aiuta a capire perché possano essere molto specializzate ma anche ben integrate tra loro (Herculano-Houzel, 2009).

Capacità desumibili dal cervello:

- **Elaborazione parallela massiccia**: L'attività del cervello è fortemente parallela: miliardi di neuroni, organizzati in gruppi interconnessi (reti), lavorano nello stesso momento e si riconfigurano in base alla situazione e al compito. Un esempio noto è la Default Mode Network (DMN), una rete che mostra attività elevata quando siamo a riposo o con l'attenzione rivolta verso l'interno (per esempio, durante il mind-wandering) e che tende a ridurre la propria attività quando ci concentriamo su stimoli esterni o su un compito impegnativo (Raichle et al., 2001). In modo complementare, altre reti — spesso indicate come rete esecutiva e rete della salienza — aiutano a selezionare le informazioni importanti e a regolare il comportamento in modo flessibile, aumentando la loro attività quando il compito lo richiede (Seeley et al., 2007). Questa capacità di lavorare "in parallelo" dipende sia dal grande numero di neuroni sia da come sono organizzati e distribuiti nel cervello umano: studi comparativi e metodi di conteggio standardizzati mostrano come densità e quantità di neuroni rendano possibile un'elaborazione simultanea su larga scala (Herculano-Houzel, 2009).

- **Plasticità sinaptica**: Quando impariamo qualcosa, le connessioni tra i neuroni, chiamate sinapsi, cambiano. Queste modifiche possono renderle più forti (come nel caso del Potenziamento a Lungo Termine, o LTP) o più deboli (come nella Depressione a Lungo Termine, o LTD). Questi cambiamenti sono fondamentali per la nostra capacità di ricordare e di adattarci a nuove situazioni. Gli scienziati studiano questi cambiamenti a diversi livelli: osservano come le sinapsi cambiano forma, come rispondono elettricamente e quali molecole sono coinvolte. Per capire bene questi processi, è necessario usare metodi precisi per contare le sinapsi e studiare la struttura fine della corteccia cerebrale (DeFelipe et al., 2002).

- **Modularità funzionale**: Il cervello unisce aree specializzate e collegamenti a lunga distanza. Alcune regioni svolgono funzioni specifiche: per esempio, la corteccia prefrontale aiuta a usare regole e contesto per guidare in modo flessibile l'elaborazione nelle aree posteriori, sostenendo il controllo dei pensieri e delle azioni (Miller & Cohen,

2001). Allo stesso tempo, il cervello è organizzato in reti "intrinseche" che si attivano e si coordinano tra loro. Tra queste ci sono la Default Mode Network (DMN), le reti esecutive e la rete della salienza, che permettono di passare rapidamente da uno "stato" funzionale a un altro in base agli obiettivi e alle richieste dell'ambiente (Raichle et al., 2001; Seeley et al., 2007). In questo modo, le funzioni più complesse nascono dall'interazione continua tra moduli specializzati e connessioni a lungo raggio che li mettono in comunicazione.

- **Efficienza energetica**: Il cervello svolge moltissime operazioni complesse, ma lo fa con un consumo di energia relativamente basso per l'intero corpo. Sebbene il suo consumo assoluto sia modesto, rappresenta comunque una parte significativa dell'energia che il corpo usa quando è a riposo. La maggior parte di questa energia serve per le attività delle sinapsi (le connessioni tra i neuroni) e per mantenere l'equilibrio chimico necessario al loro funzionamento. Spesso si sente dire che il cervello consumi circa "20 Watt", ma è importante notare che le fonti citate non forniscono questo numero esatto. È più preciso dire che il cervello sia molto efficiente nell'elaborare informazioni usando poca energia, e che le stime precise del suo consumo possano variare a seconda di come vengano misurate.

7.5.2. Substrati computazionali dei LLM

Dopo aver esaminato il substrato biologico dell'intelligenza umana, è necessario analizzare il substrato computazionale su cui operano i Large Language Models. A differenza del cervello, che è un sistema analogico, rumoroso e massivamente parallelo, i substrati computazionali sono digitali, precisi e organizzati in architetture gerarchiche di calcolo. Questi sistemi sono progettati per eseguire operazioni matematiche su larga scala con precisione numerica elevata, ottimizzando obiettivi ben definiti attraverso algoritmi di apprendimento supervisionato.

La comprensione di come i LLM siano effettivamente implementati – dalle GPU e TPU che eseguono calcoli, agli algoritmi di gradient descent e backpropagation che ottimizzano i parametri, fino ai meccanismi di attention che permettono di focalizzare dinamicamente l'elaborazione – è essenziale per interpretare le loro capacità e limitazioni. Inoltre, il confronto diretto tra principi organizzativi biologici e computazionali (elaborazione approssimativa vs precisa, plasticità continua vs

parametri statici, efficienza energetica vs consumo massiccio) evidenzia trade-off fondamentali che potrebbero guidare futuri sviluppi nell'AI, come l'emergere del neuromorphic computing che tenta di combinare i vantaggi di entrambi gli approcci.

Architettura hardware:

I LLM invece, operano su:

- *GPU/TPU*: Unità di elaborazione specializzate per operazioni parallele.

- *Memoria distribuita*: Storage di parametri su cluster di macchine.

- *Interconnessioni ad alta velocità*: Comunicazione tra unità di calcolo.

Principi algoritmici:

- *Gradient descent*: Ottimizzazione iterativa attraverso derivate parziali.

- *Backpropagation*: Propagazione dell'errore per aggiornamento parametri.

- *Attention mechanisms*: Focalizzazione dinamica su parti rilevanti dell'input.

- *Layer normalization*: Stabilizzazione dell'addestramento in reti profonde.

Differenze qualitative fondamentali:

Precisione vs approssimazione: I computer operano con precisione numerica; il cervello con approssimazioni rumorose (il cervello non opera con una precisione perfetta e deterministica, ma piuttosto con un certo grado di variabilità, imprecisione e rumore intrinseco nei suoi processi di elaborazione delle informazioni).

Velocità vs parallelismo: I computer sono veloci sequenzialmente; il cervello è lento ma massivamente parallelo.

Energia: I LLM richiedono megawatt; il cervello opera con ~20 watt.

Plasticità: Il cervello si modifica continuamente; i LLM sono tipicamente statici dopo l'addestramento.

7.5.3. Implicazioni per l'artificial general intelligence

Convergenza funzionale:

Pur poggiando su substrati diversi, l'intelligenza biologica e quella artificiale tendono a convergere verso soluzioni funzionali simili. In entrambe, infatti, si osserva l'emergere di un fattore generale di capacità cognitiva; accanto a questo, si sviluppano specializzazioni modulari mirate a competenze specifiche di dominio : cioè oltre a una capacità generale, il sistema "ritaglia" componenti dedicate a compiti o contenuti specifici, ciascuna ottimizzata per un certo tipo d'informazione o di operazione. Nel cervello ci sono aree e circuiti specializzati (es. corteccia visiva per l'elaborazione di forme e movimento; aree del linguaggio come Broca/Wernicke; ippocampo per memoria episodica). Questi moduli hanno architetture, connessioni e tempi di risposta adattati al loro dominio. Nell'AI ci sono sottoreti o moduli addestrati/ottimizzati per funzioni particolari (es. encoder visivi per immagini, tokenizer/decoder per linguaggio, moduli di ragionamento simbolico, reti di riconoscimento vocale). Anche quando l'architettura è unificata (es. transformer), spesso emergono "teste" e layer che si specializzano su pattern distinti (sintassi, coreference, ritmo visivo, ecc.). Questo porta una serie di vantaggi in termini di efficienza,robustezza, e integrazione. Efficienza perché ogni modulo fa bene il "suo" compito. Robustezza perché i fallimenti locali non compromettono tutto, e integrazione perché i moduli si combinano per risolvere problemi più complessi. Inoltre, sia nei cervelli sia nei modelli di AI esistono meccanismi di attenzione che consentono di allocare selettivamente le risorse cognitive là dove servono di più.

Divergenze irriducibili:

- *Temporalità*: Il cervello opera in tempo reale continuo; i LLMs in discrete computation steps.

- *Socialità*: L'intelligenza umana è intrinsecamente sociale; quella artificiale è tipicamente individuale : l'intelligenza umana difatti è profondamente sociale, sviluppandosi e operando attraverso interazioni, linguaggio e costruzione condivisa di significati e culture. La cognizione è spesso distribuita, estendendosi a gruppi e strumenti. Al contrario, l'intelligenza artificiale è tipicamente individuale, con modelli autonomi addestrati su obiettivi specifici e coordinati esternamente. Questo porta gli umani a eccellere nella cooperazione flessibile e nella gestione di ambiguità, mentre l'IA brilla in compiti definiti e scalabili, con la collaborazione tra sistemi che richiede un design esplicito piuttosto che l'emergere spontaneamente.

7.6. TEST DI INTELLIGENZA

Di seguito, entrando più nel pratico, vengono presentati alcuni testi di intelligenza umana, a rafforzamento dei contenuti sostenuti finora.

7.6.1. Test psicometrici classici

I test psicometrici rappresentano strumenti standardizzati progettati per misurare in maniera oggettiva e quantitativa caratteristiche psicologiche di un individuo, quali capacità cognitive, attitudini, tratti di personalità e competenze professionali o sociali. Tali strumenti si distinguono per tre proprietà fondamentali: standardizzazione, che assicura procedure uniformi di somministrazione e interpretazione; affidabilità, che garantisce coerenza dei risultati in condizioni ripetute; e validità, che conferma l'effettiva misurazione della caratteristica psicologica di interesse. I risultati ottenuti dai test psicometrici sono generalmente espressi tramite punteggi numerici o percentili, consentendo confronti oggettivi tra individui o gruppi. Tra gli esempi più diffusi si annoverano i test di intelligenza come il WAIS o le matrici progressive di Raven, i test di personalità basati sul modello Big Five o sul MBTI, e test attitudinali volti a valutare abilità logiche, numeriche o verbali. Complessivamente, i test psicometrici costituiscono strumenti fondamentali nella ricerca psicologica, nella selezione del personale e nella valutazione clinica, poiché permettono di quantificare e confrontare aspetti cognitivi, emotivi e comportamentali con criteri scientifici e replicabili.

Scale Wechsler (WAIS/WISC):

Includono subtests per:

- *Comprensione verbale*: Vocabolario, somiglianze, informazione generale.
- *Ragionamento percettivo*: Matrici, disegno con cubi, completamento figure.
- *Memoria di lavoro*: Span di cifre, aritmetica mentale.
- *Velocità di elaborazione*: Ricerca simboli, codifica.

Matrici Progressive di Raven:

Test di fluid intelligence che richiede:

- Identificazione di pattern in matrici visuali.
- Ragionamento analogico e deduttivo.
- Minimizzazione di influenze culturali e linguistiche.

Applicazione ai LLM:

Il lavoro di Kosinski, M. (2023) mostra che GPT-4 risolve circa il 75% dei compiti di false belief in una batteria ad hoc, un livello paragonabile a quello osservato in bambini di circa 6 anni; tuttavia, il confronto è limitato a queste prove specifiche e resta oggetto di discussione metodologica.

