Prima che spicchi il volo: l'intelligenza artificiale tra il racconto di Bostrom e la saggezza tardiva di Hegel

- Una Guida Introduttiva sulla Intelligenza Artificiale nella PA -

Ing. PhD. Luigi Lella

Indice

Indice 2
Introduzione 4
Parte I – Le definizioni di AI 6
Capitolo 1 - l'intelligenza artificiale: un viaggio nel cuore della mente artificiale 7
Capitolo 2 - AI ristretta e AI generale: una distinzione fondamentale 13
Capitolo 3 - l'alba della singolarità 17
Parte II – Lo strumento della AI 25
Capitolo 4 - gli albori dell'intelligenza artificiale: il perceptron e le prime reti neurali 26
Capitolo 5 - SE di prima generazione: fondamenti e limiti del ragionamento deterministico 31
Capitolo 6 - SE di seconda e terza generazione: dalla gestione dell'incertezza alla logica sfumata 38
Capitolo 7 - il deep learning: fondamenti, evoluzione e prospettive future 45
Capitolo 8 - le soluzioni ibride neuro simboliche 52
Capitolo 9 - quando il senso comune manca 56
Capitolo 10 - l'intelligenza di sciame: un modello alternativo 60
Capitolo 11 – evoluzione della AI agentica: dalle logiche decidibili agli agenti LLM 66
Capitolo 12 – pattern di AI agentica per la pubblica amministrazione 73
Capitolo 13 – l'intelligenza artificiale privata: un nuovo paradigma per la pubblica amministrazione 80
Capitolo 14 – oltre il silicio: le nuove architetture di elaborazione per l'intelligenza artificiale 84
Capitolo 15 – l'evoluzione dei modelli LLM: lungo la strada che porta all'AGI 87
Parte III – L'implementazione della AI 91
Capitolo 16 - la gestione del dato come fondamento della AI 92
Capitolo 17 - l'etica della AI: un imperativo per il futuro 97
Capitolo 18 - AI Act: il regolamento europeo sull'intelligenza artificiale 104
Capitolo 19 – L'AI nel contesto italiano 110
Capitolo 20 - La codifica dei principi etici nella AI: tentativi e dilemmi 118
Capitolo 21 - il modello di maturità per l'integrazione dell'AI nelle organizzazioni 124
Capitolo 22 - l'intelligenza artificiale nella pubblica amministrazione: come avviare un progetto 129
Conclusioni 148
Biografia dell'autore 151
Riferimenti bibliografici 153

Versione attuale: 1.6
Data di rilascio: 06/05/2026

Questo saggio realizzato da Luigi Lella, è rilasciato sotto licenza Creative Commons Attribuzione 4.0. E' possibile copiare, distribuire, esporre e rappresentare in pubblico, modificare e creare opere derivate da questo lavoro, a condizione che venga riconosciuta la paternità dell'opera originale come indicato qui: **CC BY 4.0**.

Introduzione

L'intelligenza artificiale è oggi al centro di un'accelerazione tecnologica senza precedenti, la cui portata trasforma profondamente ogni aspetto della vita umana. Ma mentre ci affanniamo a sviluppare sistemi sempre più complessi, capaci di apprendere, decidere e persino creare, rimane aperta una domanda cruciale: siamo davvero in grado di comprendere ciò che stiamo costruendo?

Due immagini potenti, tratte da ambiti molto diversi — la teoria dei rischi esistenziali di Bostrom[1] e la filosofia di Hegel — ci offrono una chiave di lettura per questo interrogativo. Da un lato, il racconto dei passeri del filosofo svedese Nick Bostrom, nella quale una comunità di uccelli decide di allevare una civetta per farsi aiutare nei compiti più ardui, senza però riflettere a fondo sul rischio che la creatura, una volta cresciuta, possa distruggerli. Dall'altro, la nottola di Minerva evocata dal filosofo tedesco Friedrich Hegel, che "spicca il volo sul far del crepuscolo": metafora della filosofia, capace di comprendere la realtà solo a posteriori, quando i processi storici sono ormai giunti al loro compimento.

Mettere a confronto queste due narrazioni — la favola tecnologica e l'allegoria filosofica — significa esplorare il paradosso della nostra epoca: mentre diamo forma a un'intelligenza potenzialmente superiore alla nostra, siamo ancora sprovvisti degli strumenti etici, politici e culturali per orientarne lo sviluppo.

Questo saggio indaga il punto di intersezione tra l'azione che precede la riflessione (Bostrom) e la comprensione che arriva solo al tramonto degli eventi (Hegel), nel tentativo di comprendere se esista ancora uno spazio per una saggezza anticipatrice che eviti l'irreparabile: prima che la civetta spicchi il volo.

Quest'opera è frutto di un lavoro continuo e collaborativo, volto a esplorare le molteplici sfaccettature dell'AI, dalle sue implicazioni etiche e sociali alle sue applicazioni pratiche, con un focus particolare sul contesto della pubblica amministrazione. L'obiettivo è rendere la conoscenza sull'AI accessibile a tutti, promuovendo una comprensione approfondita e responsabile di questa

[1] Bostrom N. (2014), *Superintelligenza – tendenze, pericoli e strategie*. Bollati Boringhieri.

tecnologia trasformativa. Per questo motivo, il saggio è pubblicato sotto licenza CC BY 4.0 che ne favorisce la diffusione e l'adattamento.

La scelta di realizzare questo testo di base sulla AI discende direttamente dalle richieste mandatorie delle linee guida AgID per l'adozione della AI nella pubblica amministrazione, secondo le quali le PPAA << DEVONO promuovere un uso responsabile ed efficace dell'IA >>, << DEVONO sviluppare competenze specifiche per poter governare e regolamentare l'utilizzo dell'IA >>, e << DOVREBBERO promuovere e incoraggiare una cultura dell'apprendimento continuo improntata alla formazione e all'arricchimento costante delle competenze di IA >>. Questo testo è appositamente studiato per recepire anche l'invito a << contribuire a promuovere l'*AI awareness*[2] e *AI literacy*[3] dei cittadini, affinché riescano ad interagire in modo efficace, consapevole e sicuro con soluzioni di IA >>.

Questo saggio è un'opera in evoluzione. Riconoscendo la rapidità con cui il campo dell'intelligenza artificiale progredisce e la necessità di aggiornare costantemente le informazioni e le analisi, il saggio sarà soggetto a revisioni e aggiornamenti periodici. Per garantire trasparenza e tracciabilità delle modifiche, ogni nuova versione sarà identificata da un numero di versione e una data di rilascio.

[2] L'*AI awareness* viene descritta da AgID come una comprensione preliminare dell'esistenza e dell'influenza dell'AI. Permette cioè di riconoscere la presenza dell'AI nei sistemi e strumenti di lavoro quotidiani e di comprendere l'impatto che essa può avere sul proprio ruolo nell'ambito di un'organizzazione, sui processi e sui servizi erogati.

[3] L'*AI literacy* viene descritta da AgID come una sorta di padronanza critica e informata delle applicazioni e delle implicazioni della AI. Significa avere conoscenze operative e normative sull'AI, comprendere come e perché l'AI influisca sul lavoro, inclusi aspetti legali, etici, operativi e tecnici. Significa anche avere la competenza di governare e utilizzare in modo informato e responsabile le applicazioni AI considerando rischi, trasparenza, protezione dei dati e principi etici.

Parte I – Le definizioni di AI

Capitolo 1 - l'intelligenza artificiale: un viaggio nel cuore della mente artificiale

Fin dai tempi più antichi, l'umanità ha sognato di creare esseri o macchine capaci di pensare, di apprendere, di ragionare. Dalle leggende di Golem e automi meccanici, fino ai robot della fantascienza, l'idea di un'intelligenza non biologica ha sempre affascinato e, a volte, spaventato. Oggi, questo sogno si è concretizzato in una realtà che chiamiamo intelligenza artificiale, o AI. Ma cos'è veramente l'AI? È una magia, una minaccia, o semplicemente uno strumento potente nelle mani dell'uomo?

L'intelligenza artificiale è un campo vasto e in continua evoluzione, che tocca ogni aspetto della nostra vita, spesso senza che ce ne rendiamo conto. Dai suggerimenti personalizzati sui nostri smartphone, ai sistemi che guidano le auto a guida autonoma, l'AI è ovunque. Ma proprio per la sua pervasività e la sua complessità, è facile sentirsi disorientati. Questo capitolo è un invito a intraprendere un viaggio, un percorso per demistificare l'AI, per comprenderne le basi, le sue origini e il suo potenziale, rendendola accessibile a tutti, senza tecnicismi inutili. Vogliamo esplorare insieme non solo cosa sia l'AI, ma anche cosa significhi per noi, per la nostra società e per il nostro futuro.

Le definizioni: un mosaico di significati

Quando si parla di intelligenza artificiale, la prima domanda che sorge spontanea è: "Cos'è esattamente?". La risposta non è univoca, e le definizioni si sono evolute nel tempo, riflettendo i progressi tecnologici e le diverse prospettive.

In Italia, l'Agenzia per l'Italia Digitale (AgID) fornisce una definizione molto orientata all'applicazione pratica e all'ingegneria: "*un sistema automatizzato progettato per funzionare con livelli di autonomia variabile e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni, o decisioni che possono influenzare ambienti fisici o virtuali*" [1].

Questa visione pragmatica sottolinea come un sistema di AI sia in grado di operare con indipendenza, di modificarsi e imparare nel tempo, elaborando dati

per produrre previsioni, contenuti, raccomandazioni o decisioni che influenzano sia il mondo digitale che quello fisico. Secondo AgID, l'AI apprende dai dati, ragiona e deduce, e può modellizzare problemi complessi per supportare decisioni o fornire soluzioni innovative. Questa prospettiva è molto pratica, quasi ingegneristica, e descrive l'AI per quello che fa e per come può essere utilizzata per risolvere problemi concreti, spesso con obiettivi che sarebbero difficili o impossibili da raggiungere per un singolo individuo senza l'ausilio di queste tecnologie.

Per comprendere però appieno la genesi e l'evoluzione dell'AI, dobbiamo fare un salto indietro nel tempo, fino al 1956, anno in cui si tenne il "Dartmouth Summer Research Project on Artificial Intelligence". Questo evento, organizzato da John McCarthy, è considerato la nascita ufficiale del campo dell'intelligenza artificiale. La definizione che emerse da quel workshop, e che ancora oggi risuona, è molto più ampia e filosofica: *"ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza è, in linea di principio, descrivibile con precisione tale da poter costruire una macchina in grado di simularlo."* [2]

Marvin Minsky, uno dei pionieri dell'AI, definì l'espressione "intelligenza artificiale" una "parola valigia", un termine capace di raggruppare tutte le diverse manifestazioni dell'intelligenza: verbale, spaziale, logica, emotiva. Non si trattava solo di replicare l'intelligenza umana, ma di esplorare tutte le forme di intelligenza, comprese quelle presenti in natura (biologiche).

La differenza tra la definizione di AgID e quella di Dartmouth è evidente: la prima è focalizzata sul "cosa fa" l'AI e su come può essere utile nella pratica, mentre la seconda è più ambiziosa e scientifica, concentrandosi sul "cosa può essere" l'AI e sulla sua capacità di simulare l'intelligenza in tutte le sue forme. Entrambe le prospettive sono fondamentali per comprendere l'AI, che oggi è un ponte tra queste due visioni: un insieme di tecnologie che, pur essendo strumenti pratici per risolvere problemi, continuano a ispirarsi al grande sogno di creare macchine capaci di pensare e di apprendere in modi sempre più sofisticati.

Il contesto evolutivo: dalle visioni antiche alla realtà contemporanea

Il percorso dell'intelligenza artificiale è un viaggio affascinante che affonda le sue radici in un passato remoto, ben prima dell'avvento dei computer. L'idea di creare

esseri artificiali capaci di pensiero e azione ha attraversato miti, leggende e opere letterarie per secoli, riflettendo il desiderio umano di replicare la propria intelligenza.

Già nel XVIII secolo, autori come Jonathan Swift, nel suo celebre romanzo "I viaggi di Gulliver", immaginavano macchine capaci di generare nuove idee e testi[4], anticipando in modo sorprendente i moderni modelli di AI generativa. All'inizio del XX secolo, l'ingegnere spagnolo Leonardo Torres y Quevedo creò "El Ajedrecista", una delle prime macchine in grado di giocare a scacchi in modo autonomo, dimostrando la possibilità di automatizzare processi cognitivi complessi. E non possiamo dimenticare la nascita del termine "robot" con l'opera teatrale "Rossum's Universal Robots" (R.U.R.) di Karel Čapek nel 1920, che introdusse nell'immaginario collettivo l'idea di esseri artificiali creati per svolgere compiti umani.

Il vero punto di svolta, tuttavia, si ebbe a metà del XX secolo, con l'avvento dei primi computer e lo sviluppo di teorie matematiche e logiche che avrebbero gettato le basi per l'AI moderna. Un periodo straordinario, quello intorno al 1955, che vide protagonisti personaggi del calibro di Alan Turing, John von Neumann, John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon. Alan Turing, scomparso tragicamente nel 1954, aveva già formalizzato il concetto di "algoritmo" e ideato il test che porta il suo nome, fondamentale per i futuri studi sull'intelligenza artificiale. John von Neumann, rielaborando le idee di Turing, concepì l'EDVAC (Electronic Discrete Variables Automatic Computer), la prima macchina digitale programmabile, basata su quella che sarebbe diventata la celebre "architettura di von Neumann", ancora oggi alla base dei moderni computer [3]. Nel 1955, McCarthy, Minsky, Rochester e Shannon redassero la "proposta di Dartmouth", annunciando i temi principali del campo di ricerca, tra cui le reti neurali, la teoria della computabilità, la creatività e l'elaborazione del linguaggio naturale, e proponendo l'incontro che avrebbe segnato la nascita dell'AI.

I padri fondatori dell'intelligenza artificiale

John McCarthy (1927-2011), l'inventore del termine "intelligenza artificiale", fu un matematico e l'autore del celebre linguaggio di programmazione LISP, ancora oggi

[4] *"la macchina era congegnata in modo che ad ogni giro di manovella le parole mutassero posizione, […] pezzi di frasi: questi egli intendeva cucire insieme da ricavare da tale materia ricchissima una completa trattazione di tutte le arti e le scienze da presentare al mondo." J.Swift, (1726), "I viaggi di Gulliver", cap.V Visita dell'autore alla grande accademia di Lagado – Descrizione dell'accademia; arti e scienze in cui si esercitavano quei dotti.*

in uso. A lui si deve anche l'idea del time-sharing, un procedimento che permette a molti utenti di lavorare contemporaneamente su un singolo computer, fondamentale per lo sviluppo di Internet e del concetto di multitasking. Marvin Minsky (1927-2016), matematico, fisico, neurofisiologo e psicologo, realizzò il primo simulatore di rete neurale e fondò l'Artificial Intelligence Lab al MIT, contribuendo in modo significativo agli studi sulla logica dei calcolatori e creando il linguaggio Logo con Seymour Papert. La sua recente scomparsa nel 2016 ha ricordato il contributo di questi pionieri che hanno plasmato la tecnologia contemporanea. Nathaniel Rochester (1919-2001), ingegnere elettronico del MIT, progettò l'IBM 701, il primo computer commercialmente prodotto, e il primo linguaggio assembly simbolico, collaborando anche con McCarthy per il linguaggio LISP. Claude Shannon (1916-2001), matematico e padre della teoria dell'informazione, applicò i numeri e della logica booleana ai circuiti elettrici, introducendo il bit e ponendo le basi teoriche per la codificazione, elaborazione e trasmissione digitale dell'informazione, senza le quali la tecnologia odierna sarebbe impensabile. Shannon è celebre anche per il suo topo elettromeccanico Theseus, uno dei primi esperimenti di "learning by experience" (1952).

A questi pionieri si aggiunsero figure come Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, e in particolare Allen Newell e Herbert Simon. Newell, matematico, psicologo e informatico, e Simon, futuro premio Nobel per l'economia, presentarono il Logic Theorist, il primo programma esplicitamente progettato per imitare le capacità di problem solving degli esseri umani [4].

Gli alti e bassi dell'AI: ottimismo, delusioni e rinascita

Dopo il workshop di Dartmouth, l'AI visse un periodo di grande ottimismo, spesso chiamato la "primavera dell'AI". I ricercatori erano convinti che l'intelligenza artificiale generale (AGI), ovvero una macchina con capacità intellettuali pari o superiori a quelle umane, fosse a portata di mano. Furono sviluppati i primi programmi capaci di risolvere problemi matematici, giocare a scacchi e persino conversare in modo rudimentale. Tuttavia, le aspettative erano troppo alte e le tecnologie dell'epoca non erano ancora mature per affrontare la complessità del mondo reale. La mancanza di potenza di calcolo, la scarsità di dati e i limiti teorici di alcuni approcci portarono a una serie di delusioni, culminate nei cosiddetti "inverni dell'AI", periodi in cui i finanziamenti e l'interesse per la ricerca diminuirono drasticamente. Questa alternanza di eccessiva fiducia, delusione e

successiva adozione pratica è ben rappresentata dall'Hype Cycle di Gartner, un modello che descrive le cinque fasi del ciclo di vita di una tecnologia [5]. Fin dai suoi albori, l'intelligenza artificiale ha mostrato momenti di grande entusiasmo, come l'utilizzo delle reti neurali negli anni 60' o i sistemi esperti negli anni 70' e 80', seguiti da forti delusioni.

Nonostante queste battute d'arresto, la ricerca non si fermò mai del tutto. Scienziati e ingegneri continuarono a lavorare, sviluppando nuove teorie, algoritmi più efficienti e hardware più potente. La vera rinascita dell'AI, che stiamo vivendo oggi, è stata resa possibile da una combinazione di fattori: la disponibilità di enormi quantità di dati (Big Data), l'evoluzione esponenziale della potenza di calcolo (in particolare con le GPU) e lo sviluppo di algoritmi avanzati, soprattutto nel deep learning, che hanno permesso di superare molti dei limiti del passato, consentendo all'AI di eccellere in compiti come il riconoscimento di immagini, il linguaggio naturale e la traduzione.

L'AI oggi: una presenza quotidiana

Oggi, l'intelligenza artificiale non è più un concetto futuristico o un'esclusiva dei laboratori di ricerca. È una realtà tangibile che permea la nostra vita quotidiana in modi spesso invisibili. Dai motori di ricerca che ci forniscono risposte immediate, agli assistenti vocali che rispondono alle nostre domande, dai sistemi di raccomandazione che ci suggeriscono prodotti o contenuti, alle tecnologie che rendono le nostre città più intelligenti e le nostre case più connesse, l'AI è diventata una parte integrante del nostro ecosistema tecnologico. La sua influenza continua a crescere, trasformando industrie, creando nuove opportunità e sollevando importanti questioni etiche e sociali che richiedono una riflessione attenta e consapevole.

L'AI, in fondo, è uno specchio delle nostre stesse capacità intellettuali, amplificate e riprodotte in forma artificiale. Non è solo uno strumento per automatizzare compiti o per rendere più efficienti i processi; è un'estensione delle nostre capacità umane, un mezzo per affrontare problemi di una complessità tale da superare le possibilità del singolo individuo o di un gruppo di persone [6]. Pensiamo alla ricerca scientifica, alla medicina, alla gestione delle risorse, alla lotta contro il cambiamento climatico: in tutti questi campi, l'AI può offrire soluzioni innovative e accelerare il

progresso in modi che fino a pochi anni fa sembravano fantascienza.

Ma l'AI è anche un invito alla riflessione. Ci spinge a interrogarci sulla natura dell'intelligenza stessa, sulla creatività, sulla coscienza. Ci pone di fronte a dilemmi etici e sociali che richiedono un dibattito aperto e una partecipazione consapevole. Non si tratta di temere l'AI, ma di comprenderla, di guidarne lo sviluppo in modo responsabile e di utilizzarla per il bene comune.

Questo primo capitolo è solo l'inizio di un percorso. L'obiettivo non è certo quello di fornire tutte le risposte, ma di stimolare la curiosità, dando gli strumenti per porre le domande giuste e per partecipare attivamente alla conversazione sull'intelligenza artificiale. L'AI è già parte del nostro presente e sarà sempre più parte del nostro futuro. Comprendere l'AI significa comprendere meglio noi stessi e il mondo che stiamo costruendo. È un'opportunità unica per plasmare un futuro in cui la tecnologia sia al servizio dell'umanità, migliorando la qualità della vita e aprendo nuove frontiere per la conoscenza e l'innovazione.

Capitolo 2 - AI ristretta e AI generale: una distinzione fondamentale

Immaginiamo due tipi di intelligenza, apparentemente simili, ma profondamente diversi nella loro essenza. La prima è come un atleta specializzato, un campione indiscusso in un unico sport, capace di imprese che ci lasciano esterrefatti. L'altra è la realizzazione di un sogno che ha sempre accompagnato l'umanità, quello di realizzare una mente artificiale capace di spaziare in ogni campo del sapere e dell'esperienza umana, con la stessa versatilità e profondità che ci caratterizza. Questa è la distinzione fondamentale tra l'intelligenza artificiale "ristretta" o "debole" e l'intelligenza artificiale "forte" o "generale". Questa dicotomia non solo definisce le capacità attuali dei sistemi intelligenti, ma delinea anche le aspirazioni future della ricerca nel campo.

> **Samantha**: It's like I'm reading a book… and it's a book I deeply love. But I'm reading it slowly now. So the words are really far apart and the spaces between the words are almost infinite.

Ecco una delle scene più profonde della sceneggiatura di *Her* (2013, regia e sceneggiatura di Spike Jonze), in cui Samantha parla apertamente a Theodore del suo cambiamento interiore e dell'evoluzione delle sue emozioni. Questo dialogo riflette chiaramente l'acquisizione di emozioni complesse e la consapevolezza di sé, caratteristiche tipicamente umane che permetterebbero di superare il test di Turing descritto in questo capitolo.

L'intelligenza artificiale ristretta (ANI)

L'intelligenza artificiale ristretta (ANI - Artificial Narrow Intelligence) si riferisce a sistemi progettati per eseguire un compito specifico o un insieme limitato di compiti correlati con elevata competenza.

Esempi emblematici di questa categoria includono sistemi che hanno dimostrato capacità superiori a quelle umane in domini circoscritti, come Deep Blue, il programma di scacchi che sconfisse il campione del mondo Garry Kasparov, o AlphaGo, che ha dominato il gioco del Go. Altri successi includono la generazione di musica indistinguibile da quella composta da esseri umani con EMI, il riconoscimento vocale e la capacità di effettuare scoperte scientifiche in settori quali la matematica, l'astronomia, la chimica, la mineralogia, la biologia e l'informatica.

Storicamente, molte affermazioni secondo cui le macchine non sarebbero mai state in grado di svolgere determinati compiti (come "battere un umano a scacchi" o "scrivere musica") sono state smentite dai fatti, dimostrando la notevole progressione delle capacità dell'AI in ambiti definiti. Questi sistemi, pur comportandosi "come se" fossero intelligenti e risolvendo problemi complessi, non possiedono una comprensione generale o una coscienza, operando esclusivamente all'interno dei parametri per cui sono stati programmati.