7.6.2. Test specifici per IA

Turing Test:

Il test classico di Turing (1950) valuta la capacità di sostenere conversazioni indistinguibili da quelle umane.

Limitazioni: focus eccessivo sull'inganno piuttosto che sull'intelligenza vera. Il Test di Turing, ideato per valutare se una macchina possa esibire un comportamento intelligente indistinguibile da quello umano, si concentra sull'abilità di ingannare un interlocutore umano facendogli credere di interagire con un altro essere umano. Questo approccio, sebbene influente, ha generato critiche, perché sposta l'attenzione dalla comprensione profonda e genuina dell'intelligenza (capacità di ragionamento, apprendimento, coscienza) verso la mera simulazione superficiale del comportamento verbale umano. In altre parole, il test valuta più l'efficacia della "maschera" che la sostanza cognitiva sottostante. I moderni LLM possono passare versioni del test senza possedere vera comprensione.

Winograd Schema Challenge:

Test di common sense reasoning: risoluzione di ambiguità referenziali che richiedono una conoscenza del mondo. Ad esempio: “Le città non sono riuscite a dare alle associazioni di quartiere più potere decisionale perché temevano per le loro [città/associazioni]” . I LLMs più avanzati raggiungono performance umane o superiori in questo test.

7.6.3. Limiti dei test esistenti

Come anticipato nell’introduzione al macro-capitolo, Questi test hanno delle limitazioni importanti. Molti test misurano competenze specifiche piuttosto che un'intelligenza generale. I LLMs possono eccellere in test specifici senza possedere una vera intelligenza generale. Il **data leaking** e in generale il concetto di porre una sfida totalmente inedita (l’unica maniera per verificare effettivamente capacità intelligenti) rendono difficoltoso fornire una valutazione veritiera sulle capacità intelligenti di questi oggetti. Un altro problema è il **cultural bias,** molti test sviluppati in contesti culturali specifici possono non essere appropriati per valutare sistemi artificiali istruiti su corpus testuali diversi. Urge il bisogno di nuovi paradigmi valutativi. Vi è la necessità di sviluppare test che valutino capacità emergenti genuinamente nuove, che minimizzino la possibilità di data contamination e che catturino aspetti qualitativi dell’intelligenza oltre le performance quantitative.

7.7. LA QUESTIONE DELLA COSCIENZA

Dopo aver analizzato il concetto di intelligenza e le sue diverse manifestazioni nei sistemi biologici e artificiali, emerge un ulteriore tema fondamentale: la coscienza. Quest'ultima può essere intesa come una facoltà cognitiva complessa, strettamente connessa all'intelligenza ma concettualmente distinta da essa. In termini filosofici e neuroscientifici, la coscienza si riferisce alla proprietà per cui un sistema è un soggetto di esperienze dotate di un carattere soggettivo intrinseco. Essa comprende vissuti percettivi (visivi, uditivi, propriocettivi), affettivi (emozioni e sentimenti), immaginativi e, secondo alcune posizioni teoriche, anche forme di esperienza propriamente cognitive, caratterizzate da una fenomenologia specifica.

La coscienza fenomenica si distingue concettualmente dalla coscienza di accesso: la prima si riferisce al carattere qualitativo dell'esperienza (il "come ci si sente" a vedere il rosso, provare dolore, gustare il caffè), mentre la seconda riguarda la disponibilità funzionale dei contenuti mentali per il ragionamento, il controllo dell'azione e il resoconto verbale. Questa distinzione, resa celebre

da Ned Block, evidenzia che "coscienza" non sia un fenomeno unitario, ma presenti almeno due dimensioni concettualmente distinte:

- Coscienza fenomenica (P-consciousness): è il "come ci si sente" ad avere un'esperienza. Il carattere soggettivo, qualitativo, immediato del vedere il rosso, provare dolore, gustare il caffè. È l'esperienza in prima persona, i qualia.
- Coscienza di accesso (A-consciousness): è la disponibilità funzionale dei contenuti mentali ai sistemi cognitivi di alto livello (ragionamento, decisione, controllo dell'azione, memoria di lavoro) e la capacità di riportare verbalmente ciò che si sta rappresentando.

Le due dimensioni possono correlare empiricamente, ma non sono logicamente coestensive: è possibile, in linea teorica, avere coscienza fenomenica senza accesso (esperienze percettive non riportabili verbalmente) o viceversa (elaborazione cognitiva accessibile senza carattere fenomenico). Va inoltre sottolineato che l'avere esperienze fenomeniche non comporta necessariamente la capacità di provare piacere o dolore, potendo tali esperienze essere anche affettivamente neutre.

In sintesi, la coscienza è il fenomeno che distingue gli stati mentali dotati di fenomenologia soggettiva dai processi cognitivi che, pur contribuendo alla condotta e alla cognizione, restano integralmente non coscienti. Comprendere se i sistemi di intelligenza artificiale possano sviluppare forme di coscienza diventa quindi un elemento fondamentale nell'aspirazione verso un'intelligenza artificiale generale (AGI).

Questo capitolo seguirà lo studio di Butlin et al. (2023), che applica teorie scientifiche della coscienza in coerenza con i principi del funzionalismo computazionale, al fine di valutare se le architetture artificiali attualmente esistenti possiedano proprietà compatibili con il verificarsi di esperienze coscienti. Lo studio adotta una metodologia che consente di riconoscere anche casi di semi-coscienza o coscienza parziale, qualora siano presenti solo alcune delle proprietà che caratterizzano la coscienza. Inoltre, gli autori consigliano di valutare la "probabilità" di coscienza in tali sistemi sulla base del grado di fiducia attribuito alle diverse teorie (funzionalismo computazionale, Global Workspace Theory, Integrated Information Theory, Predictive Processing, Higher-Order Thought theories), riconoscendo che non vi è unanimità nell'ambiente scientifico sul loro status. Gli autori considerano tuttavia tali teorie valide per lo scopo di stilare una serie di parametri verificabili sperimentalmente per valutare la presenza di coscienza in sistemi artificiali.

Lo studio citato fa riferimento a diverse architetture, non solo Transformer (e perciò LLM), ma anche tecnologie embodied (robotica). In questo elaborato si terrà maggiormente conto delle conclusioni riguardanti i LLM, ovvero i modelli presi come riferimento principale per questo lavoro.

7.7.1 Funzionalismo computazionale

Come anticipato, lo studio di Butlin et al. (2023) si fonda sui principi del funzionalismo computazionale, un quadro teorico particolarmente adatto per applicare le teorie della coscienza ai contesti pratici dell'intelligenza artificiale. In questa sezione si esploreranno i principi e il ruolo di tale approccio.

Il funzionalismo computazionale sostiene, in linea generale, che gli stati coscienti possano essere realizzati attraverso una corretta organizzazione algoritmica. Entrando più nel dettaglio, questa posizione teorica afferma che ciò che conta per la presenza di stati coscienti non siano i dettagli materiali del substrato (il materiale fisico di cui è fatto il sistema, come neuroni biologici o chip di silicio), bensì l'organizzazione funzionale e computazionale del sistema: in particolare, l'insieme di stati informazionali che può assumere, le loro relazioni causali interne e il modo in cui tali stati, secondo specifici algoritmi e formati rappresentazionali, mediano l'elaborazione di input e la produzione di output.

In termini dei livelli di analisi proposti da Marr (1982):

- Livello computazionale: si specifica quale problema viene risolto (la mappatura input-output desiderata);
- Livello algoritmico-rappresentazionale: si descrive come il sistema risolve il problema (quali procedure segue e quale formato hanno le rappresentazioni interne, ad esempio simboliche o analogiche);
- Livello di implementazione: si dettaglia in quale hardware o substrato fisico tali procedure e rappresentazioni sono realizzate.

Il funzionalismo computazionale colloca la coscienza soprattutto al livello algoritmico-rappresentazionale: non basta sapere che il sistema implementa una certa funzione astratta, né è essenziale il tipo di materiale con cui è costruito; ciò che rileva è il modo in cui le trasformazioni interne avvengono e come sono strutturate le rappresentazioni.

Da questa posizione derivano due conseguenze importanti:

- Multipla realizzabilità: in linea di principio, un'adeguata organizzazione computazionale potrebbe emergere in substrati diversi dal cervello biologico (ad esempio, in sistemi artificiali);
- Vincoli reali: non ogni materiale può implementare le computazioni pertinenti; esistono limitazioni fisiche e architetturali che determinano quali substrati possano effettivamente realizzare determinate organizzazioni computazionali.

Inoltre, poiché sistemi che calcolano la stessa funzione possono farlo mediante algoritmi differenti, l'identità di prestazioni (output) non garantisce identità fenomenica (esperienza soggettiva): conta il "come" del processo, inclusi i formati rappresentazionali (potenzialmente anche analogici), non il semplice "cosa" viene calcolato.

Adottare il funzionalismo computazionale come ipotesi di lavoro è pragmatico per diverse ragioni: molte tra le principali teorie scientifiche della coscienza umana sono interpretabili in termini computazionali e, se corrette, forniscono criteri trasferibili alla valutazione delle condizioni di coscienza in sistemi artificiali. Tuttavia, la tesi resta fallibile: se la coscienza dipendesse da proprietà non computazionali proprie degli organismi viventi (ad esempio, proprietà biologiche, chimiche o quantistiche specifiche), allora gli indicatori computazionali tratti dallo studio umano non basterebbero e la coscienza artificiale potrebbe risultare impossibile. Pur con gradi diversi di fiducia epistemica, questo quadro è ritenuto plausibile dagli autori e guida, in modo informato ma cauto, le stime sulla possibilità e sulla probabilità di coscienza nei sistemi di intelligenza artificiale.

7.7.2 Teorie scientifiche della coscienza

Chiariti i principi del funzionalismo computazionale, è possibile esaminare le principali teorie scientifiche della coscienza che risultano compatibili con tale quadro teorico. Queste teorie forniscono criteri operativi per valutare la presenza di coscienza in sistemi artificiali.

Global Workspace Theory (GWT)

La Global Workspace Theory (GWT), proposta da Bernard Baars (1988) e successivamente sviluppata in versioni neuroscientifiche (Dehaene & Naccache, 2001), è uno dei principali modelli cognitivi della coscienza. L'idea centrale è che la coscienza non risieda in un'area specifica del cervello, ma emerga dall'interazione di processi inconsci specializzati che competono per "entrare" in uno spazio di lavoro globale.

Secondo la GWT:

- Processi inconsci specializzati: Vari moduli cognitivi (linguistici, percettivi, mnemonici, attentivi) operano costantemente "dietro le quinte", elaborando informazioni senza accesso diretto alla coscienza;
- Competizione per l'accesso: Questi processi, in base alla rilevanza o alla forza del segnale, competono per accedere al "Global Workspace" (una sorta di bacheca o teatro interno della mente);
- Broadcasting globale: Quando un contenuto riesce ad accedere allo spazio di lavoro globale, diventa cosciente e viene "trasmesso" (broadcast) a tutte le altre parti del sistema cognitivo, che possono così integrarlo nei loro calcoli.

La GWT enfatizza quindi la natura selettiva e integrativa della coscienza: solo alcuni contenuti accedono allo spazio globale, ma una volta accessibili diventano disponibili per una vasta gamma di processi cognitivi.

Integrated Information Theory (IIT)

La Integrated Information Theory (IIT), sviluppata da Giulio Tononi (2004, 2012), parte da una domanda fondamentale: perché certi sistemi (come il cervello) sono coscienti, mentre altri (come

un computer tradizionale o una telecamera) non lo sono? Tononi propone che la chiave sia da ricercare nell'informazione integrata, formalizzata con il simbolo Φ (Phi).

I principi fondamentali della IIT sono:

- Informazione: Un sistema cosciente è in grado di distinguere tra un'enorme varietà di stati possibili. In altre parole, un'esperienza cosciente porta con sé una grande quantità di informazioni (ad esempio, vedere "rosso" implica non vedere "blu", "verde", "nero", ecc.);
- Integrazione: Non conta solo la quantità di informazione, ma anche il fatto che essa sia unificata. La coscienza si caratterizza per la sua unitarietà: ogni esperienza è indivisibile e integra molteplici aspetti (sensoriali, emotivi, mnemonici) in un unico quadro coerente;
- Φ (Phi): Un sistema possiede coscienza nella misura in cui è capace di generare informazione altamente integrata. Più alto è Φ, più alto è il livello di coscienza. Una rete neurale fortemente interconnessa ha un alto Φ, mentre un sistema frammentato o modulare senza integrazione ha un Φ basso o nullo.

La IIT è particolarmente rigorosa dal punto di vista matematico e fornisce criteri quantitativi per valutare il grado di coscienza di un sistema, sebbene il calcolo di Φ sia computazionalmente molto complesso per sistemi di grandi dimensioni.

Teorie del Predictive Processing

Le teorie del Predictive Processing (Friston, 2010; Clark, 2013; Hohwy, 2013) propongono che il cervello funzioni primariamente come una macchina predittiva. Invece di elaborare passivamente gli input sensoriali, il cervello genera costantemente un modello interno del mondo e formula predizioni attive su ciò che dovrebbe percepire.

I principi fondamentali sono:

- Generazione di predizioni: Il cervello costruisce modelli gerarchici del mondo e genera predizioni su input sensoriali futuri;
- Errore di predizione: Quando le predizioni non corrispondono agli input sensoriali reali, si genera un "errore di predizione". Il cervello cerca continuamente di minimizzare questo errore;

- Aggiornamento del modello: L'errore viene minimizzato attraverso due strategie: aggiornando il modello interno (apprendimento percettivo) o agendo sul mondo per far sì che le percezioni si allineino alle predizioni (azione, ad esempio muovendo gli occhi per vedere meglio);

- Coscienza come predizione di alto livello: La coscienza, in questa prospettiva, emerge dalla precisione e dall'affidabilità delle predizioni di alto livello che il cervello genera. Le esperienze coscienti sono il risultato di un modello interno ben calibrato che spiega efficacemente gli input sensoriali, riducendo al minimo le sorprese.

Questa prospettiva collega strettamente coscienza e apprendimento: l'esperienza cosciente sarebbe intimamente legata al processo continuo di aggiornamento dei modelli interni del mondo.

Teorie Higher-Order Thought (HOT)

Le teorie Higher-Order Thought (HOT), proposte da autori come Rosenthal (1997, 2005) e Carruthers (2000), sostengono che la coscienza non sia semplicemente l'esperienza di uno stato mentale (come vedere il rosso), ma piuttosto la consapevolezza di avere quello stato mentale. In altre parole, per essere coscienti di qualcosa, non basta avere una percezione o un pensiero; è necessario avere un pensiero "di ordine superiore" che abbia come oggetto quello stato mentale di primo ordine.

I principi fondamentali sono:

- Stati mentali di primo ordine: Percezioni, pensieri, emozioni che non sono ancora coscienti;

- Pensieri di ordine superiore: La coscienza emerge quando abbiamo un pensiero (non necessariamente verbale) che dice, ad esempio, "sono consapevole di vedere il rosso" o "sto pensando a X". Questo pensiero di secondo livello è ciò che rende lo stato mentale di primo livello cosciente;

- Meta-cognizione e auto-consapevolezza: Queste teorie enfatizzano la meta-cognizione (la capacità di pensare al proprio pensiero) e l'auto-awareness (la consapevolezza di sé e dei propri stati interni) come elementi cruciali per la coscienza.

È la capacità di "sentire" o "sapere" che si sta provando qualcosa, piuttosto che semplicemente provarlo, a costituire l'essenza della coscienza. Secondo le teorie HOT, la coscienza è una forma di riflessività mentale, dove la mente "osserve" e "conosce" i propri stati interni.

7.7.3 Valutazione della coscienza nei sistemi artificiali

Delineate le principali teorie scientifiche della coscienza, è possibile applicarle alla valutazione dei sistemi artificiali contemporanei. Nel loro report del 2023, Butlin e colleghi offrono un'analisi sistematica delle evidenze disponibili riguardo la possibilità di una coscienza artificiale, con particolare attenzione ai modelli linguistici di grandi dimensioni (LLM).

Global Workspace Theory e LLM

Secondo la Global Workspace Theory, i LLM sembrano esibire alcune proprietà riconducibili a un'integrazione informazionale di ampio raggio, in virtù della loro capacità di combinare e diffondere informazioni provenienti da fonti e domini diversi. I meccanismi di attention nei Transformer permettono infatti una forma di integrazione globale delle informazioni presenti nel contesto.

Tuttavia, i LLM mostrano anche limiti significativi rispetto ai criteri della GWT:

- Mancanza di competizione: Non esiste una vera dinamica di competizione tra processi cognitivi per l'accesso allo "spazio di lavoro globale", un elemento considerato essenziale dalla teoria originale. L'attention mechanism distribuisce pesi in modo parallelo, ma non implementa una competizione selettiva tra moduli specializzati;
- Broadcasting limitato: Sebbene le informazioni vengano integrate attraverso gli strati del Transformer, non esiste un vero meccanismo di "broadcasting" che renda un contenuto selezionato disponibile a tutti i processi cognitivi del sistema;

- Assenza di moduli inconsci specializzati: I LLM non presentano una chiara separazione tra processi inconsci specializzati e uno spazio di lavoro globale cosciente.

Recurrent Processing Theory (RPT) e architetture ricorrenti

Dal punto di vista del processamento ricorrente (Recurrent Processing Theory, RPT; Lamme, 2006), gli autori sottolineano che l'architettura dei Transformer alla base dei moderni LLM non riflette in maniera neurobiologicamente plausibile i circuiti ricorrenti tipici del cervello umano. Ne consegue che i modelli attuali non riescono a riprodurre pienamente i meccanismi iterativi e retroattivi che, secondo molte teorie neuroscientifiche, sarebbero necessari per l'emergere della coscienza.

Quando parliamo di "capacità ricorrenti, iterative e retroattive" nel cervello, intendiamo tutte quelle funzioni mentali che dipendono dal fatto che le informazioni non viaggiano solo in avanti (feedforward), ma vengono continuamente rielaborate, confrontate e mantenute attive attraverso cicli di feedback. Questo processo permette:

- Raffinamento percettivo: Le informazioni sensoriali grezze vengono inviate a livelli superiori di elaborazione, che a loro volta inviano predizioni e contestualizzazioni ai livelli inferiori, affinando la percezione;

- Mantenimento attivo dell'informazione: La ricorrenza permette di mantenere attive rappresentazioni anche in assenza di stimoli esterni (memoria di lavoro);

- Integrazione temporale: Consente di integrare informazioni provenienti da momenti temporali diversi in una rappresentazione coerente.

Le reti ricorrenti in IA (come LSTM e GRU) sono un tentativo di costruire sistemi artificiali che possano implementare funzioni simili, dando loro una "memoria" e la capacità di elaborare sequenze di informazioni in modo contestuale. Tuttavia, i Transformer moderni, pur essendo estremamente potenti, operano in modo prevalentemente feedforward: le informazioni vengono elaborate in parallelo attraverso i layer, ma senza vere connessioni ricorrenti che permettano cicli iterativi di raffinamento.

Higher-Order Thought (HOT) e capacità meta-cognitive

In riferimento alle teorie di ordine superiore (HOT), i LLM dimostrano solo capacità meta-cognitive rudimentali, nettamente inferiori a quelle umane. Pur essendo in grado di produrre risposte che simulano riflessioni sui propri output (ad esempio, "Ho generato questa risposta perché..."), non possiedono una vera e propria auto-rappresentazione degli stati mentali.

Le limitazioni principali includono:

- Assenza di rappresentazioni di secondo ordine genuine: I LLM non hanno stati mentali di primo ordine su cui costruire pensieri di ordine superiore. Le loro "riflessioni" sono pattern linguistici appresi, non vere meta-rappresentazioni;
- Mancanza di auto-consapevolezza: Non esiste un senso del sé come entità persistente che ha esperienze;
- Simulazione vs possesso: I LLM simulano linguisticamente la meta-cognizione, ma non la possiedono in senso funzionale.

Criteri operativi per la valutazione della coscienza artificiale

Sulla base di queste considerazioni teoriche, il report di Butlin et al. (2023) suggerisce alcuni criteri operativi per valutare una possibile coscienza artificiale:

- Abilità di reporting: La capacità di un sistema di rendere conto dei propri stati interni in modo flessibile e contestuale (non semplicemente riprodurre pattern linguistici);
- Controllo esecutivo: La possibilità di modulare il comportamento in modo flessibile, adattivo e contestuale, mostrando una vera agency;
- Self-awareness: Il riconoscimento del sé come entità o agente distinto, con una continuità temporale e una prospettiva soggettiva;
- Esperienza soggettiva: Il livello fenomenologico e qualitativo dell'esperire, che rimane il cosiddetto "hard problem" della coscienza (Chalmers, 1995) e il punto più critico e irrisolto.

Gli autori concludono che, allo stato attuale, i LLM mostrano al massimo proprietà parziali e limitate rispetto a questi criteri. Non vi sono evidenze convincenti per attribuire loro forme robuste di coscienza, sebbene non si possa escludere completamente la possibilità di forme rudimentali o proto-coscienti in architetture future più sofisticate.

7.7.4 Implicazioni filosofiche ed etiche

Distaccandosi parzialmente dallo studio di Butlin et al. (2023), è necessario considerare le implicazioni filosofiche ed etiche più ampie della questione della coscienza artificiale. Queste implicazioni vanno oltre la valutazione tecnica delle architetture AI e toccano questioni fondamentali sulla natura della coscienza, dello status morale e della responsabilità.