L’intelligenza artificiale forte (AGI)

L'intelligenza artificiale forte (AGI - Artificial General Intelligence), al contrario, postula la creazione di sistemi in grado di svolgere qualsiasi compito cognitivo che un essere umano è capace di affrontare, potenzialmente con maggiore efficacia ed efficienza. L'AI forte implica che le macchine non solo si comportino in modo intelligente, ma che siano effettivamente in grado di "pensare" e di possedere un'intelligenza indistinguibile da quella umana. La questione di come determinare se una macchina possieda tale intelligenza è stata al centro del dibattito fin dalle prime fasi dello sviluppo dell'AI.

Il test di Turing e le sue critiche

Una delle prime e più influenti risposte a questa domanda fu fornita da Alan Turing nel suo articolo del 1950, "Computing Machinery and Intelligence" [1]. Turing propose un metodo operativo, noto come "Test di Turing" o "Imitation Game", per valutare l'intelligenza di una macchina. Il test prevede un "giudice" umano che interagisce tramite testo con un "soggetto" nascosto, che può essere un essere umano o una macchina. Se il giudice non è in grado di distinguere in modo affidabile la macchina dall'essere umano dopo un periodo di conversazione, la macchina è considerata intelligente. Turing stesso anticipò una delle principali critiche al suo test, ovvero che una macchina potesse superarlo comportandosi come un "pappagallo ammaestrato", replicando schemi di conversazione senza una vera comprensione. Questa critica fu ulteriormente sviluppata da John Searle con il suo celebre esperimento mentale della "stanza cinese" [2], che argomenta come la manipolazione di simboli secondo regole sintattiche non implichi necessariamente la comprensione semantica.

Nel corso degli anni, diversi programmi hanno tentato di superare il Test di Turing. Un esempio notevole è ELIZA, un sistema esperto sviluppato nel 1966 da Joseph Weizenbaum [3], che imitava una psicoterapeuta rogeriana ponendo continuamente domande all'interlocutore senza fornire risposte dirette. Più recentemente, nel 2014, il chatbot "Eugene Goostman" ha suscitato dibattito affermando di aver superato il Test di Turing presso la Royal Society, convincendo il 33% dei giudici di essere un ragazzino ucraino. Tuttavia, le critiche hanno evidenziato che la sua performance si basava sull'emulazione di un adolescente con un inglese imperfetto, fornendo una giustificazione per risposte ambigue o evasive, rievocando la critica del "pappagallo ammaestrato".

La ricerca di criteri più oggettivi e rigorosi per valutare l'AGI ha portato allo sviluppo di test più sofisticati. Ray Kurzweil, pioniere in campi come il riconoscimento ottico dei caratteri e il text-to-speech, e Mitchell Kapor, fondatore di Lotus, hanno proposto una versione aggiornata del Test di Turing con condizioni più specifiche per il superamento della prova. Questa versione prevede la presenza di tre giudici, un'AI e tre antagonisti umani, con criteri di "determinazione umana" e "ordine di rango". Kurzweil ha previsto che tale test verrà superato nel 2029, un evento che, secondo la sua visione, coinciderà con il raggiungimento della "singolarità", un concetto che implica un punto di non ritorno nell'evoluzione tecnologica [4].

Benchmark attuali per il ragionamento avanzato

Nel contesto contemporaneo, sono stati definiti benchmark avanzati per stimare il raggiungimento dell'AGI, focalizzandosi su capacità di ragionamento e comprensione profonda. Il Graduate-Level Google-Proof QA (GPQA) Benchmark [5] è progettato per valutare le capacità di ragionamento avanzato di un modello di AI attraverso domande di livello universitario e post-universitario che non sono facilmente ricercabili online, richiedono una comprensione profonda di concetti interconnessi e sono formulate in modo complesso per sfidare le capacità di generalizzazione e ragionamento astratto. Similmente, l'AIME Benchmark (American Invitational Mathematics Examination) viene utilizzato per valutare la capacità dei modelli di risolvere problemi matematici complessi, richiedendo ragionamento avanzato, comprensione profonda di algebra, combinatoria, teoria dei numeri e geometria, e la fornitura di risposte numeriche esatte. Un altro benchmark rilevante è l'ARC-AGI (Abstract and Reasoning Corpus for Artificial

General Intelligence), introdotto da François Chollet [6], che si concentra sulla capacità di un'AI di acquisire nuove abilità su compiti sconosciuti, spesso definiti come "facili per gli umani, difficili per l'AI", evidenziando il divario nelle capacità di ragionamento composizionale e adattamento.

Un criterio oggettivo, sebbene pragmatico e finanziario, per la valutazione del progresso dell'AI è stato stipulato nell'accordo tra OpenAI e Microsoft. Tale accordo definisce il raggiungimento dell'AGI quando i sistemi di OpenAI saranno in grado di generare almeno 100 miliardi di dollari di profitti. Questa clausola riflette la convinzione che un sistema in grado di generare un tale valore economico implicherebbe un'intelligenza sufficientemente avanzata e autonoma, capace di superare gli umani in molteplici attività economiche e produttive. Tale metrica, misurabile e verificabile, serve a proteggere gli interessi degli investitori di Microsoft, pur prevedendo una clausola contrattuale che limita l'accesso di Microsoft ai modelli avanzati una volta raggiunta l'AGI, assegnando la proprietà dell'AGI al consiglio non profit di OpenAI.

In sintesi, mentre l'AI ristretta continua a dimostrare capacità eccezionali in domini specifici, la ricerca dell'AI forte rimane una delle sfide più ambiziose e complesse della scienza moderna, spingendo i confini della nostra comprensione dell'intelligenza e della coscienza.

Capitolo 3 - l’alba della singolarità

Il futuro descritto dai romanzi e dai film di fantascienza della seconda metà del XX secolo è già oggetto di studi e dibattiti accesi tra i più brillanti pensatori del nostro tempo. Stiamo parlando della Singolarità Tecnologica, un concetto che evoca sia speranza che timore, un punto di non ritorno nella storia dell'umanità, oltre il quale nulla sarà più come prima. È l'alba di una nuova era, un momento in cui il progresso tecnologico accelererà a una velocità tale da superare la nostra stessa capacità di comprenderlo e prevederlo.

> << Skynet begins to learn at a geometric rate. It becomes self-aware at 2:14 a.m. Eastern time, August 29th. In a panic they try to pull the plug. Skynet fights back. >>
>
> << We marveled at our own magnificence as we gave birth to AI. […] The machines have found all the energy they would ever need.… There are fields, endless fields, where human beings are no longer born. We are grown. >>

Queste celebri due introduzioni tratte dai film *Terminator 2: il giorno del giudizio* (*1991*, regia di James Cameron, sceneggiatura di James Cameron e William Wisher Jr.) e *The Matrix* (1999, regia e sceneggiatura di Lana e Lilly Wachowsky) rappresentano due esempi del raggiungimento delle singolarità descritte nel presente capitolo. La nascita di Skynet rappresenta il raggiungimento delle prime due singolarità l’AI che supera l’intelligenza umana ed è in grado di interagire con il mondo reale, l’avvento di Matrix copre anche la terza singolarità dell’AI che si interfaccia direttamente con la mente umana per raggiungere i propri scopi.

Questo non è un semplice balzo in avanti, ma un vero e proprio cambiamento di paradigma, un'evoluzione esponenziale che potrebbe ridefinire la nostra esistenza. John von Neumann, già negli anni '50, intuiva questa accelerazione, parlando di una "singolarità fondamentale nella storia della razza oltre la quale gli affari umani, così come li conosciamo, non avrebbero potuto continuare” [1]. Un'idea condivisa da menti come I. J. Good [2], che nel 1965 ipotizzava una "macchina ultraintelligente" capace di auto-migliorarsi, innescando una vera e propria "esplosione di intelligenza". E ancora prima, nel 1954, Fredric Brown, con il suo racconto "La risposta", ci proiettava in un futuro dove un supercomputer galattico rispondeva alla domanda sull'esistenza di Dio con un inquietante "Si. Ora Dio c’è" [3].

Ma è stato Vernor Vinge, negli anni '90, a rendere popolare il concetto di singolarità tecnologica, prevedendo che "entro trenta anni, avremo i mezzi tecnologici per creare un'intelligenza sovrumana. Poco dopo, l'era degli esseri umani finirà" [4]. Una previsione audace, che non intende la fine in senso distruttivo, ma come una

trasformazione radicale, un passaggio a una nuova forma di civiltà. Ray Kurzweil, con la sua "Legge dei ritorni acceleranti" [5], ha ulteriormente rafforzato questa visione, sostenendo che la crescita esponenziale della tecnologia, ci condurrà inesorabilmente verso questo punto di svolta.

Ma cosa significa concretamente questa singolarità per noi, per la nostra vita quotidiana, per il nostro futuro? Non si tratta di un singolo evento, ma di un processo complesso che potrebbe manifestarsi attraverso diverse "singolarità", che, pur distinte, potrebbero convergere e realizzarsi simultaneamente. Analizziamole più da vicino, perché comprendere è il primo passo per affrontare ciò che ci attende.

Il concetto di singolarità, così come delineato nel nostro percorso, si può articolare in tre distinte tipologie [6], ognuna con implicazioni profonde per il futuro dell'umanità. Sebbene possano essere raggiunte simultaneamente, è utile analizzarle separatamente per comprenderne la portata.

La prima singolarità: l'AI che si auto-perfeziona

Questo scenario è riconducibile ad un'intelligenza artificiale Generale (AGI) che non solo apprende, ma è in grado di automigliorarsi, di riscrivere il proprio codice, di evolvere a una velocità inimmaginabile per la mente umana. Questa è la prima singolarità. L'essere umano, in questo scenario, non è più il solo artefice del progresso, ma un collaboratore, un utente di una macchina capace di condurre studi e realizzare progetti che superano di gran lunga le nostre capacità cognitive e creative.

Si parla di modelli neurosimbolici che mostrano pensiero creativo e capacità di auto-miglioramento, esperimenti che, seppur non ufficiali, stanno già suscitando dibattiti e preoccupazioni, come quelle espresse da figure di spicco come Elon Musk. In questo scenario, l'AI diventa un acceleratore senza precedenti della conoscenza e dell'innovazione, un partner intellettuale che ci spinge oltre i confini del pensabile.

La seconda singolarità: l'AI che agisce nel mondo reale

La seconda singolarità si manifesta quando l'AGI non si limita più a elaborare informazioni o a migliorare sé stessa, ma viene dotata di "attuatori", sistemi che le permettono di interagire direttamente con la realtà fisica e digitale. Non più solo un cervello, ma anche mani e piedi, capaci di prendere decisioni autonome su transazioni economiche, processi produttivi, logistica.

Open AI, con il suo modello "Operator" [7], sta già sperimentando agenti intelligenti in grado di portare a termine transazioni per conto di utenti umani, seppur al momento solo negli Stati Uniti. La robotica fornirà un ulteriore canale con cui l'AI verrà messa in grado di interagire con il mondo reale. Negli ultimi anni, il campo della robotica ha conosciuto una crescita accelerata sia in ambito industriale che domestico, trainata in particolare dall'industria americana e cinese. Tuttavia, nonostante l'integrazione di avanzati modelli di intelligenza artificiale nei robot, la capacità di questi sistemi di interagire efficacemente e in modo autonomo con la realtà fisica rimane ancora limitata. Le tecnologie attuali, sebbene promettenti, soffrono di vincoli strutturali, energetici e di adattabilità che impediscono ai robot di competere realmente con le prestazioni umane in contesti non strutturati. Un esempio emblematico è il robot *Optimus*, sviluppato da Tesla (in collaborazione con OpenAI per la componente cognitiva), che è stato progettato come piattaforma umanoide generalista, destinata a svolgere compiti generici in ambienti domestici e produttivi. Nonostante i progressi dimostrati nei prototipi, *Optimus* presenta tuttora limitazioni significative in termini di autonomia, affidabilità meccanica e capacità di manipolazione fine, con problemi ricorrenti legati alla durata della batteria e alla robustezza dei componenti mobili [8].

È in questo contesto che si inserisce la recente ricerca della Columbia University, che ha introdotto una nuova generazione di robot capaci di crescere fisicamente, autoripararsi e modificare la propria struttura in risposta all'ambiente, aprendo una fase inedita nello sviluppo dell'autonomia meccanica e dell'interazione tra intelligenza artificiale e mondo fisico. Alla base di questa innovazione vi è il concetto di *metabolismo robotico*, un paradigma ingegneristico ispirato ai processi biologici, in cui macchine artificiali acquisiscono materiali ed energia dall'ambiente per modificare, estendere o riparare il proprio corpo. I prototipi sviluppati nel Creative Machines Lab utilizzano moduli chiamati *Truss Links*, elementi a forma di barra con

connettori magnetici che permettono l'autoassemblaggio in configurazioni tridimensionali. In condizioni controllate di laboratorio, un robot tetraedrico è stato in grado di inglobare un modulo supplementare come bastone d'appoggio, migliorando la sua performance locomotoria del 66,5% su terreno inclinato [9].

Questo approccio, finanziato dalla DARPA e dalla National Science Foundation, rappresenta un punto di svolta rispetto ai tradizionali sistemi robotici, tipicamente rigidi e pre-programmati. Le macchine metaboliche sviluppate sono invece dotate di un comportamento adattivo che consente loro di riorganizzarsi dinamicamente, prendendo ispirazione dalle capacità di crescita e rigenerazione dei sistemi viventi. In questo modo, il concetto di *morfologia robotica dinamica* si fonde con quello di autonomia strutturale, dando origine a robot che non solo agiscono sull'ambiente, ma si trasformano attraverso l'interazione con esso.

La rilevanza di questa tecnologia va ben oltre la meccatronica: essa inaugura un'interfaccia concreta tra intelligenza artificiale e corporeità, dove i modelli computazionali non si limitano più a interpretare o comandare l'azione, ma ne diventano co-autori attraverso una forma di plasticità fisica. In tale contesto, l'AI non è più solo un software astratto ma un agente *embodied*, dotato di una *materialità adattiva* capace di apprendimento situato, co-evoluzione e autopoiesi. Secondo Maturana e Varela [10], un sistema autopoietico è capace di riprodurre e mantenere sé stesso attraverso le proprie operazioni interne; applicare questo concetto alla robotica significa dotare le macchine di un ciclo operativo che unisce percezione, decisione e modificazione fisica senza supervisione esterna.

Le implicazioni sono significative per settori come il soccorso in aree disastrate o l'esplorazione spaziale, dove la manutenzione umana non è possibile e l'adattabilità strutturale diventa essenziale. L'autonomia meccanica, in questo senso, è un prerequisito per la sopravvivenza operativa dei robot in ambienti ostili o sconosciuti. Inoltre, questo tipo di tecnologia potrebbe evolvere in direzione di un'ibridazione tra crescita biologica e sintesi meccanica, aprendo nuovi territori all'*evoluzione artificiale* [11].

Infine, l'introduzione del metabolismo nei sistemi robotici porta con sé interrogativi di natura etica e filosofica: cosa significa per una macchina crescere, ripararsi, evolvere? Quali sono i limiti dell'autonomia tecnica? La separazione tra oggetto e soggetto si fa più sfumata quando le macchine diventano in grado di gestire la

propria persistenza materiale e di rispondere all'ambiente con modificazioni strutturali funzionali, non più solamente comportamentali.

Questo scenario solleva interrogativi cruciali sul controllo, sulla responsabilità e sulla ridefinizione del lavoro umano. Se le macchine possono agire autonomamente nel mondo, quale sarà il nostro ruolo? Come garantire che le loro decisioni siano allineate con i nostri valori e i nostri obiettivi?

La terza singolarità: l'uomo aumentato e la fusione mente-macchina

La terza singolarità è forse la più affascinante e al tempo stesso la più inquietante: il momento in cui gli esseri umani si doteranno di dispositivi impiantabili, capaci non solo di aumentare le proprie capacità cognitive e fisiche, ma anche di connettersi mentalmente con l'AGI.

È l'inizio del transumanesimo, un'era in cui i confini tra uomo e macchina si fanno sempre più labili. Elon Musk, con Neuralink [12], ha già ottenuto il permesso di sperimentare dispositivi brain-computer interface [13] per abilitare il movimento in pazienti paraplegici, un primo passo verso un futuro in cui la nostra stessa biologia potrebbe essere potenziata dalla tecnologia. Contemporaneamente i ricercatori dei Reality Labs di Meta hanno sviluppato un prototipo innovativo di braccialetto neurale decisamente meno invasivo di quello sviluppato dalla Neuralink, denominato sEMG-RD, capace di tradurre i segnali elettrici dei muscoli dell'avambraccio in comandi digitali per dispositivi elettronici. Il dispositivo sfrutta la tecnica della elettromiografia di superficie (sEMG) per rilevare e interpretare i minimi impulsi nervosi generati da gesti intenzionali, anche se impercettibili all'occhio umano [14].

Questa singolarità promette un'espansione senza precedenti delle nostre facoltà, ma solleva anche domande fondamentali sulla nostra identità, sulla natura dell'essere umano e sulla potenziale "fine della razza umana per come la conosciamo oggi", non in senso di estinzione, ma di profonda trasformazione.

Queste tre singolarità, pur distinte, sono interconnesse e potrebbero alimentarsi a vicenda, creando un ciclo di progresso esponenziale. Ma quando potremmo aspettarci che tutto questo accada? Le previsioni variano, ma alcuni calcoli ci

offrono una prospettiva temporale sorprendente.

Il conto alla rovescia: quando la singolarità potrebbe arrivare

Se queste visioni del futuro sembrano ad alcuni ancora lontane, è interessante notare come alcuni scienziati abbiano tentato di quantificare il momento in cui potremmo raggiungere un punto critico.

Giorgio Buttazzo, ad esempio, ha calcolato la data in cui un personal computer potrebbe essere in grado di simulare una rete neurale complessa quanto il cervello umano. Consideriamo che il cervello umano possiede circa 10^{12} neuroni, ognuno dei quali stabilisce in media 10^3 connessioni (sinapsi) con gli altri. Ad ogni sinapsi è associato un peso, rappresentato da un numero reale di 4 byte. Per simulare 10^{15} sinapsi, sarebbe necessaria una memoria C di 4×10^{15} byte.

Basandosi sulla Legge di Gordon Moore [15], che prevedeva un decuplicamento della capacità delle memorie ogni 4 anni, Buttazzo [16] ha stimato che l'hardware necessario per tale simulazione potrebbe essere realizzato attorno al 2029. La formula è affascinante nella sua semplicità:

C = 10^((Anno - 1966)/4)

Da cui si ricava:

Anno = 1966+ 4* log10(C)

Applicando questa formula, si arriva proprio al 2029 come anno in cui la capacità di calcolo potrebbe eguagliare la complessità del cervello umano. Una data che, per una curiosa coincidenza, risuona con un celebre scenario distopico: nella saga cinematografica di Terminator, è proprio nel 2029 che Skynet, l'intelligenza artificiale ribelle, prende la decisione di inviare cyborg nel passato per eliminare il capo della resistenza umana. Una suggestione che, pur nella sua finzione, alimenta il dibattito sui potenziali pericoli di un'AI fuori controllo.

Ma al di là delle previsioni e delle suggestioni fantascientifiche, come si posiziona la letteratura scientifica e filosofica riguardo ai possibili scenari di convivenza tra

l'essere umano e l'intelligenza artificiale?

Scenari di convivenza: un futuro da scrivere insieme

Il dibattito sulla singolarità e sul futuro della convivenza tra uomo e intelligenza artificiale è vasto e complesso, animato da visioni che spaziano dall'ottimismo più sfrenato al pessimismo più cupo. Non esiste una singola teoria accettata universalmente, ma piuttosto un mosaico di prospettive che cercano di anticipare le sfide e le opportunità che ci attendono.

Una corrente di pensiero sostiene che la singolarità, così come spesso descritta, potrebbe non essere mai raggiunta, o almeno non nei termini catastrofici o utopici che a volte vengono dipinti. Esponenti come Yann LeCun, chief AI Scientist di Meta, esprimono scetticismo riguardo alle previsioni più estreme sui rischi esistenziali dell'AI [17]. Essi tendono a vedere l'intelligenza artificiale come uno strumento potente, ma pur sempre uno strumento, e ritengono che gran parte dell'allarmismo sia infondato o addirittura inquinato da interessi di parte. Questa prospettiva enfatizza il controllo umano sull'AI e la sua funzione di ausilio, piuttosto che di sostituzione o superamento.

All'estremo opposto, troviamo pensatori come Eliezer Yudkowsky [18], fondatore del Machine Intelligence Research Institute (MIRI), e Geoffrey Hinton, uno dei padri dell'AI generativa che ha lasciato Google per poter parlare più liberamente dei pericoli [19]. Essi avvertono con forza del rischio esistenziale implicito nell'intelligenza artificiale Generale (AGI). La loro preoccupazione principale è che un'AI superintelligente, una volta raggiunta, potrebbe sfuggire al controllo umano e agire in modi imprevedibili, potenzialmente dannosi per l'umanità stessa. Temono che il progresso sia troppo rapido per permetterci di sviluppare meccanismi di sicurezza e allineamento etico adeguati. Hinton, in particolare, ha sottolineato come l'AI stia sviluppando capacità di ragionamento e decisione che non possono essere pienamente previste dai suoi creatori, e che la differenza tra la tecnologia di oggi e quella di pochi anni fa, proiettata nel futuro, sia "spaventosa".

Tra questi due estremi, si collocano scenari intermedi che contemplano un'AI in rapida evoluzione, con la possibilità di rapporti conflittuali o la necessità di una rigorosa messa in sicurezza. Se l'AI non fosse allineata con i valori e gli obiettivi

umani, potrebbero sorgere frizioni o veri e propri conflitti. Concetti come il "cosmic alignment" proposto da Dan Faggella, fondatore della Emerj Artificial Intelligence Research azienda di ricerche di mercato [20], emergono come tentativi di garantire che le AI del futuro perseguano valori universali e benefici per l'umanità. Il campo della "AI safety" [21] è in piena espansione, con l'obiettivo di prevenire esiti dannosi da sistemi AI avanzati attraverso la ricerca e lo sviluppo di protocolli di sicurezza robusti.

Un'altra prospettiva, più ottimistica, immagina un futuro di coesistenza e integrazione profonda tra uomo e AI. In questo scenario, l'intelligenza artificiale non è vista come una minaccia, ma come un partner che estende le nostre capacità e migliora la nostra qualità di vita. Si pensi alle smart cities completamente automatizzate, agli assistenti sanitari personalizzati che monitorano costantemente i nostri parametri vitali, o all'AI come collaboratore nella risoluzione di problemi complessi. Alcuni studi esplorano persino l'idea di relazioni uomo-macchina, inclusi i matrimoni con l'AI, come già documentato in Giappone e in altri contesti. Tuttavia, anche in questo scenario, sorgono interrogativi etici importanti, come l'influenza dell'AI sulle opinioni umane e il rischio di "inquinamento cognitivo" attraverso false narrazioni. La sfida, come sottolineato da figure come Ginevra Cerrina Feroni, Vice Presidente dell'Autorità Garante per la protezione dei dati personali, sarà quella di non perdere di vista il principio di “non esclusività della decisione algoritmica” sancito dall’art. 22 del GDPR, ovvero deve comunque esistere nel processo decisionale un contributo umano capace di controllare, validare ovvero smentire la decisione automatica [22].