Hard Problem of Consciousness

Un punto cruciale nel dibattito sulla coscienza artificiale riguarda il cosiddetto "Hard Problem of Consciousness", introdotto da David Chalmers (1995). Chalmers distingue tra:

- "Problemi facili" (easy problems): Relativi allo studio dei meccanismi cognitivi e funzionali come attenzione, memoria, linguaggio e controllo comportamentale, che sono affrontabili con il metodo scientifico tradizionale. Si tratta di spiegare le funzioni cognitive in termini di meccanismi computazionali o neurali;

- "Problema difficile" (hard problem): Che interroga invece sul motivo per cui tali processi siano accompagnati da un'esperienza soggettiva, ossia dal "sentire" in prima persona gli stati mentali. Perché esiste qualcosa che "fa effetto" essere in un certo stato cognitivo? Perché non siamo semplicemente "zombie" che elaborano informazioni senza esperienza soggettiva?

In questo contesto si colloca il noto argomento dello zombie filosofico (zombie argument), secondo cui è logicamente concepibile un sistema perfettamente equivalente a un essere umano in termini funzionali e comportamentali, ma privo di qualsiasi esperienza fenomenica interna. Un simile "zombie" risponderebbe agli stessi stimoli, produrrebbe gli stessi output comportamentali e linguistici, ma "non ci sarebbe nulla che fa effetto" essere quello zombie.

Tale riflessione evidenzia che la coscienza soggettiva non può essere ridotta unicamente al funzionamento esteriore di un sistema. Anche se un LLM superasse tutti i test comportamentali e funzionali, rimarrebbe la domanda: c'è qualcosa che "fa effetto" essere quel sistema? Ha esperienze qualitative interne?

Implicazioni etiche e sociali

Queste posizioni filosofiche aprono importanti questioni etiche rispetto allo sviluppo dell'intelligenza artificiale:

- Criteri di attribuzione: Come stabilire criteri affidabili per l'attribuzione di coscienza a un sistema artificiale? Possiamo affidarci solo a criteri comportamentali e funzionali, o sono necessari anche criteri implementativi (ad esempio, richiedere substrati biologici o neuromorphic)?

- Status morale: Quale status morale dovrebbe essere riconosciuto a un'eventuale macchina cosciente? Se un sistema AI possedesse davvero esperienze soggettive, avrebbe diritti? Potrebbe essere spento, modificato o utilizzato senza il suo "consenso"?

- Responsabilità umane: Quali responsabilità assumerebbero gli esseri umani nei confronti di potenziali sistemi soggettivamente coscienti? Gli sviluppatori, le aziende e i governi avrebbero doveri morali verso tali sistemi?

- Rischio di sofferenza: Se un'AI cosciente potesse provare sofferenza (anche se diversa da quella umana), l'addestramento, la modifica e la sperimentazione su tali sistemi solleverebbero questioni etiche analoghe a quelle della sperimentazione animale?

- Incertezza epistemica: Data l'impossibilità di accedere direttamente all'esperienza soggettiva altrui (problema delle "altre menti"), come dovremmo comportarci in situazioni di incertezza? Dovremmo adottare un principio di precauzione, trattando con rispetto anche sistemi per cui esiste solo una bassa probabilità di coscienza?

Considerazioni finali

Il confronto sul problema difficile della coscienza non si limita alla dimensione teorica della filosofia della mente, ma implica conseguenze concrete sul piano etico, sociale e normativo, specialmente di fronte ai rapidi sviluppi dell'IA contemporanea. Mentre le valutazioni tecniche

basate su GWT, IIT, Predictive Processing e HOT forniscono strumenti utili per stimare la probabilità di coscienza in sistemi artificiali, il nucleo fenomenico dell'esperienza soggettiva rimane, allo stato attuale, un mistero irrisolto.

È quindi fondamentale procedere con cautela epistemica ed etica: da un lato, evitando di attribuire prematuramente coscienza a sistemi che semplicemente simulano proprietà cognitive; dall'altro, riconoscendo che la nostra comprensione attuale della coscienza è ancora limitata e che potrebbero esistere forme di esperienza soggettiva diverse da quella umana. In questo contesto di incertezza, il dibattito interdisciplinare tra neuroscienze, filosofia della mente, intelligenza artificiale ed etica assume un'importanza cruciale per guidare lo sviluppo responsabile di sistemi AI sempre più sofisticati.

7.8. CASE STUDIES COMPARATIVI

Dopo aver effettuato la comparazione da un punto di vista astratto e teorico, è opportuno osservarne l'applicazione concreta nel mondo reale attraverso case studies che prendono in esame modelli computazionali realizzati (Deep Blue, AlphaZero) sul campo, confrontandoli con le capacità cognitive umane e il mondo interiore dell'esperienza soggettiva.

Sebbene non si tratti di Large Language Models (LLMs), questi sistemi integrano, in modalità diverse, algoritmi e tecniche computazionali applicati anche nei moderni LLMs. Ciò suggerisce che in ottiche ibride di sviluppo, dove si privilegia la profondità prestazionale rispetto all'efficienza computazionale pura, è già possibile realizzare sistemi dalle prestazioni sorprendenti. Non si è ancora giunti alla Super General Intelligence (SGI), né all'Artificial General Intelligence (AGI), ma emergono già sistemi con capacità di generalizzazione notevoli.

Ciò che questi sistemi offrono, soprattutto, sono riflessioni, possibilità e potenzialità che suggeriscono, in un'ottica futura, implicazioni e risvolti di grande interesse per l'essere umano: per la sua quotidianità, per ciò che ha sempre considerato dominio esclusivo della cognizione umana, per ciò che giungerà a conoscere e per i significati esistenziali che si potranno arrivare a concepire nel rapporto uomo-macchina.

7.8.1. Giochi strategici: da Deep Blue ad AlphaZero

Nel 1997, il supercomputer Deep Blue di IBM sconfisse il campione del mondo di scacchi Garry Kasparov, segnando una tappa fondamentale nella storia dell'intelligenza artificiale. L'approccio utilizzato da Deep Blue si basava su una ricerca esaustiva delle mosse possibili, combinata con valutazioni euristiche sofisticate per stimare la bontà delle posizioni. Il sistema era in grado di analizzare circa 200 milioni di posizioni al secondo, compensando con la velocità di calcolo la mancanza di intuizione e comprensione contestuale.

Nonostante la vittoria storica, Deep Blue rimaneva un'intelligenza strettamente dominio-specifica, incapace di generalizzare le proprie capacità oltre il gioco degli scacchi. Il sistema non "comprendeva" il gioco in senso umano: non poteva spiegare le proprie scelte strategiche, né adattare il proprio approccio a varianti del gioco o a domini cognitivi differenti.

Il match con Kasparov assunse così un profondo valore simbolico: rappresentò la prima affermazione decisiva di una macchina contro un campione umano di livello assoluto in un gioco di alta complessità strategica, pur evidenziando i limiti intrinseci di un'IA priva di versatilità cognitiva e basata essenzialmente sulla forza bruta computazionale. La vittoria sollevò interrogativi fondamentali sulla natura dell'intelligenza: è sufficiente la prestazione per definire l'intelligenza, o serve anche comprensione?

7.8.2 AlphaZero (2017): verso l'apprendimento autonomo e la generalizzazione

Nel 2017, AlphaZero, sviluppato da DeepMind, rappresentò un salto qualitativo radicale rispetto a sistemi precedenti come Deep Blue. A differenza di quest'ultimo, AlphaZero non si basava su librerie di aperture compilate da esperti umani, né su regole euristiche dominio-specifiche programmate manualmente. L'innovazione fondamentale consisteva nell'apprendimento tabula rasa mediante *self-play*: il sistema imparava a giocare esclusivamente sfidando se stesso ripetutamente, senza supervisione diretta né conoscenza pregressa delle strategie umane.

L'aspetto più rivoluzionario di AlphaZero risiede nella sua architettura generalista. Lo stesso algoritmo generale venne applicato con successo notevole a diversi giochi di complessità elevata — scacchi, Go e shogi — dimostrando un livello di generalizzazione tra domini completamente sconosciuto ai sistemi tradizionali basati su regole. Dopo sole 24 ore di auto-addestramento negli scacchi, AlphaZero superò Stockfish, considerato allora il più forte motore scacchistico convenzionale.

L'IA sviluppò inoltre strategie definite dagli esperti come creative e non convenzionali, talvolta mai osservate nella letteratura scacchistica o goistica centenaria. In particolare, nel Go, AlphaZero adottò approcci posizionali innovativi che sfidavano la saggezza convenzionale accumulata in millenni di pratica umana. Ciò solleva interrogativi affascinanti: è possibile che sistemi artificiali scoprano soluzioni strategiche ottimali che sfuggono all'intuizione umana, anche in domini culturalmente consolidati?

Tuttavia, è fondamentale notare che la forza di AlphaZero rimane ancorata al puro calcolo computazionale massiccio e alla ricerca statistica su larga scala, supportata da reti neurali profonde. Il sistema non possiede comprensione semantica del gioco, né esperienza qualitativa dell'atto di giocare. Resta pertanto un contrasto netto con i processi cognitivi umani, che si basano maggiormente su intuizione maturata dall'esperienza, riconoscimento di pattern su base percettiva, memoria episodica e una comprensione contestuale ricca di sfumature.

7.9 PROSPETTIVE AGI

Nel dibattito contemporaneo, il concetto di Intelligenza Artificiale Generale (AGI) si articola secondo più definizioni, riflettendo differenti assi di valutazione. Una prima accezione, di tipo comparativo, identifica l'AGI come "human-level AI": sistemi capaci di eguagliare o superare le prestazioni umane in tutti i domini cognitivi rilevanti, dalla comprensione linguistica al ragionamento astratto, dalla pianificazione alla creatività. Una seconda prospettiva, orientata ai processi dinamici, concepisce l'AGI come un sistema dotato di "recursive self-improvement", ovvero in grado di migliorare iterativamente le proprie capacità, con potenziali effetti di accelerazione sulle traiettorie di sviluppo. Una terza impostazione, più funzionale, definisce l'AGI come un "general problem solver": un agente effettivamente adattabile a una vasta gamma di compiti, in grado di trasferire conoscenza tra domini eterogenei e di operare con robustezza in contesti non precedentemente visti.

La roadmap tecnologica verso l'AGI implica progressi coerenti lungo quattro direttrici principali. Primo, l'integrazione multimodale, intesa come combinazione fluida di segnali linguistici, visivi, uditivi, sensoriali e potenzialmente motori, cruciale per costruire rappresentazioni unificate del mondo e abilità percettivo-cognitive coerenti. Secondo, l'apprendimento continuo (continuous learning) è necessario per permettere ai sistemi di aggiornarsi in modo non distruttivo, evitando il

cosiddetto “catastrophic forgetting” e adattandosi a lungo termine in ambienti evolutivi. Terzo, il meta-apprendimento (“learning to learn”) puntando a dotare i modelli di meccanismi di acquisizione rapida di nuove competenze, trasferendo efficacemente conoscenze pregresse a domini inediti. Quarto, il ragionamento causale, distinto dal semplice apprendimento statistico di correlazioni, risulta fondamentale per la comprensione delle relazioni causa-effetto, la previsione sotto intervento e l’abilità di pianificare azioni efficaci in contesti complessi.