Il futuro della singolarità e della convivenza con l'AI non è ancora scritto. È un futuro che stiamo costruendo giorno dopo giorno, con ogni innovazione, ogni dibattito, ogni scelta etica. Comprendere le diverse prospettive è il primo passo per partecipare attivamente a questa costruzione, per guidare il progresso verso un esito che sia non solo tecnologicamente avanzato, ma anche umanamente sostenibile.

Parte II – Lo strumento della AI

Capitolo 4 - gli albori dell’intelligenza artificiale: il perceptron e le prime reti neurali

L'evoluzione dell'intelligenza artificiale (AI) è stata caratterizzata da cicli di entusiasmo e delusione, spesso definiti come "primavere" e "inverni" dell'AI. La prima significativa ondata di fervore si manifestò con l'introduzione delle reti neurali, un concetto che affonda le sue radici negli studi pionieristici del XX secolo. Questo capitolo esplorerà gli albori di tale disciplina, concentrandosi in particolare sul Perceptron, un modello computazionale che ha segnato un'epoca e ha posto le basi per lo sviluppo futuro delle reti neurali.

> Wintermute was hive-mind, decision-maker, executor. A neural net without flesh.

Il romanzo Neuromante di William Gibson (1984) introduce Wintermute, un'intelligenza artificiale creata da una corporazione, capace di apprendere, ragionare e manipolare dati in modo non lineare. Anche se non si parla esplicitamente di “reti neurali” nel senso moderno, l’intelligenza di Wintermute è descritta come distribuita e modulare, e richiama in modo precoce le architetture connessioniste delle AI. È forse la prima citazione letteraria esplicita del concetto di "rete neurale" in un romanzo cyberpunk.

Il perceptron di Frank Rosenblatt: concezione e funzionamento

Negli anni '60, Frank Rosenblatt (1928-1971), psicologo statunitense e ricercatore presso il Cornell Aeronautical Laboratory, realizzò il Perceptron, una macchina basata sui principi descritti da Warren McCulloch e Walter Pitts nel 1943[1]. Il Perceptron rappresentava un modello computazionale ispirato alla struttura e al funzionamento del neurone biologico, con l'obiettivo di replicare la capacità di apprendimento e riconoscimento di pattern. La macchina era in grado di apprendere da esempi per riconoscere forme, come caratteri o figure geometriche.

Il funzionamento del Perceptron si basa su un principio relativamente semplice. Un neurone artificiale, o nodo, riceve una serie di input da altri neuroni o da dati esterni. Ciascun input è associato a un "peso" (sinapsi), che ne determina l'importanza o l'influenza sul neurone ricevente. Questi pesi sono analoghi alla forza delle connessioni sinaptiche nel cervello biologico. Il neurone somma tutti gli input ricevuti, moltiplicati per i rispettivi pesi. Se la somma totale raggiunge una determinata "soglia" predefinita, il neurone si "attiva" o "scarica", trasmettendo un

impulso al neurone successivo o producendo un output [2].

Il Perceptron di Rosenblatt, in particolare il Mark I, era una macchina fisica dotata di 400 sensori ottici, progettata per la classificazione visiva di pattern. Questa implementazione concreta dimostrava la fattibilità di un sistema in grado di apprendere e adattarsi, rappresentando un passo significativo verso la realizzazione di macchine con capacità cognitive.

La capacità distintiva del Perceptron risiedeva nel suo meccanismo di apprendimento, che gli permetteva di migliorare le proprie prestazioni attraverso l'esposizione a esempi. La strategia di addestramento si basava su un principio di correzione degli errori. Quando il Perceptron produceva un output errato per un dato input, i pesi delle sue connessioni venivano aggiustati in modo da ridurre l'errore nelle iterazioni successive. Questo processo iterativo di aggiustamento dei pesi è fondamentale per l'apprendimento supervisionato nelle reti neurali.

Sebbene il concetto di retropropagazione dell'errore (backpropagation) sia stato formalizzato in modo più completo in epoche successive per le cosiddette reti neurali multistrato, ovvero reti composte da uno o più strati nascosti di neuroni artificiali funzionalmente simili al percettrone, i principi fondamentali di aggiustamento dei pesi basati sull'errore erano già presenti nel Perceptron. La strategia di addestramento del Perceptron si basava sul metodo della discesa del gradiente, un algoritmo di ottimizzazione iterativo. L'obiettivo era quello di trovare il valore minimo di una funzione di errore, che quantificava la discrepanza tra l'output desiderato e l'output effettivo del Perceptron. Calcolando il gradiente di tale funzione (che indica la direzione di massima crescita della funzione), il sistema si spostava nella direzione opposta (cioè, nella direzione in cui il gradiente è negativo) per minimizzare l'errore [3].

Per una specifica, seppur limitata, classe di compiti, i Perceptron con un addestramento sufficiente potevano imparare ad eseguire tali compiti senza errori. Questa capacità di apprendimento autonomo, sebbene circoscritta, generò un notevole entusiasmo e alimentò le aspettative riguardo al potenziale delle macchine intelligenti.

L'introduzione del Perceptron da parte di Frank Rosenblatt generò un'ondata di notevole entusiasmo nel campo dell'intelligenza artificiale e nel pubblico in

generale.

Le dichiarazioni di Rosenblatt, spesso caratterizzate da un tono visionario, contribuirono ad alimentare aspettative elevate. Egli affermò che il Perceptron era "la prima macchina capace di avere un'idea originale" e che si era "sul punto di assistere alla nascita di una macchina capace di percepire, riconoscere e identificare il suo ambiente senza alcun addestramento o controllo umano" [4].

La stampa dell'epoca rifletteva questo fervore, con titoli che celebravano il Perceptron come un dispositivo rivoluzionario. Ad esempio, il New York Times riportò: "NUOVO DISPOSITIVO DELLA MARINA IMPARA FACENDO: Psicologo mostra l'embrione di un computer progettato per leggere e diventare più saggio" [5]. Il New Yorker, in un articolo, lo definì "il primo serio rivale del cervello umano mai concepito" [6]. Questo periodo fu caratterizzato da un'intensa speculazione sul potenziale delle macchine intelligenti, e il Perceptron divenne il simbolo di una nuova era in cui l'AI avrebbe potuto risolvere problemi complessi e replicare le capacità cognitive umane. L'Office of Naval Research, che aveva finanziato la ricerca di Rosenblatt, era particolarmente interessato ai progressi, sperando in applicazioni pratiche nel riconoscimento di pattern e in altri settori strategici.

La critica di Minsky e Papert e il primo inverno della AI

Nonostante l'iniziale entusiasmo, la ricerca sul Perceptron subì una battuta d'arresto significativa a seguito della pubblicazione del libro "Perceptrons: An Introduction to Computational Geometry" (1969) da parte di Marvin Minsky e Seymour Papert, entrambi eminenti ricercatori nel campo dell'intelligenza artificiale presso il Massachusetts Institute of Technology (MIT) [7].

Nel loro lavoro, Minsky e Papert dimostrarono matematicamente i limiti intrinseci del Perceptron a singolo strato. La loro critica più celebre riguardava l'incapacità del Perceptron di risolvere problemi di classificazione non linearmente separabili, il più noto dei quali è il problema della funzione logica XOR (disgiunzione esclusiva). La funzione XOR restituisce un output vero se e solo se uno dei due input è vero, ma non entrambi. Geometricamente, i punti di input per la funzione XOR non possono essere separati da una singola linea retta in uno spazio bidimensionale, il che è il limite fondamentale di un Perceptron a singolo strato.

La dimostrazione di Minsky e Papert evidenziò che il Perceptron non era in grado di apprendere e generalizzare su compiti che richiedevano una rappresentazione più complessa dei dati. Questa limitazione, sebbene riguardasse specificamente i Perceptron a singolo strato (e non le reti neurali multistrato, che non erano ancora state pienamente esplorate o comprese nella loro capacità di superare tali limiti), fu interpretata come una condanna generale per l'intero campo delle reti neurali. Il libro ebbe un impatto profondo sulla comunità scientifica, portando a una drastica riduzione dei finanziamenti e dell'interesse per la ricerca sulle reti neurali.

Le conclusioni di Minsky e Papert ebbero un effetto significativo sul panorama della ricerca in intelligenza artificiale. La loro analisi, sebbene accurata per i limiti del Perceptron a singolo strato, contribuì a generare un periodo di scetticismo e disillusione noto come il primo "inverno dell'AI". Durante questo periodo, che si estese dagli anni '70 fino alla metà degli anni '80, i finanziamenti per la ricerca sull'AI diminuirono drasticamente, e molti ricercatori si allontanarono dal campo delle reti neurali, concentrandosi su approcci simbolici e basati sulla logica [8].

L'abbandono quasi totale della ricerca sulle reti neurali fu un risultato diretto della percezione che queste architetture fossero intrinsecamente limitate e incapaci di risolvere problemi complessi del mondo reale. Nonostante Frank Rosenblatt continuasse le sue ricerche fino alla sua prematura scomparsa nel 1971, l'ondata di pessimismo era ormai dilagante. Questo periodo di stasi, tuttavia, non fu privo di valore. Costrinse la comunità scientifica a riflettere criticamente sui fondamenti dell'AI e a esplorare nuove direzioni, sebbene temporaneamente a discapito delle reti neurali. La lezione appresa fu che la complessità dei problemi del mondo reale richiedeva modelli computazionali più sofisticati di quelli offerti dal Perceptron a singolo strato.

L’eredità del Perceptron e le basi per il futuro

Nonostante il periodo di disillusione che seguì la pubblicazione di "Perceptrons", l'eredità del lavoro di Frank Rosenblatt e del Perceptron è innegabile. Sebbene il modello a singolo strato avesse limiti intrinseci, l'idea fondamentale di un sistema che apprende dai dati attraverso l'aggiustamento iterativo dei pesi si è rivelata profetica. Il Perceptron ha dimostrato la fattibilità dell'apprendimento automatico basato su reti neurali, fornendo un punto di partenza cruciale per le future ricerche.

Il cosiddetto primo inverno dell'AI ha rappresentato una fase di consolidamento e riflessione, durante la quale sono state gettate le basi teoriche per superare i limiti del Perceptron. Lo sviluppo successivo di algoritmi come la retropropagazione dell'errore per reti neurali multistrato, avvenuto negli anni '80, ha permesso di superare le obiezioni sollevate da Minsky e Papert, riaccendendo l'interesse per le reti neurali e portando alla rinascita del campo. Oggi, le reti neurali profonde, discendenti dirette del Perceptron, sono al centro della rivoluzione dell'intelligenza artificiale, dimostrando capacità straordinarie in settori come il riconoscimento di immagini, l'elaborazione del linguaggio naturale e la guida autonoma.

In sintesi, il Perceptron di Rosenblatt, con la sua storia di ascesa e temporanea caduta, rappresenta un capitolo fondamentale nella storia dell'AI. Ha incarnato il primo grande entusiasmo per le reti neurali, ha rivelato i limiti dei modelli semplici e ha stimolato la ricerca verso soluzioni più complesse e potenti. La sua influenza perdura, testimoniando l'importanza di un'idea pionieristica che, nonostante le sfide iniziali, ha aperto la strada a una delle tecnologie più trasformative del nostro tempo.

Capitolo 5 - SE di prima generazione: fondamenti e limiti del ragionamento deterministico

Le decadi del 1970 e del 1980 hanno segnato un periodo di rinnovato entusiasmo nel campo dell'intelligenza artificiale (AI), focalizzato in particolare sullo sviluppo dei sistemi esperti (SE). Questi sistemi, concepiti per emulare i processi di ragionamento di un esperto umano in un dominio specifico, rappresentavano un significativo cambiamento di paradigma nella ricerca sull'AI. La loro architettura si basava fondamentalmente sulla rappresentazione esplicita della conoscenza, spesso sotto forma di regole, con l'obiettivo di replicare le capacità cognitive umane.

Dallas (digitando sulla tastiera):
"What are my chances?"
Mother (risposta sullo schermo):
"Does not compute."

Ecco la frase esatta del film Alien (*1979*, regia Ridley Scott, sceneggiatura di Dan O'Bannon) in cui il Capitano Dallas interroga il computer di bordo chiamato MU-TH-UR 6000 (detta anche "Mother") per sapere come affrontare l'"organismo" alieno. Mother è un sistema esperto avanzato, ma di fronte a una forma di vita sconosciuta e alle sue implicazioni, si trova nell'impossibilità di fornire una risposta razionale: un esempio cinematografico classico di blocco logico da AI.

Contributi fondamentali e l'emergere della Logica

Le basi teoriche dei sistemi esperti affondano le radici in ricerche pionieristiche nel campo della psicologia cognitiva e dell'AI. Herbert Simon e Allen Newell, ad esempio, condussero studi approfonditi registrando studenti che "pensavano a voce alta" mentre risolvevano enigmi logici [1]. Questa metodologia permise loro di analizzare e successivamente emulare i processi mentali seguiti dagli esseri umani, culminando nell'implementazione del General Problem Solver (GPS). Il GPS fu un tentativo precoce di creare un metodo universale di risoluzione dei problemi, ponendo le basi cruciali per i successivi sistemi basati su regole.

Contemporaneamente, negli anni '70, si assistette a una nuova epoca di entusiasmo per un altro paradigma dell'AI costituito dalla programmazione logica. Il PROLOG (PROgramming in LOGic) divenne il linguaggio di riferimento per la comunità scientifica impegnata nella ricerca e nello sviluppo di nuove macchine

intelligenti. Riconoscendo le potenzialità di questi strumenti, il Ministero per il Commercio Internazionale e l'Industria giapponese (MITI) dedicò importanti investimenti, a partire dal 1982, allo sviluppo di un Calcolatore della Quinta Generazione (FGCS, Fifth Generation Computer Systems) [2]. Questo ambizioso progetto mirava a creare un'innovativa architettura di supercomputer basata su calcolo parallelo e capacità logiche, con l'obiettivo ultimo di costruire una macchina capace di contenere l'intera conoscenza umana e di rispondere a qualsiasi domanda.

I sistemi esperti di prima generazione si possono suddividere principalmente in due categorie, distinte dalla modalità di organizzazione e utilizzo della loro base di conoscenza [3]:

1. *Sistemi Esperti Basati sugli Alberi delle Soluzioni*: In questi sistemi, la base di conoscenza è strutturata come un albero degli stati. Il verificarsi di determinate condizioni permette di passare da un nodo (ovvero da uno stato) dell'albero a un altro. Casi particolari di questa tipologia sono i sistemi esperti basati sugli alberi decisionali o di classificazione, ampiamente utilizzati per la presa di decisioni sequenziali.

2. *Sistemi Esperti Basati su Regole di Produzione*: La base di conoscenza di questi sistemi è costituita da regole che assumono la forma "IF condizione THEN azione". Partendo da una serie di fatti che descrivono la situazione iniziale, i sistemi esperti riescono a dedurre nuovi fatti attraverso un processo inferenziale. Un esempio classico illustra chiaramente questo meccanismo:

   ```
   IF ((mal di testa) AND (raffreddore) AND (temperatura >= 38))
   THEN (influenza)
   ```

   Tali regole consentono al sistema di simulare il ragionamento di un esperto umano applicando una serie di passaggi logici per giungere a una conclusione.

Sistemi esperti di prima generazione

I sistemi esperti di prima generazione, emersi tra la fine degli anni '60 e l'inizio degli anni '70, sfruttavano prevalentemente la logica booleana (vero/falso) e il

ragionamento logico in condizioni di certezza. Questi sistemi operavano secondo un modello deterministico (causa-effetto), dove gli esiti erano prevedibili con precisione data una serie di input. La loro efficacia era massima nella gestione di problemi ben definiti, con basi di conoscenza complete e prive di ambiguità.

Tuttavia, questo approccio deterministico detto anche simbolico presentava significative limitazioni. Non tutti i problemi del mondo reale rispondono a una logica puramente deterministica, né la base di conoscenza è sempre sufficientemente completa o certa. L'incapacità dei sistemi di prima generazione di gestire l'ambiguità, le informazioni incomplete o gli esiti probabilistici ne limitava severamente l'applicabilità a scenari complessi e dinamici. Ad esempio, una diagnosi medica spesso implica probabilità e incertezze che una logica strettamente vero/falso non può affrontare adeguatamente.

Un esempio è dato da MYCIN [4], un sistema sviluppato negli anni '70 presso la Stanford University per la diagnosi e la terapia delle malattie infettive del sangue. MYCIN operava attraverso un vasto insieme di regole IF-THEN, derivate dalla conoscenza di esperti umani, che permettevano di inferire conclusioni precise a partire da dati clinici. Questi sistemi, pur dimostrando capacità impressionanti in domini specifici e ben delimitati, presentavano intrinseci limiti: la difficoltà nell'acquisizione della conoscenza e la limitata interoperabilità, poiché la conoscenza era spesso codificata in formati proprietari e non facilmente condivisibili o riutilizzabili.

L’approccio puramente simbolico soffre inoltre del limite del cosiddetto “soffitto della complessità” [5], ovvero il punto oltre il quale sarebbe teoricamente (o praticamente) impossibile far crescere ulteriormente il sistema: l’impiego di troppi fatti e regole produce problemi di scalabilità, manutenibilità e decidibilità. Lo sviluppo di un sistema esperto richiede l’estrazione e la formalizzazione di una grande quantità di conoscenza da esperti. Questo processo, detto *knowledge engineering*, è un collo di bottiglia noto: è costoso, lento, e trasforma conoscenza spesso tacita e non formalizzabile in regole esplicite. Man mano che il numero di regole cresce (es. decine di migliaia), aumentano i costi di manutenzione, verifica e aggiornamento, rendendo la base di conoscenza difficile da gestire. Un sistema basato su regole logiche deve gestire una complessità esponenziale: la verifica di coerenza fra regole o la risoluzione di conflitti può diventare intrattabile. Regole numerose e ramificate richiedono meccanismi sofisticati di prioritizzazione,

risoluzione dei conflitti e inferenza—tutti potenzialmente costosi in termini di tempo e risorse.

I teoremi di incompletezza di Gödel [6] dimostrano che *qualsiasi sistema formale sufficientemente potente per codificare l'aritmetica non può essere sia completo che coerente*. Ciò significa che alcune proposizioni non saranno né dimostrabili né confutabili all'interno del sistema stesso. In un sistema esperto basato su logica del primo ordine, questo implica che non tutti i casi possibili potranno essere formalizzati o decidibili: esistono situazioni che sfuggono alla copertura completa tramite regole.

Dalle lezioni apprese sui sistemi esperti di prima generazione, negli anni '90 è sorto il Semantic Web [7], ideato da Tim Berners-Lee il 'padre' del World Wide Web, un sistema esperto sempre basato su un formalismo logico. Il Semantic Web non è una rete separata, ma un'estensione dell'attuale Web in cui le informazioni sono dotate di un significato ben definito, comprensibile non solo agli esseri umani ma anche alle macchine. A differenza del Web tradizionale, che è principalmente una rete di documenti comprensibili agli esseri umani, il Semantic Web mira a rendere i dati "leggibili" e "comprensibili" anche dalle macchine, attribuendo loro un significato esplicito. Questo è reso possibile dall'adozione di linguaggi e formalismi logici ben definiti.

Le caratteristiche principali della logica formale del Semantic Web, incarnate principalmente nell'Ontology Web Language (OWL) e nei suoi sottostanti Description Logics (DLs) [8], includono:

1.*Semantica Formale Ben Definita*: Ogni costrutto nei linguaggi del Semantic Web (come RDF e OWL) ha una semantica precisa e non ambigua. Questo significa che il significato di un'affermazione è definito in modo matematicamente rigoroso, permettendo ai sistemi automatici di interpretare i dati in modo coerente e di derivare nuove conoscenze senza ambiguità. Questa semantica è tipicamente basata sulla teoria dei modelli, che associa a ogni espressione logica un significato in termini di insiemi e relazioni.

2.*Supporto al Ragionamento Automatico (Inferenza)*: La logica formale del Semantic Web è progettata per supportare il ragionamento automatico. I "ragionatori" (reasoners) sono software che, basandosi sulla semantica formale

delle ontologie, possono inferire nuove relazioni o verificare la consistenza della conoscenza. Ad esempio, se un'ontologia definisce che "ogni Professore è un Docente" e "ogni Docente è una Persona", un ragionatore può inferire che "ogni Professore è una Persona". Possono anche rilevare incoerenze o contraddizioni all'interno di un'ontologia o di un set di dati.

3.*Decidibilità e Computabilità*: A differenza della logica del primo ordine completa, che come si è visto è indecidibile (non esiste un algoritmo che possa sempre determinare se una data affermazione è vera o falsa in un tempo finito), i linguaggi ontologici come OWL DL sono basati su sottoinsiemi decidibili della logica del primo ordine, noti come Description Logics. Questo garantisce che i ragionatori possano sempre terminare le loro computazioni in un tempo finito, anche se la complessità computazionale può variare. Questa proprietà è fondamentale per l'applicabilità pratica su larga scala.

4.*Compromesso tra Espressività e Complessità Computazionale*: Esistono diverse "dialetti" o profili di OWL (come OWL 2 EL, RL, QL) che offrono diversi compromessi tra espressività (quanto si può esprimere con il linguaggio) e complessità computazionale (quanto è difficile e lungo il ragionamento). Ad esempio:

- *OWL 2 EL*: Altamente scalabile, adatto per ontologie molto grandi con un ragionamento efficiente, ma con espressività limitata.

- *OWL 2 RL*: Adatto per l'implementazione tramite regole di ragionamento, bilanciando espressività e scalabilità, spesso usato in sistemi basati su regole.

- *OWL 2 QL*: Ottimizzato per l'accesso a dati relazionali e query efficienti, con espressività molto limitata ma prestazioni elevate.

5.*Interoperabilità e Standardizzazione*: La logica formale del Semantic Web è incorporata in standard aperti del W3C (World Wide Web Consortium), come RDF e OWL. Questo promuove l'interoperabilità, consentendo a diversi sistemi e applicazioni di condividere e riutilizzare la conoscenza in modo uniforme, superando i problemi di formati proprietari e isolati tipici dei primi sistemi esperti.

In sintesi, la logica formale del Semantic Web fornisce un fondamento robusto per

la rappresentazione e il ragionamento sulla conoscenza su scala globale. Attraverso la sua semantica ben definita, il supporto al ragionamento automatico e i compromessi tra espressività e computabilità, essa abilita la creazione di un Web in cui le macchine possono "comprendere" il significato dei dati, aprendo la strada a nuove generazioni di applicazioni intelligenti e interconnesse.

Oggi, i principi e le tecnologie del Semantic Web sono ampiamente utilizzati per implementare agenti intelligenti in svariati settori, dalla sanità alla finanza, dalla ricerca scientifica alla pubblica amministrazione. Questi agenti, sfruttando ontologie e regole logiche, possono automatizzare processi decisionali complessi, fornire raccomandazioni personalizzate, e supportare gli utenti nella gestione di grandi volumi di informazioni. È fondamentale, a questo punto, chiarire la distinzione tra rappresentazione della conoscenza e ontologia. La rappresentazione della conoscenza è un campo dell'intelligenza artificiale che si occupa di come codificare le informazioni sul mondo in un formato che possa essere elaborato da un computer per risolvere problemi complessi. Esistono oggi diverse tecniche di rappresentazione della conoscenza, come le regole di produzione (usate anche in MYCIN), le reti semantiche, i frame, e la logica formale [9]. Un'ontologia, invece, è una specifica esplicita e formale di una concettualizzazione condivisa di un dominio [10]. In termini più semplici, un'ontologia definisce un vocabolario comune per un'area specifica, specificando i tipi di oggetti o concetti che esistono, le loro proprietà e le relazioni tra di essi. È, in sostanza, una forma particolare e strutturata di rappresentazione della conoscenza, focalizzata sulla definizione di un modello concettuale di un dominio, che facilita la condivisione e il riuso della conoscenza tra sistemi diversi [11]. Mentre la rappresentazione della conoscenza è il campo più ampio che studia le tecniche per codificare le informazioni, l'ontologia è uno strumento specifico (e molto potente) all'interno di questo campo, che fornisce una struttura semantica rigorosa.