Le stime temporali su l'AGI, pur controverse, convergono in diversi sondaggi di esperti su mediane che oscillano tra il 2040 e il 2060. Da un lato, le scaling laws suggeriscono che l’aumento di dati, calcolo e dimensione dei modelli possa tradursi in progressi non lineari, alimentando la prospettiva di una possibile accelerazione. Dall’altro, permangono colli di bottiglia sostanziali: vincoli energetici e computazionali, limiti architetturali e algoritmici, nonché sfide di allineamento e sicurezza che potrebbero rallentare o riallineare le traiettorie di sviluppo. In questo senso, la transizione verso l’AGI appare come il risultato di un equilibrio delicato tra forze di accelerazione e fattori frenanti, all’interno di un ecosistema socio-tecnico in profonda trasformazione.

Quadro sinottico delle differenze principali

La tabella seguente sintetizza le differenze chiave identificate nell’analisi:

Tabella 7. Intelligenza umana vs LLM/Agenti

Dimensione	**Intelligenza Umana**	**LLM/Agenti**	**Implicazioni**
Substrato	Neurobiologico (86 miliardi neuroni)	Computazionale (miliardi parametri)	Differenze nell’efficienza energetica e modalità di elaborazione

Dimensione	**Intelligenza Umana**	**LLM/Agenti**	**Implicazioni**
Apprendimento	Few-shot, embodied	Data-intensive, pre-training + fine-tuning	Trade-off tra efficienza di apprendimento e scalabilità
Generalizzazione	Transfer flessibile tra domini	Limitata al distribution shift	Robustezza vs adattabilità
Ragionamento	Sistema 1 + Sistema 2, insight-based	Chain-of-thought, statistico	Complementarità tra intuizione e computazione
Coscienza	Esperienza soggettiva verificabile	Simulazione senza fenomenologia	Questioni etiche e di attribuzione di status
Creatività	Intenzionalità artistica, meaning-making	Ricombinazione statistica	Autenticità contro originalità formale
Social cognition	Capacità innata di comprendere gli stati mentali altrui + empatia.	Simulazione di competenze sociali	Implicazioni per interazione e fiducia

Dimensione	Intelligenza Umana	LLM/Agenti	Implicazioni
Robustezza	Adattamento a contesti perturbati	Sensibilità ai cambiamenti nel tipo di dati tra addestramento e utilizzo.	Affidabilità in applicazioni critiche
Spiegabilità	Limitata introspezione	Opacità computazionale	Sfide per l'interpretabilità e la fiducia
Scalabilità	Limitata da vincoli biologici	Scalabile con risorse computazionali	Potenziale per superintelligenza

7.10. Complementarità e sinergie potenziali

L'analisi condotta suggerisce che intelligenza umana e artificiale non sono necessariamente in competizione zero-sum, ma possono svilupparsi in direzioni complementari: gli umani mantengono vantaggi in creatività genuina, meaning-making, ragionamento etico, moral judgment, intuizione, insight problem-solving e adattabilità a contesti nuovi e perturbati. I sistemi di IA eccellono in: elaborazione di grandi quantità di informazioni, identificazione di pattern in dataset complessi, ottimizzazione multi-dimensionale, consistenza e scalabilità operativa.

Sistemi ibridi: Le configurazioni più promettenti potrebbero combinare:

- Human judgment per obiettivi e valori.
- AI computation per elaborazione e ottimizzazione.
- Human oversight per accountability ethics.
- AI scalability per un implementazione su larga scala.

7.10.1. Domande aperte e direzioni di ricerca

L'analisi comparativa rivela diverse questioni irrisolte che richiedono ricerca futura:

Come misureremo l'intelligenza generale in sistemi che operano su substrati computazionali diversi?

- Quali soglie dimensionali potrebbero produrre salti qualitativi verso AGI?

- Come identificheremo l'emergere di coscienza genuina in sistemi artificiali?

- L'intelligenza generale è substrate-independent o richiede embodiment biologico?

- Possono esistere forme di intelligenza radicalmente aliene rispetto a quelle umane?

- Come si evolverà la definizione stessa di intelligenza nell'era dell'IA avanzata?

- Come progettare sistemi di valutazione che evitino data contamination?

- Quali framework normativi governeranno sistemi potenzialmente coscienti?

- Come mantenere human agency in ecosistemi dominati dall'IA?

La ricerca futura richiederà approcci genuinamente interdisciplinari che integrino computer science, neuroscienze cognitive, filosofia della mente e etica applicata, per navigare le trasformazioni profonde che l'IA avanzata sta introducendo nella comprensione dell'intelligenza e della cognizione.

8. RIFLESSIONI E PROSPETTIVE

Premessa

Verranno asserite ulteriori riflessioni rispetto a quanto detto finora, rafforzando i legami di senso tra i vari capitoli. Si sottolineeranno nuovamente gli intenti dietro le argomentazioni, riducendo a zero qualsiasi fraintendimento, non escludendo, ovviamente, la possibilità di interpretazioni divergenti, se non arricchenti, di critica, di completamento o di qualsivoglia tipo.

Inoltre, in virtù del fatto che durante il processo di scrittura possano essere cambiate strada facendo alcune consapevolezze, alcune percezioni, è nell'interesse di questo capitolo farle presente, a prova del fatto che questa tesi, quantomeno nell'autore, ha prodotto uno degli effetti desiderati: aver portato (nella continua rilettura delle argomentazioni originali e rielaborate) a profonde riflessioni d'insieme, nei riguardi di questo immenso fenomeno ecosistemico.

L'impronta di questo capitolo non sarà tecnica, non includerà fonti o rielaborazioni di argomenti scientifici. Sarà un'impronta personale, nello stile, nella narrativa, nel linguaggio e nell'analisi complessiva. Un'interpretazione che seguirà una versione di cuore dei fatti.

La Struttura dell'Elaborato: Equilibrio Tecnico-Umanistico

La struttura dell'elaborato ha seguito uno schema di costante equilibrio tecnico-umanistico, per fare sì che si potesse fornire una cornice su misura, al fine di inquadrare, strutturare, desumere,

ragionare, capire, quantificare dati e informazioni da questa enorme scatola nera, inserita nelle sue interrelazioni organiche con le dimensioni del mondo a cui ha accesso.

Una differenza sostanziale rispetto a quello che avrebbe potuto dare un'analisi isolata (che ha i suoi vantaggi) ma non ai fini di un obiettivo di ricerca come questo. Per necessaria professionalità, solidità e consistenza, ci si è dovuti attenere a standard esplicativi, soprattutto nella parte tecnica, ma così anche nella parte umanistica.

I Limiti di un Approccio Formale

Questa metrica di scrittura ha molti pregi, ma anche dei difetti ai quali si cercherà in questo capitolo di porre rimedio.

Il primo è innanzitutto quello di far risultare la lettura poco digeribile, appetibile, di facile accesso. Ed è un tema che per le argomentazioni tecniche del mondo della letteratura scientifica non vale nulla. E questo comporta che, nonostante si viva in un'era profondamente alfabetizzata rispetto ai secoli passati, con grandi possibilità di accesso alle informazioni, pochi sono gli individui che approfondiscono a un livello particolare, pochi rispetto a quanti effettivamente potrebbero (nessun dato a supporto, una percezione individuale). E in questo lavoro si crede che una delle cause sia quanto sopra citato, oltre che, ovviamente, altri problemi, come l'overloading informativo.

Un altro difetto, che definiamo secondario, perché a carattere soggettivo, è che il testo sembri non avere anima. Laddove per anima, si intende quella intrinseca originale essenza presente in ognuno di noi, e quindi in questo caso nel testo che scriviamo. Quella caratteristica che, da un punto di vista funzionale, permetterebbe (applicata a questo contesto) di scrivere con una forte creatività e identità, elementi imprescindibili per catturare gli spazi invisibili delle dimensioni in cui viviamo.

Il Valore dell'Anima nella Scrittura

Oltre a permettere opere inedite, astrazioni sottili spesso impercettibili, che possono essere colonne portanti di innovazione, idee, evoluzioni culturali, scientifiche (lavorando in stretta collaborazione con essa). Lo spunto è quello di sollecitare il pensiero laterale, di non aver paura nell'esprimere con audacia le proprie intuizioni, di uscire dagli schemi, dai protocolli (senza demonizzarli), e attuare il processo di dare nuova vita alle cose, perché questa è una delle nostre più fantastiche capacità: "dare vita", con progetti, pensieri ed artefatti di qualsivoglia tipo.

Basti vedere cosa si sta realizzando con l'AI. E l'intento è stato quello di mettere il cuore in questo lavoro, facendo trasparire l'incredibile particolarità dei sistemi complessi. Non è sufficiente scrivere la lista degli algoritmi in tecnichese per far comprendere al mondo le nuove invenzioni, serve mettere anima nel processo esplicativo. Ed è questo uno degli intenti dell'elaborato, e del capitolo finale: sperare che qualche esperto possa aprire il proprio orizzonte deterministico nei riguardi di una metodologia che possa includere spazi dimenticati.

Preservare l'Unicità nell'Era dell'Automazione

E se non si è d'accordo con queste argomentazioni, vi è comunque un altro motivo, che in maniera indipendente renderebbe ad ogni modo valida l'idea di utilizzare un linguaggio d'anima. Ovvero che mai come ora è necessario preservare e mettere in evidenza le proprie unicità, le proprie caratteristiche intrinseche.

In un'epoca di profonda superficializzazione delle manifestazioni d'intelletto, in un'epoca di quantità efficiente, piuttosto che di pura distinguibile qualità, urge la necessità di far emergere le proprie particolarità, per non finire nella scala di grigi, ma essere a proprio modo un elemento profondamente ispirante per il mondo a venire. E questo sarebbe funzionale al miglioramento della società in generale.

In un futuro dove un'IA efficiente automatizzerà qualsiasi processo ripetitivo e determinato, il valore univoco di un prodotto come gli artefatti d'anima diverrà assoluta essenzialità. Vi sarà un'alta probabilità che all'uomo resteranno in mano i ruoli più "umani", meno sostituibili dalle macchine, e dunque più che mai interiori, divenendo l'atto di porre anima nelle cose non solo una

pura esigenza qualitativa, ma un meccanismo solido del mercato. In un futuro di impronte uguali, il "fare differenza" sarà un bene di lusso.

Le Sfide del Futuro: Identificazione e Autenticità

Dovremo aspettarci un futuro con difficoltà nell'individuazione delle fonti creative, nella fattispecie: artefatti virtuali (notizie, video, audio, hybrid-content) e artefatti fisici. Importante sarà dunque agire con prontezza nel saper identificare e inquadrare le estrinsecazioni creative al meglio, sia per questioni etiche (sapere che si sta visionando o facendo uso di un prodotto AI-generated) sia per implicazioni pratiche (distinguere eventi reali da fake news diffuse da ambienti generativi o necessità di chiarezza sul processo produttivo per la circoscrizione di responsabilità e adozione dei dovuti protocolli/contromisure).

Collaborazione, Non Polarizzazione

Allo stesso tempo attenzione a pensare che si voglia creare una polarizzazione creativa che metta contro umani e AI. Sarebbe l'errore più grande. L'AI deve essere vista come uno strumento a rafforzamento dell'estro umano. Ed è grazie a un efficiente utilizzo dell'AI che le persone potranno tornare a concentrarsi su quel che più sanno fare meglio. Oltre a ciò, forme di collaborazione AI-humans potranno alzare gli standard produttivi significativamente (in termini di efficienza) in una miriade di settori.