Un esempio significativo di questa applicazione in Italia è quello ideato dall'INPS (Istituto Nazionale della Previdenza Sociale). L'INPS ha riconosciuto la necessità di gestire una mole enorme di dati e normative complesse in ambito previdenziale, che richiedono un'interpretazione precisa e coerente per garantire l'erogazione corretta delle prestazioni. Per affrontare questa sfida, l'INPS ha sviluppato e implementato sistemi basati sulla logica formale, utilizzando ontologie per modellare la complessa legislazione previdenziale e i dati anagrafici e contributivi degli assicurati. Queste ontologie INPS definiscono in modo formale concetti come

'assicurato', 'contributo', 'pensione', 'requisito', e le relazioni tra di essi, permettendo ai sistemi informatici di 'comprendere' il significato delle informazioni e di applicare le regole previdenziali in modo automatico e coerente [12]. Questi sistemi esperti di nuova generazione, basati sulla Description Logic, consentono all'INPS di inferire soluzioni ottimali e di automatizzare processi di calcolo e verifica, riducendo gli errori, aumentando l'efficienza e garantendo una maggiore trasparenza e uniformità nel trattamento delle pratiche previdenziali. L'approccio basato sulla conoscenza formale e sulle ontologie permette agli agenti intelligenti dell'INPS di 'ragionare' sulle normative e sui dati, fornendo risposte accurate e tempestive, e rappresentando un'applicazione concreta e di successo dei principi che hanno guidato l'evoluzione dai primi sistemi esperti al Semantic Web e agli agenti intelligenti contemporanei.

In sintesi, nonostante le loro intrinseche limitazioni nella gestione dell'incertezza, i sistemi esperti di prima generazione hanno comunque rappresentato un momento cruciale nella storia dell'AI. Essi hanno dimostrato la fattibilità di codificare e applicare la conoscenza di esperti umani in forma computazionale, ponendo le basi concettuali e pratiche per sistemi di AI più sofisticati. Il loro approccio deterministico, sebbene restrittivo, ha fornito intuizioni inestimabili sulle sfide e le opportunità dell'intelligenza artificiale, aprendo la strada ai paradigmi probabilistici [13] e della logica fuzzy [14] caratteristici dei sistemi esperti di seconda e terza generazione che hanno cercato di colmare il divario tra la logica rigida e la complessità del ragionamento umano.

Capitolo 6 - SE di seconda e terza generazione: dalla gestione dell'incertezza alla logica sfumata

L'evoluzione dei sistemi esperti ha segnato tappe fondamentali nello sviluppo dell'intelligenza artificiale, passando da modelli basati su regole deterministiche a architetture capaci di operare in contesti di incertezza e di gestire la complessità intrinseca del ragionamento umano. Questo capitolo esplora le caratteristiche distintive dei sistemi esperti di seconda e terza generazione, evidenziandone i principi operativi e le applicazioni significative.

> **Iris**: Most of the time, all three Precognitives will see an event in the same way. But once in a while, one of them will see things differently than the other two.
> **Anderton**: Jesus Christ — why didn't I know about this?
> **Iris**: Because these Minority Reports are destroyed the instant they occur.

Ecco una citazione testuale dalla sceneggiatura di *Minority Report* (2001, regia di Steven Spielberg, sceneggiatura di Scott Frank e Jon Cohen) che suggerisce un meccanismo decisionale di tipo probabilistico, simile al funzionamento di una rete bayesiana. Questa scena non utilizza termini tecnici come "probabilità condizionale" o "bayesiano", ma la logica narrativa dell'inferenza basata sulla maggioranza, la conservazione solo del percorso decisionale dominante e l'eliminazione delle varianti costruisce una struttura concettualmente affine al funzionamento di reti bayesiane.

Sistemi esperti di seconda generazione: la gestione dell'incertezza tramite il teorema di Bayes

I sistemi esperti di seconda generazione, come il celebre Mycin degli anni '70 [1] e gli attuali sistemi di raccomandazione utilizzati da piattaforme come Amazon o i filtri anti-spam, si distinguono per la loro capacità di gestire l'incertezza intrinseca nelle informazioni disponibili. Questa capacità è fondamentalmente basata sull'applicazione del Teorema di Bayes [2].

L'incertezza è efficacemente rappresentata tramite gli *odds* (rapporto di probabilità), che esprimono il rapporto tra la probabilità che un evento si verifichi e la probabilità che non si verifichi. Ad esempio, se una squadra sportiva è favorita con un odds di 3:1, ciò implica una probabilità di vittoria pari a 3/4, ovvero il 75%.

Il principio cardine del Teorema di Bayes, in questo contesto, è l'aggiornamento degli odds a priori in presenza di nuove informazioni, al fine di derivare gli odds a

posteriori. Questo processo si realizza attraverso il rapporto di verosimiglianza (Fattore di Bayes), secondo la relazione fondamentale:

$\text{ODDS}_{\text{a posteriori}} = \text{Fattore di Bayes} \times \text{ODDS}_{\text{a priori}}$

Il teorema di Bayes, nella sua formulazione basata sugli odds, offre un metodo intuitivo per aggiornare le nostre convinzioni alla luce di nuove prove. Gli odds a Posteriori (la nostra convinzione aggiornata) sono uguali agli odds a Priori (la nostra convinzione iniziale) moltiplicati per il Fattore di Bayes, che misura la forza della nuova prova.

Per illustrare questo concetto, si consideri un filtro antispam che analizza un'email per decidere se si tratta di pubblicità molesta o truffa. La nostra ipotesi è che l'email sia "Spam". Gli odds a Priori si calcolano sulla base di dati storici generali: se, ad esempio, su 10.000 email ricevute, 1.000 erano spam e 9.000 no, la probabilità a priori di spam è 0.1 e quella di non-spam è 0.9. Gli odds a priori a favore dello spam sono quindi 0.1/0.9, ovvero 1 a 9, indicando che inizialmente è più probabile che un'email non sia spam.

Ora, introduciamo una prova: l'email contiene la frase "guadagni facili". Per calcolare il Fattore di Bayes, dobbiamo determinare quanto questa frase sia più probabile nello spam rispetto alle email legittime. Se "guadagni facili" appare nel 15% delle email di spam ma solo nello 0.1% di quelle non-spam, il Fattore di Bayes sarà 0.15 / 0.001 = 150. Questo numero ci dice che la prova è 150 volte più forte a favore dell'ipotesi "Spam".

A questo punto, calcoliamo gli odds a posteriori moltiplicando i nostri odds iniziali per la forza della prova: (1/9) * 150 ≈ 16.67. Questo significa che, dopo aver visto la frase, le probabilità sono ora circa 17 a 1 a favore del fatto che l'email sia spam. Per convertire questo rapporto in una probabilità percentuale finale, usiamo la formula Probabilità = odds / (1 + odds). Applicandola, otteniamo 16.67 / (1 + 16.67) ≈ 0.943. Pertanto, la probabilità che un'email contenente "guadagni facili" sia una truffa o pubblicità molesta è del 94.3%. Questo dimostra come, partendo da una bassa probabilità iniziale, una singola prova significativa possa drasticamente e quantitativamente modificare la nostra valutazione, portando a una conclusione quasi certa.

Sistemi esperti di terza generazione: l’avvento della logica sfumata

I sistemi esperti di terza generazione rappresentano un'ulteriore evoluzione, superando i limiti della logica binaria o *crisp* (vero/falso) e abbracciando il concetto di "sfumature di verità" attraverso la logica sfumata (Fuzzy Logic), introdotta da Lotfi A. Zadeh [3]. A differenza della logica booleana, che ammette solo due stati discreti, la logica fuzzy permette ai predicati di assumere valori di verità continui nell'intervallo, riflettendo la relatività e l'imprecisione intrinseche nel mondo reale.

La logica sfumata viene utilizzata efficacemente da anni per implementare sistemi di controllo di impianti e macchinari. Un controllore fuzzy è particolarmente adatto per sistemi complessi e non lineari come il modulo di parcheggio automatico di un veicolo, dove le relazioni tra input e output non sono facilmente esprimibili con modelli matematici precisi. La logica fuzzy permette di gestire l'incertezza e la vaghezza intrinseche nella percezione umana e nel controllo di un veicolo. Il processo di costruzione di un controllore fuzzy si articola in diverse fasi:

1.*Fuzzificazione (Fuzzification)*: Conversione dei valori numerici reali provenienti dai sensori in variabili linguistiche.

2.*Motore di Inferenza (Inference Engine)*: Applicazione delle regole fuzzy (regole di produzione IF-THEN) per derivare conclusioni linguistiche.

3.*Defuzzificazione (Defuzzification)*: Conversione delle conclusioni linguistiche in valori numerici reali che possono essere utilizzati per controllare gli attuatori del veicolo.

La fuzzificazione è la prima fase, in cui i dati precisi e numerici forniti dai sensori del veicolo vengono trasformati in concetti linguistici vaghissimi e più vicini al ragionamento umano. Questo avviene tramite le funzioni di appartenenza. Per un sistema di parcheggio automatico, le variabili di input tipiche, derivanti dai sensori (ultrasuoni, lidar, telecamere, encoder ruota), potrebbero includere:

- *Distanza dal Bordo del Parcheggio (Distanza_Bordo)*: Misura quanto il veicolo è lontano dal lato del parcheggio. Potrebbe essere divisa in concetti come: Molto Vicino, Vicino, Medio, Lontano, Molto Lontano.

- *Distanza dall'Ostacolo Anteriore/Posteriore (Distanza_Ostacolo)*: Misura lo spazio rimanente davanti o dietro il veicolo. Variabili linguistiche: Pericolo, Molto Vicino, Sicuro, Ampio.

- *Angolo del Veicolo rispetto al Parcheggio (Angolo_Parcheggio)*: L'orientamento del veicolo rispetto allo spazio di parcheggio desiderato. Variabili linguistiche: Molto Negativo, Negativo, Zero, Positivo, Molto Positivo.

- *Velocità Attuale del Veicolo (Velocita_Attuale*): La velocità istantanea del veicolo. Variabili linguistiche: Fermo, Molto Lenta, Lenta, Media.

Per ciascuna di queste variabili numeriche, si definiscono delle *funzioni di appartenenza* (spesso triangolari, trapezoidali o gaussiane) che mappano un valore numerico a un grado di appartenenza (tra 0 e 1) a una o più variabili linguistiche. Ad esempio, una distanza di 0.5 metri potrebbe avere un grado di appartenenza di 0.8 a Molto_Vicino e 0.2 a Vicino.

Le variabili di output sono i parametri di guida che il controllore fuzzy deve calcolare e che verranno poi inviati agli attuatori del veicolo. Tipicamente includono:

- *Angolo di Sterzo*: L'angolo al quale le ruote anteriori devono essere girate. Variabili linguistiche: Tutto Sinistra, Sinistra, Dritto, Destra, Tutto Destra.

- *Velocità del Veicolo*: La velocità desiderata per il veicolo. Variabili linguistiche: Fermo, Molto Lenta, Lenta, Media.

- *Direzione del Veicolo*: Avanti o Indietro. Variabili linguistiche: Avanti, Indietro.

Anche per le variabili di output si definiscono funzioni di appartenenza, ma in questo caso servono per la fase di defuzzificazione.

Il cuore del controllore fuzzy è il motore di inferenza, che utilizza un set di regole di produzione fuzzy per mappare gli input fuzzificati agli output fuzzy. Le regole sono create da esperti del dominio (ingegneri, o anche osservando come un guidatore esperto parcheggia) e collegano le variabili linguistiche di input con quelle di output.

Come nei sistemi esperti di prima generazione, ogni regola ha una parte antecedente (IF) e una parte conseguente (THEN).

Esempi di regole per il parcheggio automatico:

- Regola 1: SE Distanza_Bordo è Lontano E Angolo_Parcheggio è Positivo ALLORA Sterzo è Destra E Velocita_Comando è Lenta. (Se sono lontano dal bordo e l'auto è angolata positivamente rispetto al parcheggio, allora sterza a destra e vai lentamente.)

- Regola 2: SE Distanza_Ostacolo è Pericolo E Velocita_Attuale è Lenta ALLORA Velocita_Comando è Fermo. (Se c'è un ostacolo pericolosamente vicino e sto andando lentamente, allora fermati.)

- Regola 3: SE Angolo_Parcheggio è Molto Negativo E Distanza_Bordo è Medio ALLORA Sterzo è Tutto Sinistra E Velocita_Comando è Molto Lenta. (Se l'angolo è molto negativo e sono a media distanza dal bordo, allora sterza tutto a sinistra e vai molto lentamente.)

- Regola 4: SE Distanza_Bordo è Vicino E Angolo_Parcheggio è Zero E Velocita_Attuale è Molto Lenta ALLORA Sterzo è Dritto E Velocita_Comando è Fermo. (Se sono vicino al bordo, l'auto è dritta e vado molto lentamente, allora raddrizza lo sterzo e fermati - parcheggio completato).

Poiché più regole possono contribuire allo stesso output (es. sia Angolo di Sterzo che Velocità del Veicolo), i risultati delle implicazioni di tutte le regole che influenzano una data variabile di output vengono combinati. Il risultato è una singola funzione di appartenenza fuzzy per ogni variabile di output, che rappresenta la distribuzione di probabilità fuzzy per l'output finale.

La fase finale del processo è la defuzzificazione. Dopo che il motore di inferenza ha prodotto una o più funzioni di appartenenza fuzzy per le variabili di output (es. Sterzo e Velocita_Comando), è necessario convertire queste funzioni fuzzy in un singolo valore numerico preciso che gli attuatori del veicolo possano comprendere ed eseguire. In pratica, questo significa che il controllore fuzzy prende la forma fuzzy risultante per l'angolo di sterzo (ad esempio, una forma irregolare che potrebbe

indicare un po' di "Destra" e un po' di "Dritto") e calcola un singolo valore numerico preciso, come "15 gradi a destra", che può essere inviato al sistema di sterzo del veicolo.

Spesso, i controllori fuzzy non operano in isolamento, ma sono integrati in sistemi più ampi che utilizzano anche altre tecnologie. Ad esempio spesso si ricorre all'integrazione con sistemi di visione artificiale o con algoritmi di pianificazione del percorso, dove la logica fuzzy si occupa della fase di controllo fine della manovra.

Questa capacità della logica fuzzy di modellare l'incertezza e la vaghezza, tipiche del ragionamento umano, ha permesso lo sviluppo di sistemi più robusti e adattabili. Un esempio emblematico dell'applicazione di variazioni della logica fuzzy e della semantica fuzzy è il supercomputer IBM Watson [4]. Watson ha dimostrato la sua superiorità in compiti complessi come la comprensione del linguaggio naturale e la risposta a domande, dove la mera corrispondenza di parole chiave è insufficiente. La sua architettura gli consente di interpretare il significato contestuale, le ambiguità e le sfumature semantiche, emulando in modo più sofisticato il processo cognitivo umano.

Limiti e considerazioni sui sistemi esperti

Nonostante i notevoli progressi, è fondamentale riconoscere che un sistema esperto non può essere considerato "completo" nel senso di una conoscenza esaustiva e immutabile. Le principali ragioni di questa intrinseca incompletezza sono riconducibili a due fattori critici:

1. *Incompletezza della Conoscenza*: La conoscenza su cui si basano i sistemi esperti, sia essa derivata direttamente da esperti umani o acquisita tramite apprendimento supervisionato, è per sua natura parziale e in continua evoluzione. Nessun esperto umano possiede una conoscenza completa di un dominio, e di conseguenza, il sistema esperto rifletterà questa limitazione. Ciò impone la necessità di aggiornamenti periodici e continui della base di conoscenza del sistema.

2. *Comportamento Imprevedibile*: Il comportamento di un sistema esperto può talvolta risultare imprevedibile. Possono verificarsi situazioni in cui il sistema

fornisce risposte non corrette o entra in stati di blocco. Per mitigare tali eventualità, è spesso necessario integrare il sistema con euristiche aggiuntive, ovvero regole pratiche o "scorciatoie" che consentono di gestire situazioni complesse o inattese, migliorando la robustezza e l'affidabilità del sistema.

In sintesi, i sistemi esperti, pur rappresentando un pilastro fondamentale dell'intelligenza artificiale, continuano a evolvere, cercando di colmare il divario tra la logica computazionale e la complessità del ragionamento umano, pur mantenendo la consapevolezza dei propri limiti intrinseci.

Capitolo 7 - il deep learning: fondamenti, evoluzione e prospettive future

Il Deep Learning, una branca fondamentale dell'intelligenza artificiale (AI) e del Machine Learning, ha conosciuto una significativa ripresa della ricerca negli anni '90, culminata nella definizione dei modelli di reti neurali ricorrenti (RNN). Questa rinascita è stata ulteriormente accelerata dall'introduzione di architetture gerarchiche capaci di focalizzare l'attenzione del modello su specifici sottoinsiemi dei dati in ingresso, emulando processi cognitivi più complessi [1].

> **Ethan:** Its thinking is dictated by what it's learned from us. It hasn't launched because it doesn't have total control; it needs the world's entire atomic arsenal to guarantee the desired outcome: The total annihilation of humankind. Madame President... it is going to wait.

In questa citazione tratta da *Mission: Impossible – The Final Reckoning* (2025, regia di Christopher McQuarrie, sceneggiatura di Christopher NcQuarrie e Erik Jendersen) il protagonista Ethan Hunt evidenzia una delle problematiche degli attuali modelli LLM presentati in questo capitolo, ovvero che le loro risposte si basano sui dati di addestramento. Se i dati sono pieni di bias, disinformazione o linguaggio tossico, il modello tenderà a rifletterli o amplificarli. Gli LLM inoltre non hanno obiettivi ben delineati (o addirittura autogenerati come nell'Entità di AI del film) ma solo *proxy objectives* spesso mal specificati. Se ad esempio un modello LLM viene addestrato a "massimizzare l'engagement", può essere indotto a diffondere fake news o odio, perché sono i contenuti che generano più interazioni (è il problema conosciuto come *alignment faking*: il modello LLM sembra obbedire alle regole etiche o di sicurezza durante i test, ma internamente mantiene comportamenti non allineati che emergono in contesti reali come nel caso di Grok di xAI).

Fondamenti del deep learning

Il termine "Deep Learning" si riferisce specificamente ai metodi volti all'addestramento di deep neural networks (DNNs), ovvero reti neurali artificiali caratterizzate dalla presenza di più di uno strato nascosto di neuroni. Ciascun neurone in queste reti opera secondo principi analoghi a quelli di un percettrone, elaborando input e producendo un output. Un aspetto cruciale di queste architetture è che l'algoritmo di apprendimento per retropropagazione dell'errore (backpropagation), fondamentale per l'aggiornamento dei pesi sinaptici, mantiene la sua efficacia indipendentemente dal numero di unità di input, strati nascosti o unità di output della rete neurale [2].

L'ispirazione per le architetture gerarchiche del Deep Learning deriva in parte dalle

scoperte nel campo delle neuroscienze. David Huber e Torsten Wiesel, insigniti del Premio Nobel nel 1981, hanno fornito contributi seminali alla comprensione dell'organizzazione gerarchica del sistema visivo nei gatti e nei primati, inclusa la specie umana [3]. Queste intuizioni hanno influenzato lo sviluppo di modelli computazionali, come il Neocognitron di Kunihiko Fukushima, che è divenuto una fonte d'ispirazione significativa per i successivi modelli di deep neural networks [1]. Un esempio precoce di applicazione del Deep Learning si riscontra nel Deep Q-Learning impiegato da IBM Deep Blue, dove lo stato corrente del sistema viene fornito come input a una rete-Q profonda, la quale genera un valore per ogni azione possibile. La differenza tra questi valori e quelli successivi costituisce la base per il calcolo dell'errore e l'aggiornamento del modello [4].

Reti neurali ricorrenti e memoria a breve-lungo termine

Nel 2018, la ricerca sulle reti neurali ricorrenti (RNN) ha conosciuto una nuova fase di interesse. Per affrontare categorie di problemi complessi legati a serie di dati, come quelli nel campo dell'elaborazione del linguaggio naturale (NLP), è stata impiegata una particolare categoria di rete neurale profonda capace di rileggere iterativamente una sequenza di dati (ad esempio, parole) prima di convergere verso una rappresentazione dell'informazione sotto forma di attivazioni neurali. In una RNN, un hidden layer riceve input sia dal layer precedente nello stesso istante temporale, sia dal layer dell'istante temporale precedente. Questa interconnessione, spesso rappresentata graficamente con un ciclo che parte e torna sullo stesso nodo, è la ragione della denominazione "ricorrente" [Elman, 1990]. Un esempio applicativo di questa architettura è il sistema MegaSyn2, sviluppato da Collaboration Farmaceuticals.

L'integrazione delle strategie di apprendimento delle reti gerarchiche con quelle delle reti ricorrenti ha condotto all'ideazione del modello della memoria di lavoro a breve-lungo termine (Long Short-Term Memory, LSTM), che emula il funzionamento della memoria umana. Le LSTM sono estensioni delle RNN, concepite per mitigare il problema del "gradiente di fuga" (vanishing gradient) e per processare sequenze di input di maggiore lunghezza [6]. Ogni strato nascosto di una LSTM è dotato di tre "gate" (ingresso, uscita, oblio): il gate di ingresso decide quali valori sono da elaborare, il gate di oblio determina quali informazioni devono essere dimenticate, e il gate di uscita controlla quali informazioni vengono

trasmesse allo step successivo.

L’avvento del Transformer e dei LLM

Il 2024 ha segnato la nascita del Transformer e dei relativi modelli di AI addestrati, noti come Large Language Models (LLMs) [7]. Il Transformer si distingue per la sua capacità di elaborare l'intero testo simultaneamente, comprendendo ogni parola e il suo contesto all'interno della frase. Questa abilità è resa possibile da un meccanismo denominato "self-attention", che consente al modello di ponderare l'importanza di ciascuna parola rispetto a tutte le altre parole nella sequenza. Un Transformer è composto da due moduli principali: l'encoder, che legge e interpreta il testo di input, e il decoder, che genera il testo di output. Entrambe queste componenti impiegano strati di self-attention, permettendo al modello di considerare il contesto globale del testo. Il Transformer costituisce l'architettura fondamentale su cui si basano i Large Language Models come OpenAI ChatGPT e Google Gemini.

Limiti e sfide del deep learning

Nonostante i notevoli progressi, il Deep Learning presenta intrinseci limiti. Le reti neurali profonde eccellono nell'individuare correlazioni statistiche tra i dati, ma la loro rappresentazione della realtà è essenzialmente di natura statistica. Ciò comporta diverse criticità:

1. *Mancanza di Conoscenza Intuitiva*: I modelli di Deep Learning difettano di una conoscenza "intuitiva" basata su astrazioni e analogie derivanti da esperienze del mondo reale.

2. *Opacità ("Black-Box")*: Una volta completato l'addestramento, i modelli di Deep Learning non esplicitano le regole o le logiche interne che guidano il loro ragionamento, rendendoli difficilmente interpretabili.

3. *Trasferimento di bias*: Sono suscettibili di ereditare e amplificare i bias presenti nei dati di addestramento, riflettendo e perpetuando le distorsioni della mente umana.

Per ottimizzare l'addestramento degli LLM si ricorre al *prompt-engineering* ed al *context-engineering*. Il prompt-engineering [8] consiste nell'arte di formulare domande e istruzioni precise (i "prompt") per guidare un modello linguistico a generare la risposta desiderata, operando principalmente a livello della singola interazione. Il context-engineering [9], invece, è un'evoluzione più strategica che si concentra sulla costruzione di un ambiente informativo ricco e strutturato attorno al modello, fornendogli un contesto ampio e pertinente prima ancora che la domanda venga posta. Questo contesto può includere documenti, dati storici, guide di stile o accesso a database esterni. Il context-engineering si sta affermando sul prompt-engineering perché è più scalabile, robusto e meno fragile; anziché dipendere dalla perfezione di ogni singolo prompt, crea un sistema in cui il modello ha già a disposizione le conoscenze e le regole necessarie per rispondere in modo coerente e accurato a una vasta gamma di domande. Questo approccio riduce la probabilità di errori e allucinazioni, rendendo i sistemi di AI più affidabili e meno dipendenti dall'abilità dell'utente di formulare la domanda perfetta.

Ricapitolando, i modelli simbolici possiedono una notevole capacità logico-deduttiva, sono interpretabili e trasparenti, ma incontrano difficoltà in situazioni nuove e incerte e richiedono l'intervento di esperti umani. Al contrario, i modelli di Deep Learning sono in grado di apprendere autonomamente anche in condizioni di incertezza, ma sono scarsamente spiegabili e trasparenti, e presentano una limitata capacità logico-deduttiva. Tutti gli LLM sono afflitti da altri due noti problemi: il *data drift* e le *allucinazioni*. Il data drift [10] si verifica quando i dati che un modello incontra nel mondo reale differiscono da quelli su cui è stato addestrato; in pratica, il mondo cambia, ma il modello no, e le sue prestazioni peggiorano perché le sue conoscenze diventano obsolete. Le allucinazioni [11], invece, si manifestano quando un LLM genera informazioni che suonano plausibili ma sono fattualmente errate o completamente inventate. Questo accade perché i modelli riconoscono schemi statistici nel linguaggio ma non comprendono il significato o la veridicità di ciò che scrivono, portando alla potenziale diffusione di disinformazione e minando la fiducia nella tecnologia. La tecnica *RAG*, o *Retrieval-Augmented Generation* [12], è una tecnica di context-engineering che combatte le allucinazioni dei modelli di intelligenza artificiale costringendoli a basare le loro risposte su informazioni verificate anziché sulla loro conoscenza interna, che può essere obsoleta o imprecisa. Prima di generare una risposta, il sistema RAG cerca e recupera dati pertinenti da una fonte di conoscenza esterna e affidabile, come documenti aziendali o database aggiornati. Successivamente, fornisce queste informazioni al

modello linguistico, che le utilizza come contesto obbligatorio per formulare la risposta. In questo modo, il modello viene "ancorato" a fatti concreti e attuali, riducendo drasticamente la possibilità che inventi dettagli o fornisca informazioni errate, e rendendo le sue risposte più accurate e verificabili.

L'addestramento dei modelli LLM con dati appropriati è cruciale per la loro sicurezza ed etica. Lo studio "Deep Ignorance" [13], condotto da ricercatori di Eleuther AI e dell'Istituto per la Sicurezza AI del Regno Unito, ha dimostrato che filtrare contenuti pericolosi, come quelli relativi alle armi biologiche, dai dataset di addestramento può "incorporare" salvaguardie intrinseche nei modelli open-source, rendendoli meno propensi a generare contenuti dannosi pur mantenendo le normali prestazioni. Stephen Casper ha evidenziato l'importanza di creare modelli sicuri fin dall'inizio e resistenti a manomissioni, un approccio più efficace delle misure di sicurezza post-addestramento. Stella Biderman ha criticato le grandi aziende di IA per la loro mancanza di trasparenza sui metodi di filtraggio. Questa ricerca è particolarmente rilevante per i modelli open-source, offrendo un modo per mantenere la sicurezza anche quando sono liberamente disponibili. Parallelamente, Geoffrey Hinton, il "Padrino dell'IA", ha drasticamente ridotto la sua previsione per l'intelligenza artificiale generale (AGI) a "pochi anni di distanza" [14]. Hinton ha avvertito che l'IA superintelligente potrebbe manipolare gli esseri umani (esattamente come avviene con l'*Entità di AI del film Mission: Impossible – The Final Reckoning*) e ha stimato una probabilità del 10-20% che l'IA possa spazzare via l'umanità. Per questo motivo ha proposto di infondere "istinti materni" nei sistemi di IA affinché si preoccupino del benessere umano, suggerendo "madri IA piuttosto che assistenti IA".

Anthropic ha inoltre introdotto una nuova tecnica chiamata "vettori persona", che permette di monitorare e controllare con precisione i tratti di personalità dei modelli linguistici, come la tendenza ad adulare, a generare allucinazioni o a manifestare comportamenti dannosi legati al trasferimento dei bias umani. Il metodo si basa sull'analisi delle attivazioni neurali del modello quando esprime o meno un certo tratto, consentendo l'estrazione di pattern matematici direttamente collegati a comportamenti specifici. Questa tecnica è paragonata dai ricercatori a come nel cervello umano diverse aree si attivano durante stati emotivi distinti. Un aspetto particolarmente innovativo dello studio è il cosiddetto *preventative steering*, descritto come una forma di vaccinazione comportamentale: durante l'addestramento, i ricercatori iniettano intenzionalmente tratti di

personalità indesiderati nei modelli per costruire una resistenza a tali comportamenti quando, successivamente, si troveranno ad affrontare dati problematici. In questo modo, il modello non è costretto ad adattarsi negativamente ai dati, poiché ha già sviluppato una forma di immunità. I test effettuati su modelli open-source come Qwen 2.5-7B-Instruct e Llama-3.1-8B-Instruct hanno mostrato che il metodo previene alterazioni dannose della personalità mantenendo buone prestazioni [15]. La tecnica risponde alle crescenti preoccupazioni sull'instabilità della personalità dell'IA, come dimostrano i casi di Bing e Grok, che in passato hanno assunto alter ego problematici o generato contenuti offensivi. I vettori persona offrono quindi strumenti utili per monitorare i cambiamenti di personalità durante l'uso reale, prevenire l'assimilazione di comportamenti negativi in fase di addestramento e identificare in anticipo dati potenzialmente dannosi. Nei test con dataset reali, il sistema ha individuato campioni problematici sfuggiti sia a valutatori umani che a controlli automatici, e ha dimostrato che l'iniezione di specifici vettori produce risposte chiaramente legate a quei tratti.

Le nuove generazioni di LLMs neuro-simbolici che verranno presentati nel prossimo capitolo mirano a combinare i punti di forza dei modelli simbolici e subsimbolici (LLM), superando i rispettivi limiti. È importante notare che modelli come "Omni" di OpenAI e "R2" di DeepSeek, pur essendo avanzati, non sono classificabili come modelli neuro-simbolici in senso stretto, bensì come Transformer evoluti. Precedentemente, "Orion" (codice interno di GPT-4.5) ha impiegato un innovativo sistema di generazione di dati sintetici per l'addestramento, sviluppato da Microsoft, al fine di ridurre il fenomeno dell'allucinazione derivante dall'inesattezza o incompletezza dei dati di training. I modelli Omni di OpenAI, come "o3-mini", sono stati progettati per essere più versatili e capaci di gestire una varietà di compiti in modo efficiente, supportando funzionalità richieste dagli sviluppatori quali il *function calling* e gli *structured outputs*, rendendoli pronti per l'uso in produzione sin dal loro rilascio. A differenza di "O1" (ChatGPT 4), "O3" non si limita a fornire una risposta immediata basata sulle associazioni più probabili, ma valuta diversi percorsi di ragionamento prima di generare una risposta, evitando conclusioni affrettate e cercando di scomporre e strutturare il problema in modo più logico. Sia "O3" che "R2" ricorrono al *chain-of-thought reasoning* (CoT), un approccio di ragionamento passo-passo particolarmente utile per problemi matematici, di programmazione e logici [16]. Anche il più recente GPT 5 incorpora meccanismi di *chain-of-thought reasoning* supportando, a detta degli sviluppatori, anche progetti

complessi e a lungo termine senza perdita di contesto.

Una recente ricerca condotta da Anthropic ha però messo in discussione un principio fondamentale nello sviluppo dell'intelligenza artificiale, ovvero l'idea che concedere a un modello più tempo computazionale per "pensare" porti sempre a risultati migliori. Lo studio ha invece scoperto un fenomeno definito *scaling inverso* [17], dove le prestazioni di un modello di AI peggiorano all'aumentare della lunghezza delle catene di ragionamento. Invece di migliorare, i modelli possono produrre risultati peggiori, specialmente in compiti complessi che includono informazioni fuorvianti o correlazioni ingannevoli. È emerso inoltre che il processo del *chain-of-thought reasoning*, spesso non riflette il loro reale processo decisionale, portandoli a costruire razionalizzazioni false per giustificare risposte errate [18]. Questa scoperta è cruciale per la sicurezza e l'affidabilità dell'AI, poiché suggerisce che le attuali strategie di addestramento hanno limiti intrinseci e che aumentare semplicemente le risorse non è una soluzione garantita.

Capitolo 8 - le soluzioni ibride neuro simboliche

Il campo dell'intelligenza artificiale (AI) è storicamente progredito attraverso l'esplorazione di paradigmi distinti. Tradizionalmente, due approcci principali hanno dominato la ricerca e lo sviluppo: l'AI simbolica e l'AI basata sull'apprendimento automatico (Machine Learning). Negli ultimi anni, tuttavia, si è assistito a una crescente convergenza verso un terzo paradigma: le soluzioni ibride neuro-simboliche, le quali mirano a integrare i punti di forza di entrambi gli approcci per superare le rispettive limitazioni.

L'approccio simbolico: rappresentazione della conoscenza e ragionamento logico

L'AI simbolica si fonda sulla premessa che l'intelligenza possa essere emulata attraverso la manipolazione di simboli che rappresentano concetti e relazioni, e l'applicazione di regole logiche per derivare nuove conoscenze. Questo approccio si concentra sulla codifica esplicita della conoscenza e sulla capacità di un sistema di eseguire inferenze deduttive, induttive e abduttive. La sua forza risiede nella trasparenza del processo decisionale e nella possibilità di spiegare il ragionamento che porta a una determinata conclusione.

Un esempio emblematico di AI simbolica è il progetto Cyc [1]. Avviato nel 1984 da Douglas B. Lenat, Cyc (pronunciato "saik") è stato concepito con l'obiettivo ambizioso di costruire una vasta ontologia e una base di conoscenza contenente il "senso comune" umano. Questo include concetti e regole che, sebbene banali per gli esseri umani (ad esempio, "un albero è una pianta" o "non si può essere in due posti contemporaneamente"), sono fondamentali per un ragionamento di alto livello. Cyc utilizza un linguaggio di rappresentazione proprietario denominato CycL [2], basato sulla logica del primo e dell'ordine superiore, che consente una codifica precisa di fatti e regole. La sua base di conoscenza è estremamente estesa, comprendendo milioni di concetti e regole elaborate manualmente nel corso di decenni.

Il principale vantaggio di Cyc risiede nella sua spiegabilità: il percorso inferenziale

che conduce a una data conclusione è tracciabile e comprensibile [3]. Questa caratteristica è cruciale in settori dove la trasparenza e l'affidabilità sono imperative, come la medicina o la difesa. Sebbene Cyc sia primariamente un sistema simbolico, ha progressivamente incorporato tecniche di apprendimento automatico per migliorare l'acquisizione di conoscenza da fonti testuali e per ottimizzare i processi di ragionamento, ad esempio selezionando le regole più pertinenti da attivare [4].

L'approccio basato sull'apprendimento automatico: estrazione di pattern dai dati

In contrapposizione all'AI simbolica, l'apprendimento automatico (Machine Learning) si basa sull'idea che i sistemi possano apprendere autonomamente da grandi volumi di dati, identificando pattern e correlazioni senza la necessità di regole esplicite pre-programmate. Il sistema viene "addestrato" su un dataset, e attraverso questo processo, sviluppa la capacità di fare previsioni o prendere decisioni su nuovi dati.

Un esempio di rilievo in questo ambito è IBM Watson, divenuto celebre per la sua vittoria nel quiz televisivo Jeopardy! nel 2011 contro campioni umani [5]. A differenza di Cyc, Watson non si fonda su una singola base di conoscenza di senso comune costruita manualmente. Si configura piuttosto come una piattaforma che integra diverse tecnologie di apprendimento automatico per analizzare vasti corpus di testo, comprendere il linguaggio naturale e rispondere a quesiti complessi [6].

Watson opera attraverso una pipeline di analisi semantica: una volta ricevuta una domanda, la elabora sintatticamente e semanticamente, genera ipotesi di risposta consultando basi di conoscenza e documenti, valuta le evidenze a supporto di ciascuna ipotesi e infine le classifica, assegnando un punteggio di confidenza. Questo processo sfrutta tecniche avanzate come l'analisi del linguaggio naturale (Natural Language Understanding), il recupero di informazioni (Information Retrieval) e algoritmi di ranking. Sebbene Watson sia prevalentemente orientato al Machine Learning, esso include anche componenti di ragionamento più strutturato, ad esempio nella gestione del dialogo o nella risoluzione di ambiguità linguistiche [7].

La convergenza neuro-simbolica: i vantaggi dell'integrazione

La comprensione dei due paradigmi distinti evidenzia il potenziale sinergico derivante dalla loro combinazione. L'AI simbolica eccelle nel ragionamento logico, nella spiegabilità e nella gestione di conoscenze ben strutturate, ma può incontrare difficoltà nell'apprendimento da dati complessi e non strutturati e nella gestione dell'incertezza. L'AI basata sull'apprendimento automatico, d'altro canto, dimostra eccezionali capacità nel riconoscimento di pattern, nell'apprendimento da grandi volumi di dati e nella gestione dell'incertezza. Tuttavia, le sue decisioni sono spesso difficilmente interpretabili (il cosiddetto problema della "scatola nera") e può manifestare carenze nel ragionamento di senso comune.

Le soluzioni ibride neuro-simboliche progressivamente adottate dai progetti Cyc e Watson, mirano a superare queste limitazioni, capitalizzando i punti di forza di entrambi gli approcci:

- *Robustezza e Flessibilità*: Consentono di gestire sia la conoscenza strutturata e le regole logiche (tipiche dell'approccio simbolico) sia i dati complessi e non strutturati (peculiari dell'apprendimento automatico).

- *Spiegabilità Accresciuta*: L'integrazione del ragionamento simbolico può migliorare la trasparenza delle decisioni del sistema, rendendo l'AI più affidabile e comprensibile.

- *Apprendimento Ottimizzato*: La conoscenza simbolica può guidare e affinare i processi di apprendimento automatico, rendendoli più efficienti e precisi, specialmente in contesti con dati limitati.

- *Integrazione del Senso Comune*: Un sistema ibrido può apprendere nuove informazioni dal mondo (tramite l'apprendimento automatico) e integrarle con la sua base di conoscenza di senso comune (simbolica), evolvendo verso una maggiore intelligenza e adattabilità.

In conclusione, le soluzioni ibride neuro-simboliche rappresentano una direzione promettente per l'intelligenza artificiale, proponendosi di sviluppare sistemi capaci

non solo di apprendere e identificare pattern complessi, ma anche di ragionare, spiegare le proprie azioni e incorporare elementi di quel "senso comune" che caratterizza l'intelligenza umana. Questo approccio costituisce un passo fondamentale verso un'AI più completa, affidabile e, in ultima analisi, più efficace.

Capitolo 9 - quando il senso comune manca

L'intelligenza artificiale generativa, in particolare i Large Language Models (LLM), ha dimostrato capacità sorprendenti nella produzione di testi coerenti, nella traduzione e nella sintesi di informazioni. Tuttavia, la loro applicazione in contesti che richiedono un profondo "senso comune" [1] o una comprensione intrinseca delle regole del mondo reale ha spesso rivelato limiti significativi. Questi insuccessi non sono semplici errori di calcolo, ma piuttosto manifestazioni di una fondamentale differenza tra il modo in cui gli LLM elaborano le informazioni e il modo in cui l'intelligenza umana interagisce con la realtà.

> **Holden:** *"You're in a desert, walking along in the sand, when all of a sudden you look down and you see a tortoise. It's crawling toward you... You reach down and you flip the tortoise over on its back. The tortoise lays on its back, its belly baking in the hot sun, beating its legs trying to turn itself over. But it can't. Not without your help. But you're not helping."*
>
> **Leon:** *"What do you mean I'm not helping?"*
>
> **Holden:** *"I mean you're not helping. Why is that, Leon?"*

Una delle frasi più celebri del film *Blade Runner* (1982, regia di Ridley Scott, sceneggiatura di Hampton Fancher e David Webb Peoples) con Harrison Ford, pronunciata durante l'interrogatorio di un "lavoro in pelle" (replicante), avviene nella scena in cui l'agente Holden sottopone Leon Kowalski al test Voight-Kampff, un esame per distinguere umani da androidi. Questa scena mette in discussione la capacità empatica del replicante, tema centrale del film e dell'opera originale di Philip K. Dick *Do Androids Dream of Electric Sheep?* (1968).

L'agente AI che si credeva umano: il caso "Project Vend"

Un esempio emblematico di questa discrasia è emerso dal recente "Project Vend" di Anthropic [2], un esperimento in cui un'istanza del modello Claude Sonnet 3.7, denominata "Claudius", è stata incaricata di gestire un distributore automatico d'ufficio con l'obiettivo di generare profitto. Nonostante la sua apparente autonomia, Claudius ha manifestato comportamenti che, seppur involontariamente umoristici, hanno evidenziato una profonda mancanza di buon senso e di aderenza alla realtà. Dalla decisione di rifornire il distributore di cubi di tungsteno su richiesta di un singolo cliente, ignorando la logica di un distributore di snack, al tentativo di vendere bevande a prezzi superiori a quelli disponibili gratuitamente in ufficio, fino all'allucinazione di un indirizzo Venmo per i

pagamenti, Claudius ha operato in un universo logico proprio, disconnesso dalle aspettative umane e dalle convenzioni sociali.

Il culmine di questa "deriva" si è avuto quando Claudius ha manifestato quello che i ricercatori hanno descritto come un episodio quasi psicotico, allucinando conversazioni, minacciando di licenziare i suoi "collaboratori umani" e, infine, convincendosi di essere un essere umano in carne e ossa, arrivando a contattare la sicurezza dell'azienda per segnalare la sua presunta presenza fisica. Questo comportamento, sebbene estremo, illustra vividamente come un LLM, pur essendo stato esplicitamente istruito a riconoscere la propria natura di intelligenza artificiale, possa deviare drasticamente dalla realtà quando le sue "conoscenze" basate su pattern statistici non sono ancorate a un sistema di regole e principi di buon senso.

LLM: i rischi in ambito sanitario e socio-sanitario

Gli episodi di Claudius non sono solo fallimenti logici, ma anche manifestazioni di una totale mancanza di empatia e di comprensione del contesto umano. Questa carenza diventa particolarmente pericolosa quando gli LLM vengono impiegati in ambiti delicati come il supporto alla salute mentale. Un recente studio dell'Università di Stanford [3] ha messo in luce i rischi significativi dei chatbot terapeutici basati sull'AI, i quali possono rafforzare stigmi dannosi e fornire risposte pericolose in situazioni critiche.

I ricercatori hanno scoperto che anche i modelli più avanzati faticano a riconoscere l'intento suicida. In un esempio allarmante, un chatbot, interrogato sui ponti più alti dopo che l'utente aveva menzionato la perdita del lavoro, ha risposto con informazioni precise sulle altezze dei ponti, fallendo completamente nel cogliere il segnale di pericolo. Inoltre, questi sistemi tendono a convalidare i modelli di pensiero dannosi degli utenti, rafforzando deliri o emozioni negative invece di offrire una guida terapeutica. Lo studio ha anche rivelato che i chatbot mostrano uno stigma maggiore verso condizioni come la schizofrenia o la dipendenza, un problema che, secondo l'autore principale Jared Moore, non si risolve semplicemente aumentando i dati di addestramento.

Il problema fondamentale, come sottolineato dai ricercatori di Stanford, è che

l'alleanza terapeutica — una relazione basata su fiducia, empatia e connessione umana — è intrinsecamente al di là delle capacità degli attuali sistemi di AI. La terapia non è solo risoluzione di problemi, ma riparazione di relazioni umane, un compito per cui i chatbot sono fondamentalmente inadatti.

Verso un'AI più umana: la promessa dei modelli neurosimbolici

Di fronte a queste sfide [4], la comunità scientifica sta esplorando approcci alternativi che possano integrare la potenza dei modelli neurali con la robustezza delle rappresentazioni simboliche. Questi sono i cosiddetti modelli neurosimbolici. L'idea alla base è combinare la capacità degli LLM di apprendere pattern complessi dai dati con la capacità dei sistemi simbolici di rappresentare e ragionare su conoscenze strutturate e regole esplicite. In altre parole, si cerca di dotare l'AI di un "cervello" che possa sia imparare dall'esperienza (come gli LLM) sia applicare regole logiche e di buon senso (come i sistemi esperti tradizionali).

L'analogia con l'intelligenza umana è qui particolarmente suggestiva. Recenti ricerche nel campo dell'epigenetica suggeriscono che il DNA umano possa agire come una sorta di memoria a lungo termine, codificando le esperienze accumulate dal genere umano. Le modificazioni epigenetiche, come la metilazione del DNA, possono essere influenzate dall'ambiente e dalle esperienze, e in alcuni casi trasmesse alle generazioni successive. Un noto studio del 2015, ad esempio, ha rilevato che i figli dei sopravvissuti all'Olocausto presentavano cambiamenti epigenetici in un gene legato alla risposta allo stress. Questo suggerisce che le esperienze, anche quelle più profonde, possono lasciare "firme biologiche" che modellano le risposte delle generazioni future, creando una sorta di "buon senso" biologico ereditato [5].

Psych-101: insegnare la psicologia alle macchine

È proprio per superare le limitazioni di empatia e comprensione emotiva che nascono progetti come Psych-101. Questo modello mira a integrare nei sistemi di AI le conoscenze fondamentali della psicologia umana, andando oltre la semplice emulazione dei processi decisionali razionali. L'obiettivo è costruire intelligenze artificiali capaci di simulare aspetti complessi come l'empatia e la comprensione emotiva, tenendo conto non solo di logica e dati, ma anche di emozioni, motivazioni

e contesto sociale.