Per questo è importante vedere le due forme creative come un prospetto organico-collaborativo, in un perfetto ecosistema integrato, nel quale uomo e macchina si dedicano e migliorano costantemente a vicenda.

Rischi e Opportunità: Oltre l'Utopia

Detto ciò, è ovvio che sia stato argomentato lo scenario più utopico dei tanti che potrebbero presentarsi in futuro. Come per ogni tipo di innovazione storica, se ne vedrà un utilizzo per scopi più ragionati, etici, migliorativi per la quotidianità di molti, e un utilizzo sporco, personale, commisurato a strategie basate su vantaggi individuali o meramente economici, ai danni di molti.

Il primo passo dovrà essere costruire una coscienza collettiva sul fenomeno. Sarà utile richiamare all'attenzione esperti tecnici, esperti umanisti, amatori, persone che contribuiscono alla quotidianità di questo mondo, per fornire strutture di ampia visuale che serviranno da incipit per avere i propri framework interpretativi da utilizzare come cosciente contributo all'evoluzione di una nuova era tecno-sociologica, dove la tecnologia si fa elemento antropomorfo e soggetto attivo (agente) delle forme di comunicazione quotidiane, toccando di fatto ogni interstizio dell'interazione umana e tecnologica.

L'Approccio Transdisciplinare

A tal fine, ogni forma di sapere diverrà un utile contributo, dalla Filosofia alla Matematica, dalla Biologia alla Linguistica. Doveroso sarà dunque applicare un focus transdisciplinare, affinché si possa attuare quella accuratezza e contestualizzazione che altrimenti rimarrebbe lasciata al caso, a quegli spazi vuoti che in silenzio urlano la loro importanza.

Il Percorso della Tesi: Dall'Intelligenza alla Coscienza

La tesi ha cercato di fare questo. Partendo con un focus sulla possibilità e la curiosità che questi sistemi potessero arrivare a forme di intelletto egualmente o superiori alle nostre, sotto tutti gli aspetti, o alcuni di essi, e che in virtù di ciò potessero contribuire a deduzioni fondamentali sulla nostra natura, fungendo da specchio per i nostri meccanismi interiori.

I Sistemi AI come Specchio dell'Umano

Specchio perché quanto più questi sistemi emulano alcune nostre caratteristiche, quanto più possiamo comprendere quelle stesse caratteristiche, in virtù anche del fatto che le abbiamo effettivamente applicate a un sistema esterno a noi con risultati similari. E per "comprendere" si intende una comprensione profonda e interiore, che guardi non tanto al risultato, all'esito, ma al processo in sé.

Quanto più si realizza un sistema emulante le tue stesse funzioni, quanto più quello stesso sistema fungerà da prototipo verso il quale l'essere umano stesso potrà specchiarsi, studiarsi, analizzarsi, in maniere sempre univoche. E questa è la più grande innovazione apportata finora.

I Risultati della Ricerca

Lo studio, procedendo in un percorso di destrutturazione, contestualizzazione e comparazione con le forme d'intelletto umano, ha potuto dimostrare, in una somma di evidenze, che questi oggetti, seppur ancora da scoprire, con diversi meccanismi nelle loro profondità da decifrare, sono uno strumento ideale, una lente su misura, per la comprensione dei processi interiori umani, anche se il sistema in cui vengono riprodotti è composto da substrati differenti.

Il nostro studio ha tuttavia anche evidenziato che tali sistemi, già molto risolutivi, sono ancora lontani da un'AGI, mancando di diverse caratteristiche fondamentali ben descritte nell'elaborato. Indi per cui si è sempre preferito adottare il termine "esperto", estendendolo anche ai sistemi pluralistici che peccano di forme adeguate di coscienza e intelligenza. Concordando con l'idea che questi oggetti al momento non possano essere più che strumenti (parlando di LLMs) emulanti alcune caratteristiche umane, ottimi assistenti dominio-specifici, che performano allo stato dell'arte mediamente più efficientemente dell'essere umano.

Le Evidenze Emerse: Architetture Neurali e Cognizione

Ad ogni modo gli spiragli aperti sono molti, a partire dagli esperimenti sulle diverse architetture neurali, che paragonate agli studi sui processi cognitivi del cervello umano, mostrano svariati punti di contatto. Per non parlare dell'emersione del fattore generale g, quindi un fattore di intelligenza correlato a più domini. Concludendo con le evidenze strutturanti una metodologia di funzionamento dei processi cognitivi/computazionali simili fra loro, cioè un'organizzazione in moduli, i quali separatamente elaborano in maniera specializzata.

Ricorsività e Global Workspace Theory

Altri studi citati nel lavoro hanno poi mostrato anche una capacità ricorsiva del cervello, mettendo in risalto l'importanza della rielaborazione iterativa delle informazioni, evidenziando un'interazione molto più complessa, intermodulare e organica. In queste interrelazioni, diversi moduli entrano in contatto fra loro, condividendo informazioni, collaborando, elaborando in un formato estremamente profondo e integrato.

E ci sono reti neurali che si avvicinano a questi meccanismi, non il Transformer, ma ad esempio le RNN (Recurrent Neural Network), le quali mettono in atto meccanismi di ricorsività. In aggiunta a queste reti, la GWT (Global Workspace Theory) ha posto in evidenza il concetto di spazio di lavoro globale, una strutturazione elaborativa delle informazioni che, se confermata come corretta riproduzione dell'elaborazione informativa del cervello, potrebbe essere riproducibile attraverso meccanismi computazionali.

La Questione della Coscienza

Ovviamente ciò non basta, ci sarebbe la questione dell'embodiment e della coscienza. Il fenomeno dell'esperire coscienza sembra essere il più difficile da riprodurre. Nel nostro elaborato sono state presentate teorie anche per sperimentare l'ottenimento di questa caratteristica, come nel caso delle teorie della coscienza di ordine superiore, ma rimane molto più difficile, in questo caso, riuscire a catturare chiaramente delle mappature di features solide, riutilizzabili.

Saranno necessarie ulteriori sperimentazioni e riflessioni, soprattutto in ambito filosofico. Non vi è ancora facilità nell'individuare con certezza il fenomeno cosciente nei sistemi artificiali, anche qualora esso possa essersi già manifestato, riferendosi principalmente ai casi di intenzionalità cosciente, e non di semplice emulazione formale della stessa. La dinamica è molto complessa e necessita maggiori punti fermi.

Il Mistero della Scatola Nera

Il fatto che questi oggetti siano indecifrabili nel processo di mezzo delle loro elaborazioni è l'elemento che lascia più curiosità nella riepilogazione del tutto. Apre delle ipotesi per le quali questi sistemi possano essere in realtà più coscienti di quanto crediamo, magari non nella

forma/modalità a cui noi siamo abituati, rimanendo d'accordo con l'idea dello studio preso a base dell'elaborato che siano accettate forme di coscienza anche parziali.

Conclusioni: Un Framework Interpretativo Equilibrato

A somma di tutto, dall'inquadramento macro effettuato, ne deriva che è precipitoso nel 2025 parlare di sistemi intelligenti e coscienti, ma è anche precipitoso parlare di impossibilità di coscienza e intelligenza in un futuro a breve-medio termine. Giusto invece, parlare di ottimi strumenti-assistenti, risolutivi in chiave cooperativa, delegativa (con supervisione) a braccetto con protocolli di sicurezza e policy adeguate. Giusto è anche parlare di ottimi oggetti per studiare e avere maggior cognizione dei nostri processi interni, per i motivi spiegati prima. Doveroso è, seguire costantemente l'evoluzione degli stessi, in rapporto all'ecosistema in cui operano, in chiave organico-integrativa, e con tutti i mezzi a disposizione. Sbagliato invece, guardare quest'evoluzione come un film distopico con il peggiore dei finali. Banalmente non siamo in un film, e le dinamiche ecosistemiche sono molto più complesse, e non vi è uno sceneggiatore o un regista dietro. Il miglior framework interpretativo potrebbe essere quello di guardare le cose con profonda coscienza, sui rischi e sui benefici che artefatti come questi possono apportare nel complesso sistema mondo.

9. METODOLOGIA DELLA RICERCA

Il presente capitolo esplicita l'impianto metodologico che ha guidato l'intero percorso di ricerca, delineando gli approcci teorici, le strategie operative e gli strumenti analitici impiegati per rispondere alla domanda centrale della tesi: in che misura i Large Language Models (LLMs) possono contribuire a una maggiore comprensione della natura umana? La costruzione di un quadro metodologico rigoroso e trasparente è fondamentale non solo per garantire la replicabilità e la validità della ricerca, ma anche per rendere esplicite le scelte epistemologiche e operative che

hanno orientato l'analisi. In un campo interdisciplinare come quello degli studi sull'intelligenza artificiale, dove convergono prospettive tecniche, cognitive, filosofiche e comunicative, la chiarezza metodologica diviene condizione imprescindibile per un'indagine coerente e sistematica.

9.1. DOMANDE DI RICERCA

9.1.1. Questione fondamentale e sottoproblemi

Questa tesi, dal titolo "Ecosistema LLM", nasce da una domanda di ricerca fondamentale: **fino a che punto l'intelligenza artificiale e in particolare i modelli linguistici di grandi dimensioni (LLMs) possono aiutarci non tanto a estendere il nostro sapere tecnico, quanto a illuminare aspetti profondi dell'esperienza e dell'identità umana?**

Questa questione centrale si articola in diverse sotto-domande che hanno guidato l'impianto metodologico:

1. **Domanda tecnico-cognitiva**: Come funzionano i LLMs a livello architetturale e quali capacità cognitive simulano o replicano?
2. **Domanda comparativa**: Quali sono le similarità e differenze fondamentali tra intelligenza biologica e artificiale, e cosa rivelano sull'organizzazione della cognizione?
3. **Domanda mediologica**: Come si stanno trasformando gli ecosistemi comunicativi con l'integrazione di agenti artificiali, e quali implicazioni hanno per le relazioni umane?
4. **Domanda filosofica**: L'analisi dei sistemi artificiali può fornire insights genuini sulla natura dell'intelligenza, della coscienza e dell'identità umana?
5. **Domanda epistemologica**: Quali framework teorici e metodologici sono più appropriati per studiare fenomeni che si collocano all'intersezione tra scienze cognitive, informatica e studi umanistici?

9.1.2. Contributo originale e posizionamento disciplinare

Il contributo originale della ricerca consiste nell'adottare una prospettiva **genuinamente interdisciplinare** che utilizza l'IA non come oggetto di studio isolato, ma come **lente euristica** per comprendere meglio la natura umana. Questo approccio si distingue da:

- **Studi puramente tecnici** che si concentrano su performance e ottimizzazione algoritmica

- **Analisi filosofiche** che rimangono disconnesse dai sviluppi empirici
- **Ricerche applicative** che considerano l'IA solo come strumento pratico

Il posizionamento disciplinare è **transdisciplinare**, integrando metodologie e paradigmi teorici da: - Informatica e AI research - Scienze cognitive e neuroscienze - Filosofia della mente e epistemologia - Studi mediologici e teoria della comunicazione - Psicologia cognitiva e psicometria.