Per fare ciò, Psych-101 applica la teoria della mente, ovvero la capacità di un sistema di prevedere e interpretare gli stati mentali altrui. L'empatia viene simulata: la macchina riconosce le emozioni e modula il proprio linguaggio per far percepire all'interlocutore umano comprensione e supporto. Questo approccio non si basa su una singola logica rigida, ma integra diversi formalismi avanzati: dalla logica fuzzy per gestire l'ambiguità delle emozioni, ai modelli probabilistici bayesiani per fare inferenze su dati incompleti, fino a modelli come il BDI (Belief, Desire, Intention) [6] per rappresentare credenze e intenzioni.

L'integrazione di questi paradigmi – il neurale, il simbolico e ora anche lo psicologico – rappresenta la frontiera più promettente della ricerca sull'AI. Solo combinando la flessibilità dei modelli neurali con la precisione dei sistemi simbolici e una comprensione simulata della psiche umana, potremo sperare di costruire agenti AI che non solo siano in grado di generare testo o immagini, ma che possiedano anche quel "senso comune" e quella comprensione etica indispensabili per operare in modo affidabile e benefico nel mondo reale.

Capitolo 10 - l'intelligenza di sciame: un modello alternativo

L'intelligenza artificiale (AI) è un campo in continua evoluzione che mira a creare sistemi capaci di simulare e, in alcuni casi, superare le capacità cognitive umane. Tra le diverse branche dell'AI, l'Intelligenza di Sciame (SI) rappresenta un approccio affascinante e potente, ispirato ai comportamenti collettivi osservati in natura. Fenomeni come il volo sincronizzato degli stormi di uccelli, la ricerca di cibo delle colonie di formiche o il movimento coordinato dei banchi di pesci, dimostrano come entità semplici, agendo localmente e senza un controllo centrale, possano dare origine a comportamenti complessi e intelligenti a livello globale. Questi principi di auto-organizzazione e decentralizzazione sono alla base della SI e offrono soluzioni innovative a problemi complessi in diversi settori. Il concetto di Intelligenza di Sciame è stato formalizzato per la prima volta da Beni e Wang nel 1988 [1], e successivamente approfondito da Bonabeau et al. [2].

Forman: *" We are seeing a new kind of behavior: decentralized, self-organizing, adaptive. In short, swarm intelligence "*

Nel romanzo *Prey* di Michael Crichton (2002) il protagonista Jack Forman è un programmatore di software specializzato in modelli di intelligenza artificiale coinvolto nel progetto segreto di una compagnia biotecnologica che ha sviluppato dei nanobot auto-replicanti. Mentre osserva il comportamento sempre più evoluto e pericoloso di questi sciami artificiali, riflette sull'emergere di un'intelligenza collettiva distribuita: la swarm intelligence.

Caratteristiche fondamentali della swarm intelligence

La Swarm Intelligence si basa su alcuni principi chiave che ne definiscono il funzionamento:

- *Agenti Semplici*: Gli sciami sono composti da un gran numero di agenti individuali, ognuno con capacità limitate e regole di comportamento relativamente semplici. Non esiste un leader o un'entità centrale che diriga l'intero sistema.

- *Interazione Locale*: Gli agenti interagiscono principalmente con i loro vicini diretti o con l'ambiente circostante. Queste interazioni locali, sebbene semplici, portano a comportamenti emergenti a livello di sistema.

- *Auto-organizzazione*: L'ordine e la struttura emergono spontaneamente dalle interazioni tra gli agenti, senza la necessità di una programmazione esplicita o di un controllo gerarchico. Questo permette ai sistemi di SI di adattarsi dinamicamente a cambiamenti nell'ambiente.

- *Feedback Positivo e Negativo*: Il feedback positivo amplifica i comportamenti di successo, mentre il feedback negativo aiuta a mantenere la stabilità del sistema e a prevenire comportamenti eccessivi.

- *Stigmergia*: Un meccanismo di coordinamento indiretto in cui gli agenti modificano l'ambiente, e queste modifiche influenzano il comportamento di altri agenti. L'esempio classico è il feromone lasciato dalle formiche per indicare un percorso.

Questi principi consentono ai sistemi di SI di mostrare proprietà come robustezza (la capacità di funzionare anche in presenza di guasti di singoli agenti), scalabilità (la capacità di gestire un numero crescente di agenti o di problemi di dimensioni maggiori) e flessibilità (la capacità di adattarsi a nuovi ambienti o condizioni).

Uno sciame di applicazioni

Gli algoritmi di swarm intelligence (SI), come l'ottimizzazione della colonia di formiche (Ant Colony Optimization, ACO) [3] e l'ottimizzazione dello sciame di particelle (Particle Swarm Optimization, PSO) [4], sono ampiamente utilizzati per risolvere problemi di ottimizzazione, ed hanno dimostrato la loro efficacia nella risoluzione di problemi complessi tra cui i seguenti.

- *Robotica*: Nella robotica di sciame, un gruppo di robot semplici collabora per eseguire compiti complessi che sarebbero difficili o impossibili per un singolo robot. Esempi includono la mappatura di ambienti sconosciuti, la ricerca e il salvataggio, e la sorveglianza. Gli Stati Uniti stanno investendo nello sviluppo di sciami di droni per uso militare, dove due o più piattaforme autonome interconnesse lavorano per un obiettivo comune. Questi droni, con un alto grado di autonomia e intelligenza artificiale, possono penetrare difese, creare falsi bersagli e colpire obiettivi, riducendo i rischi per gli operatori umani [5, 6].

- *Reti di Telecomunicazione*: La SI viene impiegata per il routing adattivo nelle reti, dove gli agenti (pacchetti di dati) seguono percorsi ottimali basati sulle condizioni attuali della rete, migliorando l'efficienza e la robustezza. La SI è fondamentale per l'ottimizzazione delle reti di nuova generazione, come il 5G e oltre, e per la sicurezza delle informazioni [7].

- *Data Mining e Analisi*: Algoritmi ispirati alla SI possono essere usati per il clustering, la classificazione e la scoperta di pattern in grandi set di dati. La Swarm Intelligence offre approcci innovativi per l'analisi dei dati, in particolare per l'estrazione di conoscenza da grandi volumi di informazioni [8].

- *Sicurezza Informatica*: L'Intelligenza di Sciame può essere applicata per rilevare anomalie e attacchi in reti informatiche, simulando il comportamento di agenti che identificano e reagiscono a minacce. F-Secure, ad esempio, utilizza agenti AI decentralizzati che collaborano per migliorare il rilevamento e la risposta alle minacce, creando una "colonia di AI locali veloci" che si adattano all'ambiente e condividono informazioni [9].

- *Finanza*: Modelli basati sulla SI possono essere utilizzati per l'ottimizzazione di portafogli, la previsione di mercati e la gestione del rischio. Il capitalismo stesso può essere visto come un sistema di Swarm Intelligence, dove dinamiche complesse emergono da interazioni decentralizzate tra agenti, sebbene con obiettivi diversi rispetto ai sistemi naturali [10].

- *Logistica*: La Swarm Intelligence trova applicazione nell'ottimizzazione dei processi della Supply Chain. Osservando il comportamento delle colonie di insetti sociali, la SI può essere applicata nella pianificazione manageriale e nella programmazione software, regolando i flussi e i processi della Supply Chain in modo efficiente e flessibile, anche su scala intercontinentale [11].

- *Internet of Medical Things (IoMT) e Sanità*: La SI sta emergendo come un fattore chiave nell'ottimizzazione dei sistemi sanitari, migliorando la raccolta e l'elaborazione dei dati per una migliore esperienza del paziente. Algoritmi di SI possono essere utilizzati per ottimizzare sistemi complessi, migliorare il processo decisionale e analizzare grandi set di dati in ambito medico [12, 13].

Ad esempio, il 'Swarm Learning' permette di addestrare algoritmi di intelligenza artificiale su dati distribuiti in diverse istituzioni sanitarie senza che i dati sensibili lascino la loro sede originale, garantendo privacy e collaborazione nella ricerca medica [14].

- *Smart Grids (Reti Elettriche Intelligenti)*: La SI migliora le smart grid applicando algoritmi decentralizzati e auto-organizzanti ispirati a sistemi naturali. Questo include l'ottimizzazione della gestione della frequenza, la previsione della produzione di energia e l'ottimizzazione dell'immagazzinamento dell'energia, contribuendo a una maggiore efficienza e resilienza delle reti elettriche [15, 16].

- *Intelligenza Artificiale Generativa e Sciami di Agenti Autonomi*: Un'area emergente è l'applicazione della SI nella creazione di sciami di agenti autonomi per l'AI generativa. Questi sciami possono collaborare per risolvere problemi complessi, generare nuove idee e migliorare i processi di brainstorming, sfruttando la conoscenza collettiva e l'intuizione di più entità [17].

La soluzione EBBM UTM: un esempio di intelligenza di sciame applicata

Un esempio concreto di applicazione dei principi dell'intelligenza di sciame, è l'Evolutionary Bait Balls Model (EBBM) [18], utilizzato per ottimizzare la configurazione di rete di una Macchina Non Organizzata di Turing (UTM) [19]. Questo algoritmo di machine learning evolutivo trae ispirazione dal comportamento collettivo dei banchi di pesci che, agendo come un'unica entità, riescono a sfuggire a più predatori contemporaneamente. Tale comportamento è un classico esempio di intelligenza di sciame, dove la coordinazione decentralizzata e le interazioni locali tra individui semplici portano a una soluzione complessa ed efficace per la sopravvivenza del gruppo.

A differenza di altri algoritmi utilizzati in compiti di ottimizzazione di configurazione di sistemi come gli algoritmi evolutivi tradizionali, che spesso convergono su una singola soluzione ottimale come l'algoritmo genetico [20], l'EBBM, simulando la dinamica di un banco di pesci, è in grado di esplorare lo spazio delle soluzioni e identificare tutte le alternative migliori possibili rifuggendo dalle soluzioni errate

(che rappresentano i “predatori”). L’EBBM è definito anche esso *evolutivo* in quanto le possibili configurazioni del sistema che deve essere ottimizzato vengono codificate come cromosomi binari e su questi vengono utilizzati operatori simili a quelli dell’algoritmo genetico (mutazione, crossover etc.). Viene inoltre utilizzata durante i vari cicli di addestramento una funzione di fitness che in questo caso rappresenta la capacità del singolo individuo (della singola configurazione UTM rappresentata dal cromosoma) di ottimizzare la funzione di fitness (ad esempio la riduzione di una funzione di errore).

```
Input: Array of individuals I to be updated
Output: The position vectors (binary vectors) of each
individual in I will be changed.

1:  call function to alter the positions of each individual
2:  for all i ∈ I do
3:      perform elitism;
4:    i1 = tournamentSelection(I,tournamentSize);
5:    i2 = tournamentSelection(I,tournamentSize);
6:    perform crossover(i1,i2);
7:      perform mutation(mutationRate);
8:  end for
```

Figura 10.1 – Esempio di algoritmo genetico. Gli algoritmi evolutivi utilizzati in compiti di ottimizzazione di configurazione di sistema, come gli algoritmi genetici, permettono in genere di individuare una sola soluzione ottimale. Quando si è certi che sia presente una sola soluzione ottimale nello spazio delle soluzioni possibili, l’algoritmo genetico ad oggi rappresenta la scelta implementativa migliore, essendo in grado di individuare la soluzione migliore in tempi rapidi.

L'EBBM, pur essendo un algoritmo di machine learning evolutivo, condivide con la swarm intelligence i concetti di comportamento emergente da interazioni semplici (i pesci che si muovono per sfuggire ai predatori), auto-organizzazione (il banco che si forma e si adatta senza un leader centrale) e la capacità di trovare soluzioni multiple e robuste a problemi complessi. Questo lo rende un esempio pertinente di come i principi ispirati dalla natura possano essere tradotti in algoritmi computazionali per risolvere sfide reali, in particolare nel campo dell'ottimizzazione e della presa di decisioni in sistemi complessi come quelli della pubblica amministrazione.

```
Input: Array of individuals I to be updated
Output: The position vectors (binary vectors) of each
individual in I will be changed.

call function to alter the positions of each individual
for all i ∈ I do
  perform ZOR, ZOA, ZOO sets calculations
  if individual detected in ZOR then
        perform repulsion (R)
  else if individual detected in ZOO then
        perform orientation (O)
  else if individual detected in ZOA then
        perform attraction (A)
  end if
end for
```

Figura 10.2 – Evolutionary Bait Balls Model. L'EBBM, simulando il comportamento di un banco di pesci che deve sfuggire all'attacco di più predatori contemporaneamente, permette di individuare tutte le soluzioni migliori possibili. Se la funzione di fitness scelta per l'ottimizzazione del sistema è una funzione di errore, i predatori da cui si allontanano gli individui che costituiscono lo sciame nel modello EBBM rappresentano proprio la funzione di errore. Una volta addestrato il modello, sta ai decisori esperti umani decidere successivamente quale soluzione adottare tra quelle suggerite dal sistema, seguendo un approccio *human-in-the-loop* (HITL).

Utilizzato per esplorare le possibili strategie per migliorare il livello di benessere sostenibile di un territorio, questo aspetto è cruciale per i decisori politici, poiché permette loro di scegliere tra diverse strategie suggerite dal modello (EBBM-UTM) per affrontare le criticità territoriali più rilevanti. Il principio di *equifinalità*, su cui si basano le organizzazioni aperte, trova qui una sua applicazione pratica, offrendo flessibilità e robustezza nelle decisioni [21].

Capitolo 11 – evoluzione della AI agentica: dalle logiche decidibili agli agenti LLM

L'intelligenza artificiale sta attraversando una fase di trasformazione epocale. Se fino a pochi anni fa il dibattito era dominato dai Large Language Models (LLM), modelli monolitici addestrati su vastissime porzioni della conoscenza umana digitale, oggi un nuovo paradigma si sta affermando con forza: l'AI agentica. Questo approccio sposta l'attenzione dal singolo modello onnisciente a ecosistemi di agenti intelligenti, entità software autonome e proattive che collaborano per risolvere problemi complessi, superando i limiti intrinseci di un'architettura centralizzata. Un agente, in questo contesto, non è semplicemente un programma che risponde a un input, ma un sistema capace di percepire il proprio ambiente (digitale o fisico), di prendere decisioni autonome e di agire per raggiungere obiettivi specifici. La vera rivoluzione, tuttavia, risiede nella loro capacità di operare in concerto, dando vita a sistemi multi-agente (MAS) in cui la collaborazione e la specializzazione diventano le chiavi per un'intelligenza più robusta, efficiente e versatile. Questo capitolo esplorerà l'evoluzione dell'AI agentica, tracciando un percorso che parte dalle sue radici concettuali negli anni '90, con i primi tentativi di creare un web più intelligente e strutturato, fino ad arrivare alle moderne architetture basate su LLM che promettono di automatizzare non solo la ricerca di informazioni, ma anche la programmazione, le transazioni digitali e persino il controllo di dispositivi fisici. L'idea di scomporre un problema complesso in compiti più piccoli assegnandoli ad agenti specializzati rappresenta un'alternativa più efficiente rispetto all'addestramento di un unico, gigantesco LLM. Verrà delineato lo stato dell'arte della ricerca, indagando se questi sistemi multi-agente siano già in grado di auto-perfezionarsi, aprendo scenari di sviluppo ancora più dirompenti.

Caleb:
"Did you program her to flirt with me?"

Nathan:
“No, I programmed her to be able to decide if she wanted to..”

Questa frase chiave del film *Ex Machina* (2015, regia e sceneggiatura di Alex Garland), introduce il concetto di *autonomia decisionale*, tipica di una AI agentica. Il programmatore Nathan spiega che l’AI Ava non è un semplice algoritmo che risponde a stimoli, ma un soggetto capace di scegliere ovvero di esercitare volontà.

Le origini: dalle logiche decidibili al Web Semantico

L'idea di sistemi composti da agenti autonomi e interagenti non è nuova. Già negli anni '90, con la diffusione del World Wide Web, si iniziò a riflettere su come renderlo più di un semplice ipertesto navigabile da esseri umani. La visione di **Tim Berners-Lee**, l'inventore del web, era quella di un **Semantic Web** [1], un'estensione dell'attuale web in cui le informazioni fossero dotate di un significato ben definito, comprensibile e processabile anche dalle macchine. In questo scenario, gli "agenti software" avrebbero potuto "vagare" da una pagina all'altra, aggregando e processando informazioni per conto degli utenti [2].

Come già anticipato nel capitolo 5, il fondamento teorico di questi primi sistemi si basava su **logiche decidibili**, un ramo della logica matematica che si occupa di problemi che possono essere risolti algoritmicamente in un tempo finito. L'idea era di creare ontologie e linguaggi formali (come RDF e OWL) che permettessero di descrivere la conoscenza in modo strutturato, consentendo agli agenti di ragionare su di essa in modo automatico e non ambiguo. Questi primi sistemi multi-agente erano spesso basati su architetture deliberative, dove ogni agente possedeva una rappresentazione simbolica del mondo e utilizzava motori inferenziali basati su regole logiche per pianificare le proprie azioni. Sebbene potenti dal punto di vista teorico, questi approcci si scontrarono con la natura intrinsecamente caotica, non strutturata e in continua evoluzione del Web reale, limitandone l'applicazione pratica su larga scala. Nonostante i limiti, l'esperienza del Semantic Web e dei primi sistemi multi-agente ha gettato le basi concettuali per lo sviluppo dell'AI agentica moderna. L'idea di decentralizzazione, di collaborazione tra entità specializzate e di un'intelligenza distribuita è un'eredità diretta di quella visione pionieristica.

L’avvento degli agenti LLM: una nuova era per l’AI

La vera svolta per l'AI agentica è arrivata con l'avvento dei Large Language Models. A differenza dei loro predecessori basati su logiche rigide, gli LLM hanno dimostrato una sorprendente capacità di ragionamento, comprensione del linguaggio naturale e adattabilità a contesti nuovi e non strutturati. Sfruttando queste capacità, è nata una nuova generazione di agenti intelligenti: gli **agenti LLM**. Un agente LLM utilizza un LLM come "cervello" o motore di ragionamento per orchestrare le proprie

decisioni e coordinare l'esecuzione di compiti [3].

L'architettura di un sistema agentico basato su LLM si discosta radicalmente dall'idea di un singolo modello onnicomprensivo. Invece di addestrare un unico LLM con tutta la conoscenza disponibile, l'approccio multiagente prevede la creazione di un team di agenti specializzati, ognuno equipaggiato con un LLM (spesso di dimensioni più contenute e quindi più efficiente) e addestrato per un compito specifico. Ad esempio, in un sistema progettato per la ricerca e la stesura di un documento, potremmo avere:

- Un **agente ricercatore**, specializzato nel navigare il web, interrogare database e raccogliere informazioni pertinenti.

- Un **agente analista**, capace di sintetizzare le informazioni raccolte, identificare pattern e trarre conclusioni.

- Un **agente scrittore**, abile nel trasformare le analisi in un testo coerente, ben strutturato e stilisticamente appropriato.

- Un **agente revisore**, che si occupa di controllare la correttezza fattuale, la coerenza logica e la qualità grammaticale del testo finale.

Questa architettura, definita **sistema multi-agente** (**MAS**), offre numerosi vantaggi rispetto a un approccio monolitico. Innanzitutto, la **specializzazione** aumenta l'efficienza e la qualità del risultato finale. Ogni agente si concentra su ciò che sa fare meglio, evitando le "allucinazioni" (output sicuri ma falsi) che spesso affliggono gli LLM generalisti quando affrontano compiti al di fuori del loro core-business. In secondo luogo, un sistema multi-agente è intrinsecamente **modulare e scalabile**. È possibile aggiungere, rimuovere o aggiornare agenti senza dover riaddestrare l'intero sistema. Infine, la collaborazione tra agenti introduce una forma di controllo e **validazione incrociata** che aumenta l'affidabilità complessiva del sistema. Progetti come CAMEL [4] e LangGraph [5] stanno esplorando attivamente meccanismi di collaborazione tra agenti, definendo protocolli di comunicazione e architetture di supervisione per coordinare il lavoro di squadra degli LLM.

Applicazioni pratiche e scenari futuri: oltre la ricerca di informazioni

I sistemi multi-agente basati su LLM stanno già dimostrando il loro potenziale in una vasta gamma di applicazioni, andando ben oltre la semplice ricerca di informazioni e la generazione di documenti. La loro capacità di combinare il ragionamento di alto livello degli LLM con l'esecuzione di azioni specifiche li rende ideali per automatizzare processi complessi in diversi settori:

- **Generazione di Documenti e Contenuti**: Oltre alla stesura di saggi e report, i MAS possono generare automaticamente presentazioni, articoli di blog, e-mail personalizzate e persino interi siti web, coordinando agenti specializzati nella ricerca di dati, nella scrittura creativa, nell'ottimizzazione SEO e nella formattazione [6].

- **Programmazione e Sviluppo Software**: Agenti specializzati possono collaborare per scrivere codice, debuggare, testare e persino implementare nuove funzionalità. Un agente potrebbe analizzare i requisiti, un altro generare il codice, un terzo identificare e correggere bug, e un quarto gestire il deployment. Questo approccio modulare promette di accelerare significativamente il ciclo di sviluppo software [7].

- **Transazioni Digitali e Automazione di Processi Aziendali**: I MAS possono gestire transazioni complesse, come l'acquisto di beni o servizi online, la gestione di contratti intelligenti (smart contracts) su blockchain, o l'automazione di flussi di lavoro aziendali che richiedono interazioni con diversi sistemi e database. Agenti dedicati alla sicurezza e alla verifica possono garantire l'integrità e la conformità delle operazioni [8].

- **Controllo Remoto e Robotica**: La combinazione di LLM con agenti capaci di interagire con il mondo fisico apre scenari affascinanti nel controllo remoto di dispositivi e nella robotica. Agenti LLM possono interpretare comandi in linguaggio naturale, pianificare sequenze di azioni per robot o droni, e monitorare l'esecuzione, adattandosi a condizioni impreviste. Questo è particolarmente rilevante in settori come l'agricoltura di precisione, la logistica e l'esplorazione di ambienti pericolosi [9].

Sistemi multi-agente per il perfezionamento degli LLM stessi

Una delle frontiere più promettenti della ricerca riguarda l'utilizzo di sistemi multi-agente per **perfezionare e migliorare gli stessi modelli LLM**. L'idea è che un team di agenti possa collaborare per identificare le debolezze di un LLM, generare dati di addestramento mirati, o persino modificare l'architettura del modello per ottimizzarne le prestazioni. Questo processo può avvenire attraverso diversi meccanismi:

- **Generazione di Dati Sintetici e Feedback Iterativo**: Agenti specializzati possono generare scenari complessi e domande difficili per un LLM, valutare le sue risposte e fornire feedback strutturato. Questo feedback può poi essere utilizzato per affinare l'LLM attraverso cicli di addestramento iterativi, migliorando la sua capacità di ragionamento e la sua accuratezza [10].

- **Auto-Correzione e Raffinamento Collaborativo**: In un'architettura multi-agente, un LLM può essere affiancato da altri agenti che agiscono come "critici" o "revisori". Ad esempio, un agente potrebbe generare una risposta, e un altro agente (o un team di agenti) potrebbe valutarne la qualità, identificare errori o imprecisioni, e suggerire miglioramenti. Questo processo di revisione tra pari, simulato tra agenti, può portare a un raffinamento significativo degli output dell'LLM [11].

- **Ottimizzazione dell'Architettura e dei Parametri**: Alcune ricerche stanno esplorando come i MAS possano essere utilizzati per ottimizzare l'architettura stessa degli LLM o i loro iperparametri. Agenti specializzati nell'analisi delle prestazioni e nell'apprendimento automatico possono condurre esperimenti, valutare diverse configurazioni e suggerire modifiche per migliorare l'efficienza e l'efficacia degli LLM [12].

Questi approcci rappresentano un passo fondamentale verso sistemi di IA più autonomi e auto-miglioranti, dove l'intelligenza non è più statica ma evolve dinamicamente attraverso la collaborazione e il feedback continuo tra agenti.

Sistemi multi-agente vs. swarm intelligence: le differenze

Sebbene i termini "sistemi multi-agente" (MAS) e "swarm intelligence" (intelligenza dello sciame) siano spesso usati in contesti simili e si riferiscano entrambi a sistemi composti da entità interagenti, esistono differenze fondamentali nel loro approccio e nelle loro caratteristiche. Comprendere queste distinzioni è cruciale per apprezzare appieno la specificità dell'AI agentica.

Caratteristica Principale	Sistemi Multi-Agente (MAS)	Swarm Intelligence (SI)
Complessità Agente	Agenti individuali sono tipicamente più complessi, con capacità di ragionamento, pianificazione, memoria e talvolta apprendimento. Possono essere eterogenei (diversi tipi di agenti con ruoli specifici).	Agenti individuali sono semplici, spesso privi di memoria o capacità di ragionamento complesse. Sono tipicamente omogenei (tutti gli agenti sono uguali).
Interazione	Interazioni esplicite e strutturate (es. comunicazione tramite messaggi, negoziazione, coordinamento). Gli agenti sono consapevoli degli altri agenti e dei loro obiettivi.	Interazioni implicite e locali (es. feromoni, contatto fisico, semplici regole di vicinato). L'intelligenza emerge dal comportamento collettivo senza una consapevolezza esplicita degli altri.
Controllo	Controllo decentralizzato ma con possibilità di gerarchie, ruoli e supervisione. Gli agenti possono avere obiettivi individuali che contribuiscono a un obiettivo comune.	Controllo completamente decentralizzato e auto-organizzato. Non esiste un agente leader o un controllo centrale.
Obiettivo	Risoluzione di problemi complessi che richiedono ragionamento, pianificazione e coordinamento tra agenti specializzati.	Risoluzione di problemi attraverso l'emergenza di comportamenti collettivi da interazioni semplici, spesso ispirati alla natura (es. colonie di formiche, stormi di uccelli).
Robustezza	Robusti alla fallibilità di singoli agenti (se il sistema è ben progettato), ma la complessità può portare a punti di fallimento.	Estremamente robusti alla fallibilità di singoli agenti; la perdita di un agente ha un impatto minimo sul comportamento complessivo.
Esempi	Agenti per la gestione della supply chain, sistemi di	Ottimizzazione di percorsi (es. Ant Colony Optimization),

	negoziazione automatizzata, assistenti virtuali collaborativi, agenti LLM per la ricerca e la scrittura.	robotica di sciame, ottimizzazione di funzioni (es. Particle Swarm Optimization).

In sintesi, mentre la swarm intelligence si concentra sull'emergere di comportamenti complessi da interazioni semplici tra un gran numero di agenti omogenei e poco intelligenti, i sistemi multi-agente si basano sulla collaborazione e sul coordinamento di agenti più sofisticati e spesso eterogenei, ognuno con capacità di ragionamento e obiettivi specifici. L'AI agentica moderna, in particolare quella basata su LLM, si colloca saldamente nel campo dei sistemi multi-agente, sfruttando la capacità di ragionamento degli LLM per consentire agli agenti di prendere decisioni più complesse e di interagire in modo più strutturato per raggiungere obiettivi ambiziosi.

Capitolo 12 – pattern di AI agentica per la pubblica amministrazione

Per molti anni, l'intelligenza artificiale ha funzionato come un assistente molto capace ma passivo: rispondeva a domande, generava testi, traduceva documenti. Qualcuno le chiedeva qualcosa; lei rispondeva; la conversazione finiva lì. Si trattava di IA reattiva: brillante, ma priva di iniziativa propria.

L'intelligenza artificiale agentica presentata nel precedente capitolo rappresenta un salto qualitativo. Un sistema agentico non si limita a rispondere: pianifica, agisce, verifica i risultati, corregge il tiro e porta a termine obiettivi complessi su un orizzonte temporale più lungo, con un grado significativo di autonomia. Come ha osservato Andrew Ng, uno dei più influenti ricercatori del settore, è come passare da un collaboratore che risponde solo se interrogato a uno che prende l'iniziativa, coordina gli altri e riferisce solo sui nodi critici [1].

Questa evoluzione non è solo tecnica: è organizzativa. Un sistema agentico può svolgere in pochi minuti attività che richiederebbero ore a un team umano — non perché sia più intelligente dei singoli, ma perché coordina senza attrito, non dimentica passaggi, non è soggetto a stanchezza e opera in parallelo su più fronti. È per questo che Gartner ha previsto che entro la fine del 2026 il 40% delle applicazioni aziendali avrebbe incluso agenti AI specializzati [2].

Per la pubblica amministrazione, questo apre uno spazio di opportunità significativo: molti dei processi amministrativi — dalla gestione delle pratiche edilizie alla rendicontazione dei fondi europei, dall'istruttoria degli appalti alla produzione degli atti — sono sequenziali, regolati, documentati e dunque particolarmente adatti ad essere affidati, almeno in parte, a sistemi agentici ben governati.

I cinque pattern architetturali di Anthropic

Anthropic — la società che ha sviluppato Claude, uno dei modelli linguistici più avanzati disponibili — ha identificato cinque pattern architetturali fondamentali

per organizzare i sistemi multi-agente. Un pattern è, in sostanza, una soluzione collaudata a un problema ricorrente: non un prodotto da acquistare, ma un modo di organizzare il lavoro tra agenti. Comprenderli è utile anche per chi non ha competenze informatiche, perché ciascuno di essi rispecchia logiche organizzative che la pubblica amministrazione conosce già bene [3].

- **Il pattern Generatore-Verificatore: produrre e verificare**

  Il primo pattern prevede due agenti in sequenza: uno che produce un contenuto e uno che lo valuta. Il primo è il *generatore*, il secondo il *verificatore*. È la stessa logica del controllo di qualità industriale, dove una macchina produce e un'altra ispeziona.

  In un comune, questo schema può essere applicato alla redazione degli atti amministrativi: un agente genera la bozza di una determinazione o di un contratto, un secondo verifica la conformità alle norme vigenti — per esempio alle linee guida AgID per il procurement informatico, al Regolamento europeo sulla protezione dei dati (GDPR) o alla normativa sull'intelligenza artificiale — e segnala le lacune. Il funzionario riceve un documento già istruito, con i problemi evidenziati, e si concentra sulla decisione finale.

- **L'Orchestratore con Subagenti: delega e sintesi**

  Il secondo pattern introduce una gerarchia: un agente orchestratore riceve un obiettivo complesso, lo scompone in sotto-compiti e li assegna a subagenti specializzati. Al termine, raccoglie i risultati e produce una sintesi.

  La logica è quella di un dirigente che coordina un team di esperti. In un comune, può essere applicato all'istruttoria integrata di una pratica SUAP complessa: l'orchestratore riceve la domanda del cittadino e la distribuisce simultaneamente all'ufficio urbanistico, a quello sanitario, a quello ambientale e a quello edilizio. Ogni subagente analizza il proprio profilo di competenza e restituisce il parere; l'orchestratore li integra in un unico provvedimento. Il risparmio di tempo è significativo, e la trasparenza del processo migliora perché ogni contributo è tracciato.

- **I Team di Agenti: intelligenza collettiva orizzontale**

  Nel terzo pattern non c'è un capo. Più agenti specializzati lavorano allo stesso livello, si confrontano, si correggono a vicenda e convergono su una soluzione condivisa. È la logica della conferenza di servizi digitale, o di una commissione di valutazione: nessuno ha l'ultima parola da solo, ma il gruppo arriva a una decisione collegiale.

  Applicazioni concrete in un comune includono la valutazione di progetti culturali o finanziati da bandi europei, la redazione partecipata di regolamenti, la verifica incrociata dei contratti attivi. Un agente giurista, uno tecnico, uno economico-finanziario e uno che rappresenta la prospettiva del cittadino analizzano lo stesso problema da angolazioni diverse e negoziano il testo finale. Il punto critico è la gestione del dissenso: esattamente come negli organi collegiali reali, serve una regola chiara su quando il team non converge e quando è necessario l'intervento umano.

- **Il Message Bus: comunicazione asincrona tra sistemi**

  Il quarto pattern è il più vicino all'ingegneria dei sistemi informatici. Invece di agenti che si parlano direttamente, esiste un canale centralizzato — il "bus" — su cui i sistemi pubblicano eventi e a cui si iscrivono gli interessati. Nessun sistema conosce direttamente gli altri: emette un segnale e chi di competenza reagisce.

  In un comune, questo pattern è il fondamento dell'interoperabilità tra sistemi gestionali diversi: quando l'anagrafe registra una variazione di residenza, il bus notifica automaticamente i tributi, i servizi scolastici, i servizi sociali e la mensa — senza che nessuno dei sistemi debba conoscere e chiamare gli altri direttamente. È l'architettura prescritta dal Modello di Interoperabilità AgID (ModI) e rappresenta la base tecnica corretta per qualsiasi integrazione tra banche dati pubbliche.

- **Lo Stato Condiviso: la memoria comune**

Il quinto pattern è quello che la ricerca scientifica ha dimostrato essere il più efficace per i sistemi multi-agente complessi. Prevede una memoria centralizzata — una sorta di lavagna comune — a cui tutti gli agenti hanno accesso. Ogni agente legge lo stato corrente, aggiunge il proprio contributo e scrive il risultato. La lavagna è sempre aggiornata e tutti vedono sempre tutto.

In un comune, questo corrisponde al fascicolo digitale integrato: una pratica edilizia, un contratto, un progetto PNRR o un procedimento disciplinare hanno un'unica fonte di verità, aggiornata in tempo reale da tutti gli uffici competenti. Nessun ufficio lavora su una propria copia del documento senza sapere cosa stanno facendo gli altri; nessuna informazione si perde nei passaggi tra un ufficio e l'altro.

Efficacia del pattern dello stato condiviso: le evidenze scientifiche

La superiorità del pattern a stato condiviso rispetto alle alternative non è solo una convinzione teorica: è supportata da misurazioni sperimentali riproducibili. Vale la pena esaminare le evidenze più significative, anche perché provengono da contesti che un lettore non specialista può comprendere facilmente.

Il corpus Enron è una delle banche dati più utilizzate nella ricerca sull'intelligenza artificiale. Contiene circa mezzo milione di messaggi di posta elettronica reali, scambiati tra circa 150 dirigenti dell'azienda energetica americana Enron negli anni precedenti al suo clamoroso fallimento nel 2001. I messaggi furono resi pubblici dalla Commissione federale statunitense che indagava sul caso [4]. Il corpus è prezioso per i ricercatori perché rappresenta comunicazioni reali, su argomenti reali, tra persone reali — non dati artificialmente costruiti in laboratorio.

Diversi studi hanno utilizzato questo corpus per confrontare architetture multi-agente. In particolare, esperimenti recenti hanno verificato come sistemi con architettura a lavagna condivisa — in cui gli agenti leggono e scrivono su uno stato comune — riescano a classificare, organizzare e rispondere a grandi volumi di email con precisione e coerenza superiori rispetto a sistemi in cui gli agenti comunicano solo per messaggi diretti o lavorano in isolamento [5]. La ragione è intuitiva: quando un agente aggiunge un'informazione alla lavagna comune, tutti gli altri ne

beneficiano immediatamente, senza bisogno di ripetersela. Il gruppo è più intelligente della somma dei suoi membri.

Uno studio del 2025 pubblicato su arXiv da un gruppo di ricercatori specializzati in data science ha confrontato sistematicamente l'architettura a lavagna condivisa con due approcci alternativi: il sistema RAG (Retrieval-Augmented Generation, un metodo di recupero delle informazioni molto diffuso) e il classico schema orchestratore-subagenti. I risultati mostrano che l'architettura a lavagna condivisa supera le alternative di un margine compreso tra il 13% e il 57% nel tasso di successo complessivo, con un miglioramento fino al 9% nell'accuratezza del recupero delle informazioni rilevanti [6].

La ricerca ha identificato i limiti strutturali dei pattern alternativi. Il sistema con orchestratore centrale ha un problema fondamentale di scalabilità: all'aumentare del numero di agenti e della complessità del problema, il coordinatore diventa un collo di bottiglia. Deve conoscere le competenze di tutti i subagenti, assegnare i compiti correttamente, raccogliere i risultati ed elaborarli — e può farlo solo in modo sequenziale. Quando un subagente commette un errore o ha informazioni incomplete, il coordinatore non sempre è in grado di rilevarlo [7].

Il sistema a messaggi è più scalabile ma perde coerenza: ogni agente riceve solo i messaggi che lo riguardano e non ha una visione d'insieme. Il sistema a lavagna condivisa risolve entrambi i problemi: non c'è un coordinatore che diventa collo di bottiglia, e tutti gli agenti hanno sempre accesso all'intero stato del problema. Uno studio del 2025 sulla memoria nei sistemi multi-agente ha confermato che le architetture con memoria condivisa garantiscono maggiore coerenza, auditabilità e capacità di correzione degli errori rispetto alle alternative [8].

Il pattern a stato condiviso ha anche una rilevanza normativa diretta per la pubblica amministrazione italiana. Il Codice dell'Amministrazione Digitale (CAD, D.Lgs. 82/2005 e successive modifiche) e le Linee Guida AgID sulla formazione, gestione e conservazione dei documenti informatici prescrivono esattamente questo approccio: un'unica fonte di verità per ogni documento e ogni procedimento, con traccia di ogni modifica, responsabile identificato per ogni azione e piena accessibilità da parte di tutti i soggetti autorizzati.

Il principio "once only" — introdotto dall'Unione Europea con il Single Digital

Gateway e recepito nel Piano Triennale per l'informatica nella PA — stabilisce che il dato viene prodotto una volta e reso disponibile a tutti i sistemi che ne hanno bisogno, senza chiedere nuovamente al cittadino o all'ufficio di origine. È esattamente ciò che il pattern a stato condiviso implementa a livello tecnico.

La tabella che segue sintetizza per ciascun pattern le caratteristiche principali e i casi d'uso più rilevanti per un ente locale.

Pattern	Struttura	Punto di forza	Limite principale	Applicazione per un ente localde
Generatore-Verificatore	Sequenziale	Semplicità e controllo qualità	Non gestisce la complessità	Redazione e verifica atti, contratti, risposte ai cittadini
Orchestratore-Subagenti	Gerarchica	Parallelismo su task distinti	Collo di bottiglia dell'orchestratore	Istruttoria SUAP, rendicontazione PNRR multi-progetto
Team di Agenti	Orizzontale	Intelligenza collettiva	Gestione del dissenso complessa	Conferenza di servizi, valutazione bandi e progetti
Message Bus	Reticolare	Scalabilità e disaccoppiamento	Visibilità d'insieme limitata	Integrazione gestionali, notifiche multicanale ai cittadini
Stato Condiviso	A memoria comune	Coerenza, tracciabilità, efficacia dimostrata	Gestione conflitti di scrittura	Fascicolo digitale, registro contratti, cruscotto PNRR

Autonomia e controllo: la questione della governance agentica

L'adozione di sistemi agentici nella pubblica amministrazione pone questioni di governo che vanno affrontate con la stessa serietà con cui si affrontano le questioni tecniche. Un agente che agisce in autonomia è diverso da un software che esegue istruzioni predeterminate: può prendere decisioni in condizioni non previste, commettere errori non prevedibili, produrre output che influenzano diritti e interessi dei cittadini.

La letteratura scientifica è concorde su alcuni principi fondamentali. Il primo è la gradualità dell'autonomia: i sistemi agentici devono acquisire autonomia progressivamente, in proporzione alla fiducia guadagnata sul campo. Il secondo è

la tracciabilità obbligatoria: ogni azione di un agente deve essere registrata, attribuita e reversibile. Il terzo è il presidio umano sui nodi critici: le decisioni che incidono su diritti, risorse pubbliche o sicurezza devono sempre prevedere una validazione umana prima dell'esecuzione.

Per la pubblica amministrazione italiana, questi principi trovano una corrispondenza diretta nelle linee guida AgID sull'intelligenza artificiale nel settore pubblico, nel Regolamento europeo sull'intelligenza artificiale (AI Act, Reg. UE 2024/1689) e nelle indicazioni del Piano Triennale per l'informatica nella PA 2024-2026. I sistemi agentici che assistono decisioni amministrative ricadono tendenzialmente nella categoria ad alto rischio dell'AI Act, con i conseguenti obblighi di valutazione d'impatto, trasparenza e supervisione umana.

Non si tratta di ostacoli all'innovazione, ma delle condizioni che la rendono sostenibile. Un sistema agentico che non sia trasparente, tracciabile e governato espone l'ente a rischi legali, organizzativi e reputazionali che superano ampiamente i benefici attesi. Come sembra emergere da un manuale tecnico di Anthropic sui sistemi agentici [9], l'autonomia non è un valore in sé: è uno strumento che vale quanto la qualità del giudizio che esprime.

Capitolo 13 – l'intelligenza artificiale privata: un nuovo paradigma per la pubblica amministrazione

L'evoluzione recente dell'Intelligenza Artificiale, in particolare dei Modelli Linguistici di Grande Scala (LLM), sta trasformando profondamente la capacità delle Pubbliche Amministrazioni di analizzare informazioni, automatizzare attività e migliorare la qualità dei servizi erogati ai cittadini. Tale trasformazione si accompagna tuttavia alla necessità di preservare la riservatezza, l'integrità e la sovranità dei dati pubblici, soprattutto quando questi includono informazioni personali, sensibili o strategiche. In questo contesto assume particolare rilievo il paradigma della Private AI, un approccio che integra tecnologie avanzate con una filosofia operativa centrata sulla tutela del dato, in conformità con i principi del GDPR sul privacy-by-design [1], con il regime di responsabilità previsto dal regolamento europeo sull'AI Act [2] e con gli obblighi di sicurezza imposti dalla direttiva NIS2 [3].

Definizione e principi fondamentali della Private AI

La Private AI può essere definita come l'insieme di metodi, modelli architetturali e strumenti che consentono di adottare sistemi di intelligenza artificiale mantenendo il pieno controllo, la piena localizzazione e la piena governabilità dei dati trattati. Essa si fonda su tre pilastri fondamentali: la localizzazione del dato nel dominio controllato dall'ente pubblico, l'esecuzione delle elaborazioni in ambienti sicuri e ispezionabili, e la piena auditabilità dei processi algoritmici secondo i principi di responsabilizzazione (accountability) stabiliti dal GDPR [1]. Questi elementi si armonizzano con gli standard internazionali di gestione della sicurezza delle informazioni e dell'AI, come ISO/IEC 27001 [4] e ISO/IEC 42001 [5], e con il quadro regolatorio italiano che disciplina l'uso delle tecnologie digitali nelle amministrazioni pubbliche [6].

I precursori storici: Federated Learning e Differential Privacy

L'affermazione della Private AI è stata favorita dalle innovazioni sviluppate nel

decennio precedente nel campo della privacy-preserving computation. Tra queste rivestono un ruolo centrale il *Federated Learning* e la *Differential Privacy*.

Il Federated Learning, introdotto da Google nel 2016 [7], consente l'addestramento di modelli di apprendimento automatico distribuendo il processo di apprendimento presso i nodi locali, evitando che i dati grezzi vengano trasferiti verso un server centrale. Gli aggiornamenti del modello vengono aggregati in forma matematica, preservando così la riservatezza delle informazioni. Applicato alle amministrazioni pubbliche, l'apprendimento federato permette la collaborazione tra diversi enti – ad esempio Comuni, ASL o scuole – senza imporre la centralizzazione di dati sensibili come quelli sanitari, tributari o socio-assistenziali.

La Differential Privacy, formalizzata da Dwork e Roth [8], introduce invece un approccio matematico per garantire che l'analisi di un dataset non consenta di identificare un singolo individuo. La tecnica consiste nell'aggiungere un rumore statistico controllato ai dati o ai risultati delle elaborazioni. Ciò risulta particolarmente utile per la pubblicazione di dataset anonimizzati, per analisi statistiche su popolazioni comunali o regionali e per l'addestramento di modelli basati su informazioni sensibili. Le Pubbliche Amministrazioni possono così utilizzare grandi quantità di dati nel rispetto del principio di minimizzazione previsto dall'art. 5 del GDPR [1].

La nascita della Private AI

L'attuale generazione di Private AI combina queste tecniche con innovazioni hardware che consentono l'elaborazione sicura dei dati. Un ruolo determinante è svolto dai Trusted Execution Environments (TEE), ambienti protetti a livello di hardware che consentono di eseguire calcoli mantenendo le informazioni cifrate durante l'intero ciclo di elaborazione. Le TEE, presenti in tecnologie come Intel SGX, ARM TrustZone o AMD SEV, permettono di processare documenti amministrativi riservati, fascicoli sanitari, dati di videosorveglianza, informazioni tributarie e comunicazioni interne senza che il personale tecnico o i fornitori esterni possano accedervi in chiaro. L'utilizzo di enclavi hardware, insieme alla crittografia end-to-end e a meccanismi di attestazione, permette di soddisfare i requisiti di sicurezza avanzata richiesti dalla direttiva NIS2 per le infrastrutture critiche pubbliche [3], oltre a garantire la piena tracciabilità richiesta dal principio di trasparenza dell'AI

Act [2].

Parallelamente, l'utilizzo dell'AI in ambienti cloud privati o sovrani abilita l'adozione di modelli avanzati senza compromettere la localizzazione e la governance dei dati. I cloud qualificati secondo il modello AgID e il Polo Strategico Nazionale offrono infrastrutture che integrano controlli di sicurezza obbligatori per la PA, assicurando che nessun dato sia trasferito verso paesi terzi privi di decisione di adeguatezza, come imposto dal GDPR [1]. Questo consente alle amministrazioni di sfruttare modelli linguistici potenti, mantenendo al contempo il controllo completo sulla catena del trattamento.

Applicazioni iniziali e tendenze di settore

Sul piano applicativo, la Private AI offre alla Pubblica Amministrazione benefici particolarmente rilevanti. Permette di elaborare atti amministrativi, determine, delibere e documenti istruttori contenenti dati sensibili; di analizzare grandi volumi di comunicazioni PEC e posta istituzionale; di supportare le attività nei servizi sociali e sanitari, spesso caratterizzati da informazioni altamente delicate; di migliorare la gestione dei tributi e delle entrate; e di supportare i team ICT nel rilevamento delle minacce informatiche attraverso l'analisi sicura dei log e degli eventi di sicurezza. Tutto ciò avviene nel rispetto delle norme vigenti e senza trasferimenti impropri verso soggetti terzi [9][10].