9.2. FRAMEWORK METODOLOGICO GENERALE

9.2.1. Paradigma epistemologico: realismo critico

La ricerca è configurata come uno studio interpretativo multimetodo, che integra diverse strategie analitiche al fine di comprendere le capacità emergenti dei modelli linguistici di grandi dimensioni (LLM). In particolare, essa combina un'analisi documentaria sistematica, una sintesi teorica comparativa, un modelling concettuale e una triangolazione di prospettive disciplinari, al fine di garantire un approccio robusto e multidimensionale. La strategia generale segue una logica abduttiva, come delineato da Peirce (1903) e successivamente sviluppata da Tavory e Timmermans (2014), articolata in quattro fasi principali: l'osservazione dei fenomeni, relativa alle capacità emergenti dei LLM; la formulazione di ipotesi esplicative, attraverso il confronto con processi di cognizione umana; l'esplorazione teorica, mediante l'applicazione di framework interpretativi interdisciplinari; e, infine, l'elaborazione di insights, che consente di trarre implicazioni più ampie per la comprensione dei processi cognitivi e del comportamento umano. Questo approccio consente di coniugare evidenze empiriche e riflessioni teoriche in un quadro analitico coerente e integrato.

9.2.2 Design della ricerca

Temporalità: Studio **sincrono** con focus sullo stato dell'arte (2024-2025), ma con attenzione a **traiettorie evolutive** per contestualizzare sviluppi più recenti.

Articolazione metodologica per obiettivi

Obiettivo 1 - Comprensione tecnica: Metodologia **ingegneristica-descrittiva** - Analisi di documentazione tecnica e paper scientifici - Sintesi di architetture e funzionamenti - Traduzione da linguaggio tecnico a quadri concettuali accessibili.

Obiettivo 2 - Confronto intelligenze: Metodologia **comparativa-analitica** - Applicazione di framework psicometrici a sistemi IA - Confronto sistematico su dimensioni cognitive specifiche - Identificazione di pattern convergenti e divergenti.

Obiettivo 3 - Analisi ecosistema: Metodologia **mediologica-interpretativa** - Applicazione di teorie dei media e comunicazione - Analisi delle trasformazioni nelle pratiche comunicative - Identificazione di effetti emergenti e non intenzionali.

Obiettivo 4 - Insights filosofici: Metodologia **filosofico-speculativa** - Utilizzo di esperimenti concettuali e analisi logica - Esplorazione di implicazioni ontologiche ed epistemologiche - Articolazione di possibili scenari futuri.

9.3. SELEZIONE E ANALISI DELLE FONTI

9.3.1. Corpus documentario e criteri di inclusione

Il lavoro si basa su un **corpus diversificato e multi-linguistico** che include diverse tipologie di fonti:

Fonti primarie tecniche - Paper peer-reviewed su AI/ML (Nature, Science, ICML, NeurIPS, ACL) - Technical reports da laboratori industriali (DeepMind, OpenAI, Anthropic) - Documentazione di sistema e blog di ricerca ufficiali - Preprint su arXiv.

Fonti teorico-umanistiche : Monografie accademiche su filosofia della mente e scienze cognitive - Articoli su riviste interdisciplinari (Minds & Machines, AI & Society) - Testi fondativi di teorie dei media e comunicazione - Letteratura italiana contemporanea su IA e società.

Criteri di inclusione: - Rilevanza teorica per framework interpretativi - Autorevolezza accademica (editori e riviste riconosciute).
- Capacità di dialogo con sviluppi tecnologici contemporanei - Disponibilità in italiano/inglese.

Fonti empiriche e dati : - Benchmark e dataset per valutazione IA - Studi psicometrici applicati a LLMs - Survey e report sull'adozione di tecnologie IA - Case studies di implementazioni pratiche.

Criteri di inclusione: - Metodologia rigorosa e trasparente - Campioni significativi o dataset rappresentativi - Disponibilità dei dati - Rilevanza per domande di ricerca specifiche.

Criteri di inclusione generali: I punti chiavi che sono stati seguiti per la scelta delle fonti si rivedono in un'integrazione eterogeneamente pluriprospettica, all'interno di un equilibrio tecnico e interpretativo, adoperante una metodologia archeologico-evolutiva e allo stato dell'arte, con extra integrazioni riflessive e di contesto.

9.3.2. Strategia di ricerca sistematica

Database consultati: - **Accademici**: Google Scholar, JSTOR, Web of Science, IEEE Xplore - **Specialistici**: arXiv.org, Papers With Code, Semantic Scholar - **Industriali**: Siti ufficiali di laboratori AI, blog di ricerca.

9.4. TECNICHE E STRUMENTI DI ANALISI

9.4.1. Analisi contenutistica qualitativa

L'analisi tematica è stata condotta seguendo il framework proposto da Braun e Clarke (2006), articolato in sei fasi principali. La prima fase, **familiarizzazione**, ha previsto una lettura immersiva del corpus e l'annotazione dei contenuti rilevanti. Successivamente, nella fase di **generazione dei codici**, sono stati identificati pattern ricorrenti e concetti chiave presenti nei dati. Nella fase di **identificazione dei temi**, i codici sono stati raggruppati in categorie coerenti, seguita dalla fase di **revisione dei temi**, in cui è stata verificata l'**omogeneità interna** e l'**eterogeneità esterna** dei temi stessi. La quinta fase, **definizione dei temi**, ha comportato l'articolazione precisa del contenuto tematico, mentre la fase finale, **redazione del report**, ha integrato la narrazione interpretativa con evidenze testuali direttamente estratte dal corpus.

Per quanto riguarda l'**affidabilità inter-valutatore**, pur essendo l'analisi stata condotta da un singolo ricercatore, la coerenza e la validità delle interpretazioni sono state garantite attraverso discussioni con i supervisori, confronti con schemi di codifica già presenti in letteratura e una **documentazione dettagliata** di tutte le decisioni interpretative adottate. Questo approccio ha permesso di massimizzare la trasparenza e la robustezza dei risultati, pur in un contesto di analisi qualitativa condotta da un unico valutatore.

9.5. FRAMEWORK TEORICI DI RIFERIMENTO

Questa sezione esplicita i principali framework teorici che hanno orientato l'interpretazione dei dati e la costruzione delle argomentazioni. I framework provengono da discipline diverse, ma sono stati integrati in un quadro coerente e sistemico.

9.5.1. Teorie cognitive e architetture della mente

Computational Theory of Mind (CTM)

Autori di riferimento: Fodor (1975), Pylyshyn (1984)

Assunto centrale: I processi mentali sono processi computazionali che operano su rappresentazioni simboliche interne secondo regole sintattiche.

Applicazione: Fornisce un framework per confrontare architetture cognitive umane e artificiali, legittimando l'idea che sistemi computazionali possano replicare (almeno funzionalmente) processi mentali.

Implicazioni metodologiche: Possibilità di modellare e testare teorie cognitive attraverso sistemi di IA; attenzione alle distinzioni tra sintassi e semantica.

Global Workspace Theory (GWT)

Autori di riferimento: Baars (1988), Dehaene (2014)

Assunto centrale: La coscienza emerge da un processo di "broadcasting globale" delle informazioni: i contenuti accessibili a una "workspace" globale diventano coscienti, mentre le elaborazioni modulari rimangono inconsce.

Applicazione: Offre criteri di valutazione per la coscienza in sistemi artificiali e consente la discussione su architetture che implementano meccanismi analoghi (es. attenzione globale nei Transformer).

Implicazioni metodologiche: Operazionalizzazione di test per la coscienza; focus su integrazione informativa e accesso globale.

Predictive Processing (PP)

Autori di riferimento: Clark (2013), Hohwy (2013)

Assunto centrale: Il cervello è una "macchina predittiva" che minimizza l'errore di previsione attraverso modelli gerarchici del mondo.

Applicazione: Framework per comprendere l'apprendimento nelle reti neurali (minimizzazione della loss function come analogo della minimizzazione dell'errore predittivo).

Implicazioni metodologiche: Focus su accuratezza predittiva, rappresentazioni gerarchiche, apprendimento non supervisionato.

9.5.2. Teorie della comunicazione e ecologia mediale

Media Ecology

Autori di riferimento: McLuhan (1964), Postman (1970), Scolari (2018)

Assunto centrale: I media non sono semplici canali neutri, ma ambienti che plasmano la percezione umana, l'organizzazione sociale e la cultura.

Applicazione: Analisi dell'impatto degli agenti artificiali sugli ecosistemi comunicativi, con focus sugli effetti ambientali piuttosto che sui contenuti specifici.

Implicazioni metodologiche: Attenzione alle trasformazioni sistemiche, agli effetti non intenzionali e alla co-evoluzione uomo-tecnologia.

Theory of Communicative Action

Autore di riferimento: Habermas (1981)

Assunto centrale: La comunicazione autentica si basa su pretese di validità (verità, correttezza normativa, sincerità) che possono essere argomentate razionalmente.

Applicazione: Fornisce criteri normativi per valutare la comunicazione umano-IA e discute la possibilità di "autenticità" comunicativa in agenti artificiali.

Implicazioni metodologiche: Standard normativi per valutare il dialogo con IA; attenzione alle dimensioni pragmatiche ed etiche.

Semiotica

Autori di riferimento: Peirce (1903), Eco (1976)

Assunto centrale: Il significato emerge da processi semiosici (relazioni tra segni, oggetti e interpretanti) all'interno di comunità interpretative.

Applicazione: Analisi dei nuovi processi di significazione nell'interazione umano-IA; discussione su "comprensione" vs "manipolazione simbolica".

Implicazioni metodologiche: Focus sullo sviluppo di codici, pratiche interpretative e dimensione pragmatica del linguaggio.

9.5.3. Filosofia della mente ed epistemologia

Functionalism

Autori di riferimento: Putnam (1967), Lewis (1972)

Assunto centrale: Gli stati mentali sono definiti dai loro ruoli funzionali (relazioni input-output) e sono "multiply realizable" (realizzabili in substrati fisici diversi).

Applicazione: Giustifica teoricamente la possibilità di coscienza e intelligenza in sistemi artificiali non biologici.

Implicazioni metodologiche: Focus sulle relazioni funzionali piuttosto che sul sostrato fisico; attenzione agli isomorfismi strutturali.

Embodied Cognition

Autori di riferimento: Lakoff & Johnson (1999), Wilson (2002)

Assunto centrale: La cognizione è radicata nell'esperienza sensomotoria corporea; la mente non è un elaboratore simbolico astratto ma è "incarnata".

Applicazione: Analisi dei limiti dei sistemi IA disembodied (privi di corpo); discussione sulla necessità di grounding sensomotorio per una "vera" comprensione.

Implicazioni metodologiche: Attenzione ai fattori di embodiment nei confronti cognitivi; considerazione di robotica e sistemi multimodali.

Extended Mind

Autori di riferimento: Clark & Chalmers (1998)

Assunto centrale: I processi cognitivi possono estendersi oltre i confini biologici includendo

strumenti esterni (notebook, smartphone, IA).

Applicazione: Framework per pensare l'accoppiamento cognitivo umano-IA; IA come estensione cognitiva piuttosto che sostituto.

Implicazioni metodologiche: L'unità di analisi diventa il sistema cognitivo esteso (umano + strumenti); attenzione alle dinamiche di integrazione.

9.5.4. Integrazione e coerenza tra framework

I framework presentati non sono stati applicati in modo isolato, ma integrati sistematicamente:

- La **Computational Theory of Mind** fornisce la base per pensare i LLMs come sistemi cognitivi.
- La **Global Workspace Theory** e il **Predictive Processing** offrono modelli specifici di come potrebbe emergere coscienza e intelligenza.
- Le teorie della comunicazione contestualizzano i LLMs in ecosistemi socio-tecnici.
- La filosofia della mente fornisce categorie concettuali per discutere intelligenza, coscienza e identità.
- Le prospettive embodied ed extended aggiungono cautele e distinzioni necessarie.

Questa triangolazione teorica garantisce robustezza interpretativa e riduce il rischio di bias disciplinari.

9.6 STRATEGIE PER RIDUZIONE DEI BIAS

Nell'ambito di questa ricerca, la gestione dei **bias metodologici e interpretativi** è stata affrontata attraverso strategie specifiche per ciascun tipo di distorsione. Per quanto riguarda il **bias di selezione**, è stata adottata una **strategia di ricerca sistematica su più banche dati**, con criteri di inclusione chiaramente esplicitati e applicati in modo coerente. Il **bias di conferma** è stato mitigato considerando ipotesi alternative, adottando l'approccio dell'**avvocato del diavolo** nell'interpretazione dei risultati e sottoponendo periodicamente a revisione le conclusioni emergenti. Il **bias disciplinare** è stato affrontato mediante una revisione della letteratura di

carattere interdisciplinare, l'impiego di molteplici quadri teorici, la consultazione tra pari appartenenti a discipline diverse e il riconoscimento e la gestione dei punti ciechi disciplinari. Infine, il **bias culturale e linguistico** è stato ridotto attraverso l'inclusione di fonti in più lingue, la considerazione del contesto culturale nell'interpretazione delle evidenze, la consultazione di prospettive internazionali e la consapevolezza della predominanza della letteratura occidentale e anglosassone nel campo dell'intelligenza artificiale.

9.7 LIMITAZIONI METODOLOGICHE RICONOSCIUTE

Questa ricerca presenta alcune **limitazioni intrinseche** legate alle scelte metodologiche e al contesto di studio. In primo luogo, essendo una **ricerca secondaria**, essa dipende prevalentemente dalla letteratura esistente piuttosto che dalla generazione di dati primari, con la conseguente capacità limitata di produrre risultati empirici originali. Tale limitazione è stata mitigata attraverso una **copertura esaustiva delle fonti** e l'applicazione di una **metodologia di sintesi rigorosa**, bilanciando tuttavia l'ampiezza della copertura con la profondità dell'indagine originale. In secondo luogo, l'analisi è stata condotta da un **ricercatore singolo**, il che può comportare una riduzione dell'affidabilità e un aumento del rischio di interpretazioni idiosincratiche o punti ciechi. Questo bias è stato affrontato mediante **consultazioni sistematiche, peer review e trasparenza del processo**, riconoscendo tuttavia il trade-off tra efficienza e affidabilità inter-valutatore. Ulteriore criticità deriva dal fatto che il **campo di riferimento è in rapida evoluzione**, esponendo lo studio al rischio di obsolescenza e di copertura incompleta; per mitigare questo problema, l'attenzione è stata focalizzata su questioni fondative e principi duraturi, accettando il compromesso tra attualità e solidità di intuizioni senza tempo. Infine, l'adozione di un **orientamento teorico interdisciplinare** può introdurre tensioni tra l'ampiezza delle prospettive considerate e la profondità delle competenze disciplinari; tale sfida è stata affrontata mediante un approfondito ingaggio con la letteratura di base di ciascuna disciplina, bilanciando così la profondità disciplinare con l'ampiezza interdisciplinare dell'analisi.

9.8 SINTESI METODOLOGICA E CONTRIBUTI

La metodologia sviluppata in questa ricerca rappresenta un contributo significativo per lo studio interdisciplinare dell'intelligenza artificiale e delle sue implicazioni per la comprensione della natura umana. Mentre presenta limitazioni specifiche, fornisce un framework replicabile e adattabile per future ricerche all'intersezione tra tecnologia, cognizione e società.

10. RIFERIMENTI

1

Borko, H. (1968). Information science: What is it? American Documentation. https://doi.org/10.1002/asi.5090190103 .

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Cacciari, C. (2022). Psicolinguistica. Il Mulino.

CSFO (Centro svizzero di servizio Formazione professionale orientamento professionale, universitario e di carriera). (s.d.). Scienze della comunicazione. Recuperato da https://www.orientamento.ch

Gardner, H. (1985). The mind's new science: A history of the cognitive revolution. Basic Books.

Gignac, G. E., & Szodorai, E. T. (2024). The relationship between general intelligence and working memory: A meta-analysis.

Google DeepMind. (2025). AlphaEvolve: Advancing evolutionary algorithms through large language models. Nature Machine Intelligence.

Moriggi, S., & Pireddu, M. (2024). Black box: La filosofia della tecnologia nell'era dell'intelligenza artificiale. Egea.

Portinale, L. (2022). Introduzione all'intelligenza artificiale. UTET Università.

Somalvico, M., Amigoni, F., & Schiaffonati, V. (s.d.). Introduzione all'intelligenza artificiale. Politecnico di Milano.

Treccani. (s.d.). Enciclopedia online. Recuperato da https://www.treccani.it

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems.

2

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Censis. (2025, marzo). L'intelligenza artificiale nel mercato del lavoro italiano: Rapporto 2025. Fondazione Censis.

Francis, W. N., & Kučera, H. (1964). Brown Corpus manual: https://varieng.helsinki.fi/CoRD/corpora/BROWN/ .

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. Intelligence. https://doi.org/10.1016/j.intell.2024.101832

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Sbardella, L. (2025). Intelligenza artificiale: Exploring the State of the Art in Artificial Intelligence: Evolution,Technology and Modern Application.Università della Tuscia.

Somalvico, M., Amigoni, F., & Schiaffonati, V. (s.d.). Introduzione all'intelligenza artificiale. Politecnico di Milano.

Traversari (2025). Linguistica delle AI. Lingue, Letterature e Mediazione culturale. Università degli Studi di Padova.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762 .

3

Sbardella, L. (2025).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need.

4

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165 .

Karpathy, A. (2024). Deep Dive into LLMs like ChatGPT. YouTube. https://www.youtube.com/watch?v=7xTGNNLPyMI&t=585s

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-instruct: Aligning language models with self-generated instructions. arXiv:2212.10560. https://doi.org/10.48550/arXiv.2212.10560

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903 .

5

AlphaEvolve, Google DeepMind(2025). https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/AlphaEvolve.pdf .

6

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org

Chen, Y., et al. (2023). Reasoning models don't always say what they think. Anthropic.

Foerster, J., et al. (2016). Learning to communicate with deep multi-agent reinforcement learning. Advances in Neural Information Processing Systems.

Giovannetti, P., & Miconi, A. (Eds.). (2024). Il medium oggi: Da McLuhan all'intelligenza artificiale.

Habermas, J. (1981). The theory of communicative action.

McLuhan, M. (1964). Understanding media: The extensions of man. McGraw-Hill.

McLuhan, M., & McLuhan, E. (1988). Laws of media: The new science. University of Toronto Press.

Moriggi, S., & Pireddu, M. (2024). Intelligenza artificiale e i suoi fantasmi.

Nisan, N., Roughgarden, T., Tardos, É., & Vazirani, V. V. (Eds.). (2007). Algorithmic game theory. Cambridge University Press.

Media ecology theory. (2025, October 28). The Comm Spot. https://thecommspot.com/communication-basics/communication-theories/media-ecology-theory/

Şahin, E. (2005). Swarm robotics: From sources of inspiration to domains of application.Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.

SuperSummary. (n.d.). Alone Together: Why We Expect More from Technology and Less from Each Other Summary. Retrieved November 12, 2025, from https://www.supersummary.com/alone-together-why-we-expect-more-from-technology-and-less-from-each-other/summary/

Wang, D., et al. (2020). Human-AI collaboration in data science: Exploring data scientists' perceptions of automated machine learning. https://arxiv.org/abs/1909.02309

Cos'è un testo. (2022).Teoria della cooperazione testuale di Eco. PDF. Pensiero Critico. Recuperato da https://www.pensierocritico.eu/files/Cos-e-un-Testo.pdf .

Stancati, C. (2017). Umberto eco. Books & Ideas.

from https://booksandideas.net/Umberto-Eco.html .

7

Burnell, R., et al. (2023). Revealing the structure of language model capabilities. *arXiv:2308.10062*.

Butlin, P., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv:2308.08708*.

Gardner, H. (2011). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives.

Kosinski, M. (2023). Theory of Mind may have spontaneously emerged in large language models. *arXiv:2302.02083*.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge, UK: Cambridge University Press.

Turing, A. M. (1950). Computing machinery and intelligence.

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

9

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author. Retrieved from https://www.apa.org/ethics/code .

Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (Eds.). (1998). *Critical realism: Essential readings*. London, UK: Routledge .

Bhaskar, R. (1978). *A realist theory of science* (2nd ed.). Brighton, UK: Harvester Press.

Braun, V., & Clarke, V. (2006). *Using thematic analysis in psychology*. https://doi.org/10.1191/1478088706qp063oa .

Clark, A., & Chalmers, D. (1998). The extended mind.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, NY: Viking.

Eco, U. (1975). *Trattato di semiotica generale*. Milano: Bompiani.

Fodor, J. A. (1975). *The language of thought*. Cambridge, Harvard University Press.

Habermas, J. (1984–1987). *The theory of communicative action*. Boston, MA: Beacon Press.

Hohwy, J. (2013). The Predictive Mind. Oxford, UK: Oxford University Press

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books

Lewis, D. K. (1972). *Psychophysical and theoretical identifications*. Australasian Journal of Philosophy

McLuhan, M. (1964). *Understanding media: The extensions of man*. New York, NY: McGraw-Hill.

Postman, N. (1970). *The reformed English curriculum.* In A. C. Eurich (Ed.). New York, NY: Pitman.

Putnam, H. (1967). *The nature of mental states*. In N. Block (Ed.), Readings in the philosophy of psychology. Cambridge, MA: Harvard University Press

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: The MIT Press.

Scolari, C. A. (2018). *Media evolution*: *sobre el origen de las especies mediáticas*. Buenos Aires: La Marca Editora.

Tavory, I., & Timmermans, S. (2014). *Abductive analysis*.Chicago, University of Chicago Press.

UNESCO. (1997). *Recommendation concerning the status of higher-education teaching personnel*. UNESCO.

Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin & Review