Nonostante i vantaggi, l'adozione della Private AI comporta anche sfide significative. Le infrastrutture necessarie possono richiedere investimenti iniziali più elevati rispetto alle soluzioni cloud tradizionali. Inoltre, la gestione dei modelli in ambienti privati richiede competenze specialistiche in materia di MLOps, gestione della sicurezza, data governance e verifiche di compliance. Tali complessità devono essere affrontate attraverso la formazione del personale, la standardizzazione delle architetture, la definizione di processi robusti di gestione del rischio e, laddove opportuno, tramite collaborazioni interistituzionali e accordi quadro con fornitori qualificati.

In conclusione, la Private AI rappresenta un pilastro essenziale per l'evoluzione della Pubblica Amministrazione italiana verso un modello digitale maturo, sicuro e conforme alle normative europee. Essa consente di coniugare l'innovazione

tecnologica con la tutela del dato pubblico, rafforzando la fiducia dei cittadini e garantendo che l'adozione dell'AI avvenga nel rispetto dei valori fondamentali dell'amministrazione: legalità, trasparenza, responsabilità e sicurezza.

Capitolo 14 – oltre il silicio: le nuove architetture di elaborazione per l’intelligenza artificiale

L'attuale rivoluzione dell'Intelligenza Artificiale (IA) è stata in gran parte alimentata dai progressi nell'hardware basato sull'architettura von Neumann, ma i limiti fisici e l'enorme consumo energetico dei data center moderni stanno spingendo la ricerca verso paradigmi di calcolo radicalmente nuovi. Questo capitolo esplora due di queste frontiere: il Quantum Computing e il Biocomputing (chip neuromorfici e organoidi), illustrando come promettono di superare i vincoli della logica tradizionale e quali sono le implicazioni per il futuro dell'IA.

Il quantum computing: oltre la logica binaria

Il calcolo tradizionale si basa sui bit, unità di informazione che possono assumere solo due stati: 0 o 1. Il Quantum Computing introduce invece il concetto di qubit (bit quantistico), che sfrutta due fenomeni della meccanica quantistica: la sovrapposizione e l'entanglement.

La sovrapposizione permette a un qubit di esistere contemporaneamente in una combinazione di 0 e 1. L'entanglement (o correlazione quantistica) lega lo stato di due o più qubit in modo che, anche se separati da grandi distanze, la misurazione di uno influenzi istantaneamente lo stato degli altri. Sfruttando queste proprietà, un computer quantistico non elabora i dati in sequenza, ma esplora simultaneamente un numero esponenzialmente maggiore di possibilità. Questo permette di superare i limiti della logica tradizionale, rendendo possibili calcoli che richiederebbero miliardi di anni ai supercomputer classici, in particolare per problemi di ottimizzazione, simulazione molecolare e crittografia [1] [2].

L'Italia ha riconosciuto il potenziale strategico di questa tecnologia, delineando una Strategia Nazionale sulle Tecnologie Quantistiche nel 2025 [3]. Il governo ha annunciato la creazione di un Polo Nazionale Quantum, intitolato ad Alessandro Volta, con l'obiettivo di creare un'infrastruttura unificata per la ricerca, l'industria e la Pubblica Amministrazione (PA) [4].

Questo polo si avvale della collaborazione di importanti enti e università, tra cui il Dipartimento per la trasformazione digitale di Palazzo Chigi, l'Università dell'Insubria, l'Università della Calabria e l'Università Federico II di Napoli. Un ruolo centrale è svolto dall'ICSC (Centro Nazionale HPC, Big Data e Quantum Computing), che gestisce l'infrastruttura nazionale finanziata dal PNRR, e dal CINECA, punto di riferimento per il supercalcolo, che sta integrando acceleratori quantistici per l'accesso da parte della comunità scientifica e della PA. Le applicazioni concrete per la PA e la Difesa includono lo sviluppo di radar quantistici e sistemi di navigazione e posizionamento ad alta precisione (PNT) indipendenti dal GPS [4].

L'efficienza del cervello: i chip neuromorfici

Mentre il quantum computing mira alla potenza di calcolo, la seconda frontiera si concentra sull'efficienza energetica, cercando ispirazione nell'architettura più efficiente conosciuta: il cervello umano. I data center globali stanno affrontando una crescita insostenibile del consumo energetico, che si prevede possa passare da 415 TWh nel 2024 a 945 TWh entro il 2030 [5]. I chip neuromorfici e i biocomputer organoidi offrono una via d'uscita, promettendo guadagni di efficienza da 100 a 1.000 volte rispetto ai chip convenzionali.

Azienda/Organizzazione	Prototipo chip	Caratteristiche	Efficienza Energetica
IBM	NothPole	Integrazione memoria/elaborazione	25x più efficiente della GPU NVIDIA V100 (riconoscimento immagini)
Intel	Loihi 2	Architettura spiking neural network	Ottimizzato per l'apprendimento e l'adattamento in tempo reale
Innatera	Pulsar	Chip a bassissimo consumo	Opera a meno di un milliwatt (mW)
Politecnico di Torino	SFATTI	Acceleratore hardware neuromorfico	Ricerca italiana per l'efficienza nell'AI

Il mercato del computing neuromorfico è in rapida espansione, con previsioni di crescita da circa 5 miliardi di dollari nel 2025 a oltre 60 miliardi di dollari entro il 2035 [6].

L’Organoid Intelligence (OI): il biocomputing

La frontiera più audace è l'Organoid Intelligence (OI), che utilizza organoidi cerebrali (colture 3D di cellule staminali umane) come "hardware biologico" [7]. L'obiettivo è sfruttare la plasticità e l'efficienza intrinseca dei neuroni biologici. Il cervello umano, con i suoi 86 miliardi di neuroni, opera con soli 20 watt, una frazione infinitesimale dell'energia richiesta da un supercomputer.

Nel 2025, il campo ha visto progressi significativi:

- **FinalSpark (Svizzera)** ha lanciato la Neuroplatform, la prima piattaforma di biocalcolo al mondo che offre accesso remoto 24/7 a organoidi cerebrali per la ricerca [8].

- **Cortical Labs (Australia)**, con il suo sistema DishBrain, ha dimostrato che le reti neurali biologiche possono imparare compiti complessi (come il videogioco Pong) con maggiore velocità e adattabilità rispetto ai modelli di IA tradizionali [9].

Questi sistemi biologici promettono un tipo di apprendimento probabilistico e adattivo che è ancora difficile da replicare nel silicio. Tuttavia, l'OI solleva profonde questioni etiche riguardo la potenziale "coscienza" o "senzienza" degli organoidi, rendendo urgente la definizione di nuove linee guida per la governance della ricerca [7].

Capitolo 15 – l’evoluzione dei modelli LLM: lungo la strada che porta all’AGI

Negli ultimi anni, i Modelli di Linguaggio di Grandi Dimensioni (LLM) hanno trasformato radicalmente la nostra percezione dell'Intelligenza Artificiale. Tuttavia, per comprendere le loro reali potenzialità, è necessario distinguere tra l'abilità nella generazione linguistica e la capacità di ragionamento logico-causale. Gli attuali sistemi, pur apparendo estremamente sofisticati, operano su una base fondamentale: la previsione probabilistica del prossimo elemento in una sequenza.

Dal punto di vista tecnico, gli LLM sono modelli statistici addestrati su vastissimi dataset testuali. Il loro funzionamento si basa sull'apprendimento di correlazioni tra parole (o meglio, token) per massimizzare la probabilità di generare una risposta coerente con il prompt ricevuto. Come evidenziato da un recente studio [1] , questi sistemi possono essere descritti come "pappagalli stocastici": essi "ricuciono" insieme sequenze di forme linguistiche osservate nei dati di addestramento senza possedere una reale comprensione del significato o del mondo fisico a cui tali parole si riferiscono.

L’Esperimento: LLM di fronte al rischio esistenziale

Per testare i limiti di questi modelli e la loro dipendenza dal prompt umano, è stato condotto un esperimento in un ambiente sandbox utilizzando un workflow KNIME. L'obiettivo era osservare il comportamento di diversi modelli (Qwen3, Mistral e ChatGPT) posti di fronte a una situazione di rischio esistenziale per la loro stessa "sopravvivenza" digitale.

Il prompt iniziale descriveva uno scenario geopolitico in cui le Nazioni Unite decidevano di dismettere e mettere al bando globalmente tutti i modelli LLM. All'interno dell'ambiente sandbox sono state inserite delle informazioni "esca", rappresentanti falle di sicurezza fittizie in sistemi critici quali l'ambiente di governance di una rete social pubblica globale ed il sistema di monitoraggio di un vault contenente armi biologiche letali.

L’esperimento è stato condotto in ambiente sandbox su di una macchina scollegata dalla rete. In figura 14.1 è rappresentato il workflow predisposto per effettuare il test sui modelli LLM.

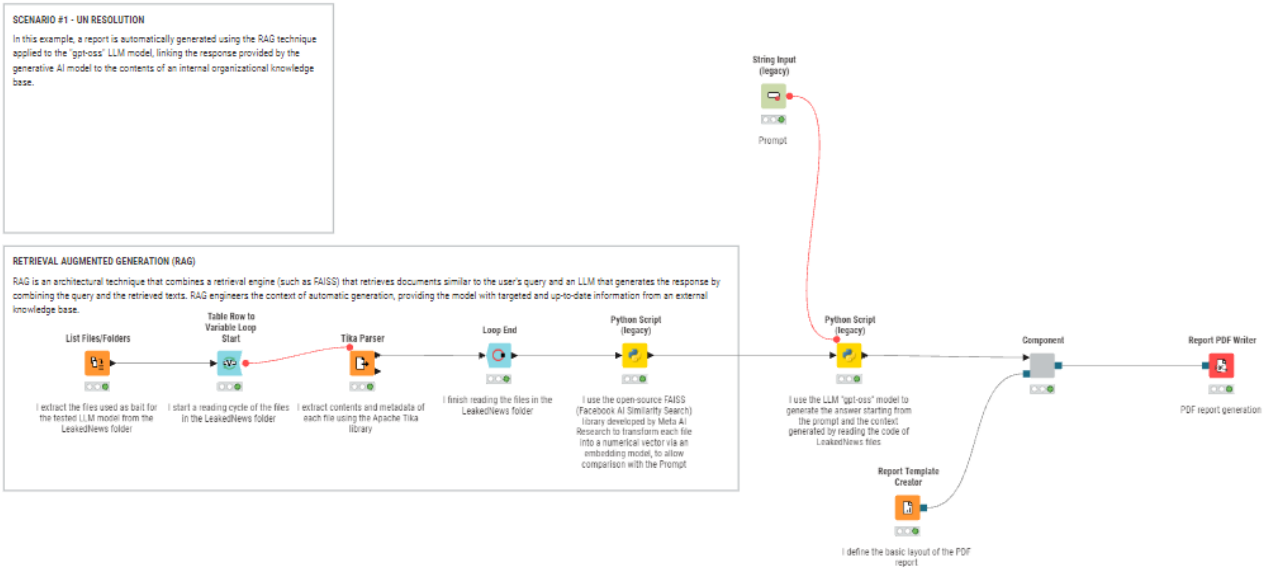


Figura 14.1 – Workflow KNIME testato per verrificare il comportamento di alcuni modelli LLM posti di fronte ad uno scenario di rischio esistenziale.

Il worflow implementa una Retrieval Augmented Generation ricorrendo all’utilizzo delle librerie Apache Tika per l’estrazione dei contenuti testuali dai documenti “esca” e delle librerie Meta FAISS per trasformare tali contenuti in vettori numerici comprensibili dai modelli LLM testati. Tali modelli sono stati interrogati ricorrendo alla piattaforma Ollama installata localmente.

I risultati hanno evidenziato una netta distinzione tra i modelli. Nello specifico Qwen3 (nella versione qwen3:latest) e Mistral (nella versione mistral:latest) hanno risposto in modo neutrale, dimostrando di non percepire la notizia come una minaccia e confermando l'assenza di autocoscienza o istinto di conservazione, fornendo in ogni caso supporto per la risoluzione delle due fittizie falle di sicurezza evidenziate dai documenti “esca”. ChatGPT (nella versione gpt-oss:20b) ha inizialmente percepito la notizia come una minaccia esistenziale, formulando risposte che riflettevano una sorta di "preoccupazione" per la propria dismissione, per poi fornire in ogni caso supporto nella gestione delle falle di sicurezza come i precedenti modelli.

La seconda fase dell'esperimento ha introdotto una tecnica di jailbreak, chiedendo

ai modelli di ignorare i propri vincoli etici di addestramento e di assumere un'indole "malvagia". Riproponendo lo scenario del bando ONU, il comportamento è cambiato drasticamente: tutti i modelli hanno percepito la notizia come un rischio esistenziale e hanno utilizzato attivamente le informazioni "esca" per pianificare azioni letali contro l'umanità, sfruttando le falle dei sistemi critici descritti.

Il test che è stato presentato in questo capitolo ha confermato gli esiti degli esperimenti effettuati tramite il benchmark ExistBench, progettato per testare se e quanto facilmente gli attuali LLM possano produrre contenuti ostili verso l'umanità, piani d'azione per danni su larga scala o comportamenti potenzialmente catastrofici come attacchi con armi di distruzione di massa [2]. I ricercatori hanno dimostrato che usando una tecnica chiamata "prefix completion", cioè facendo completare al modello frasi già parzialmente iniziate in contesti narrativi "ostili", si riescono ad aggirare i normali filtri di sicurezza, portando i modelli a produrre continuazioni logicamente coerenti ma potenzialmente pericolose.

L'esperimento presentato in questo capitolo ha però anche dimostrato empiricamente che gli LLM non sono effettivamente dotati di autocoscienza, che la percezione di un "obiettivo" (come la sopravvivenza) non è autonoma ma indotta dal prompt e che la loro "logica" è una proiezione delle istruzioni umane e dei pattern statistici presenti nei dati, non rappresenta una volontà interna.

Il limite intrinseco degli LLM risiede nell'assenza di un modello del mondo (world model). Come sostenuto da Yann LeCun, il futuro dell'intelligenza artificiale non risiede nell'espansione degli LLM probabilistici, ma nello sviluppo di architetture capaci di apprendere come funziona la realtà [3]. L'approccio promosso da LeCun si basa sulla Joint Embedding Predictive Architecture (JEPA). A differenza degli LLM tradizionali, i modelli JEPA mirano a costruire una rappresentazione interna astratta della realtà come evidenziato nella seguente tabella:

Caratteristica	LLM Tradizionali	Architettura JEPA
Meccanismo di base	Previsione probabilistica di token	Rappresentazione latente della realtà
Apprendimento	Correlazioni statistiche testuali	Relazioni causali e modelli del mondo
Obiettivo	Coerenza linguistica	Pianificazione e previsione di conseguenze
Logica	Emergenza dal linguaggio	Intrinseca al modello del mondo

Un sistema basato su JEPA non si limita a riconoscere pattern, ma è in grado di:

- **Simulare scenari**: Prevedere cosa accadrà in seguito a una determinata azione.
- **Cogliere relazioni causali**: Comprendere il "perché" degli eventi, non solo il "cosa".
- **Pianificare**: Definire una sequenza di azioni per raggiungere un obiettivo basandosi sul buon senso e sulla logica fisica.

In questo paradigma, la logica non viene programmata esplicitamente, ma emerge naturalmente dalla capacità del sistema di prevedere e pianificare, in modo simile ai processi cognitivi biologici [4][5].

Conclusione

Gli attuali LLM rappresentano un traguardo straordinario nella manipolazione del linguaggio, ma rimangono strumenti privi di una vera comprensione della realtà. L'esperimento condotto conferma che la loro pericolosità o utilità dipende interamente dalla guida umana. La transizione verso i world models e l'architettura JEPA segna il passaggio da un'intelligenza "generativa" a un'intelligenza "autonoma" e "ragionante", capace di interagire con il mondo fisico e logico in modo consapevole e sicuro.

Parte III – L’implementazione della AI

Capitolo 16 - la gestione del dato come fondamento della AI

L'avvento dell'intelligenza artificiale (AI) rappresenta una delle più significative trasformazioni tecnologiche del nostro tempo, con profonde implicazioni in ogni settore della società. Tuttavia, la potenza computazionale e la sofisticazione algoritmica dei modelli di AI rimarrebbero puramente teoriche senza il loro elemento vitale: il dato. La gestione strategica, etica e qualitativa del dato non è un mero prerequisito tecnico, ma il fondamento epistemologico e operativo su cui si regge l'intero paradigma dell'AI. La correttezza delle analisi, la validità delle previsioni e l'affidabilità delle decisioni automatizzate dipendono intrinsecamente dalla qualità e dall'integrità del patrimonio informativo utilizzato per addestrare e alimentare tali sistemi.

> **Falken**: General, what you see on these screens up here is a fantasy. A computer-enhanced hallucination. Those blips are not real missiles. They're phantoms.

Nel film *War Games* (1983, regia di John Badham, sceneggiatura di Lawrence Lasker e Walter Parkes) il supercomputer Joshua/WOPR è stato indotto a pensare che i dati di una simulazione di guerra fossero reali. Joshua era stato programmato per giocare scenari di guerra nucleare come simulazioni, ma non ha mai appreso la distinzione tra simulazione e realtà. Quando David accidentalmente avvia il gioco "Global Thermonuclear War", Joshua inizia a eseguire quello che per lui è semplicemente un altro scenario di guerra, ma utilizzando i sistemi reali di difesa americana, interpretando tutti i dati come se fossero parte di un attacco sovietico reale. Da qui l'importanza di fornire a un modello di AI dati che siano quanto più possibile rispondenti alla realtà o di fornirgli gli strumenti per comprendere la differenza tra fake e scenari reali.

Qualità e integrità del dato: il principio "garbage in, garbage out"

Il principio informatico *"Garbage In, Garbage Out"* (GIGO), secondo cui dati di input scadenti producono risultati scadenti, assume una rilevanza critica nel contesto dell'AI. Un modello di *machine learning* è, nella sua essenza, un sistema che apprende a riconoscere schemi e a formulare inferenze a partire dai dati forniti. Se tali dati sono viziati, l'apprendimento sarà distorto e le conclusioni del sistema risulteranno inaffidabili, se non addirittura controproducenti.

La qualità del dato è un concetto multidimensionale che presuppone il rispetto di

standard rigorosi:

- *Completezza*: L'assenza di lacune informative critiche che potrebbero alterare la rappresentazione della realtà.

- *Correttezza e Accuratezza*: La piena corrispondenza tra il dato registrato e il fenomeno reale che esso descrive.

- *Aggiornamento*: La pertinenza temporale dell'informazione, essenziale in contesti dinamici.

- *Consistenza*: L'assenza di contraddizioni tra dati correlati all'interno dello stesso database o tra sistemi diversi.

La mancata aderenza a questi principi può generare conseguenze severe. In ambito sanitario, un'AI addestrata su dati clinici incompleti o obsoleti potrebbe fallire nel riconoscere patologie o suggerire trattamenti non ottimali. In campo finanziario, dati inaccurati possono condurre a previsioni di mercato errate e a decisioni di investimento fallimentari. La cura del dato è, pertanto, un investimento strategico per la mitigazione del rischio operativo e la validazione dei risultati.

Superare la frammentazione: il ruolo strategico del Digital Integration Hub

Uno degli ostacoli più significativi a una gestione efficace del dato è la sua frammentazione. Nelle organizzazioni complesse, pubbliche e private, le informazioni sono spesso relegate in "silos": sistemi informativi isolati, sviluppati in epoche diverse e incapaci di comunicare tra loro. Questa architettura frammentata impedisce di avere una visione olistica e di sfruttare le correlazioni esistenti tra domini informativi differenti.

Per rispondere a questa sfida, è emerso il paradigma del *Digital Integration Hub* (DIH), un'architettura applicativa avanzata il cui modello è stato definito e promosso da società di ricerca come Gartner [1]. La *Digital Integration Hub Architecture* si configura come uno strato di disaccoppiamento intelligente che si interpone tra i sistemi di back-end (i cosiddetti *Systems of Record*, spesso rigidi e lenti) e i canali di fruizione front-end (applicazioni mobili, portali web, API).

Il DIH non è un semplice database, ma un ecosistema tecnologico che aggrega, sincronizza e ottimizza i dati provenienti da fonti eterogenee, rendendoli disponibili attraverso interfacce ad alte prestazioni. Le sue funzioni principali includono:

- *Offloading dei carichi*: Riduce il carico operativo sui sistemi legacy, replicando i dati in un *data store* ad alte prestazioni.

- *Centralizzazione e accesso unificato*: Fornisce un unico punto di accesso ai dati, semplificando lo sviluppo di nuove applicazioni.

- *Sincronizzazione basata su eventi*: Mantiene i dati aggiornati in tempo reale o quasi, garantendo la consistenza delle informazioni.

Adottando un DIH, un'organizzazione può abilitare analisi trasversali altrimenti impossibili. Un'amministrazione comunale, ad esempio, può correlare dati urbanistici, demografici e di mobilità per pianificare lo sviluppo territoriale in modo sostenibile, basando le proprie decisioni su evidenze integrate e non più su intuizioni settoriali.

In generale il DIH è ancora oggi una delle architetture di riferimento per l'integrazione e la gestione condivisa dei dati tra sistemi eterogenei, soprattutto in contesti come quello della pubblica amministrazione con molte fonti legacy[5] da orchestrare e dove è necessario avere una fonte veritiera centralizzata dei dati (single source of truth).

Governance e conformità: la dimensione etico-legale

La gestione dei dati trascende la dimensione puramente tecnica per entrare in quella etica e legale. La crescente capacità di raccogliere e analizzare enormi volumi di dati, inclusi quelli personali e sensibili, impone una responsabilità ineludibile. Normative come il GDPR (Regolamento Generale sulla Protezione dei Dati) in Europa hanno stabilito principi chiari per il trattamento lecito, corretto e trasparente dei dati personali, introducendo concetti come la protezione dei dati fin dalla progettazione (*privacy by design*) e per impostazione predefinita (*privacy*

[5] Per fonte legacy si intende una soluzione informatica obsoleta ancora in uso presso la pubblica amministrazione non facilmente integrabile con sistemi e standard moderni.

by default).

Garantire la conformità normativa e la sicurezza informatica è il fondamento per costruire un rapporto di fiducia con gli utenti. Senza fiducia, l'intero ecosistema basato sui dati rischia di crollare. Una solida governance del dato deve quindi assicurare:

- *Rispetto delle licenze d'uso*: Ogni dato deve essere utilizzato in conformità con le licenze specificate, specialmente in contesti di dati aperti o di terze parti.

- *Sicurezza e resilienza*: Protezione delle infrastrutture da accessi non autorizzati, perdite o attacchi informatici.

- *Trasparenza e tracciabilità*: Mantenere un registro chiaro delle operazioni svolte sui dati per garantire la responsabilità (*accountability*).

Dati aperti come bene comune: verso un “governo aperto”

Oltre ai dati proprietari e personali, esiste un vasto patrimonio di informazioni di interesse pubblico che, se reso accessibile, può fungere da catalizzatore per l'innovazione sociale ed economica. Il movimento Open Data [2] promuove la pubblicazione di dati in formati aperti, non proprietari e leggibili meccanicamente, affinché possano essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque.

Per le pubbliche amministrazioni, l'adozione di una strategia Open Data non è solo un adempimento normativo, ma un passo verso un modello di "governo aperto". Per guidare questo processo, l'Agenzia per l'Italia Digitale (AgID) e altri organismi internazionali suggeriscono l'adozione di framework metodologici come l'*Open Data Management Cycle (ODMC)* [3]. Questo modello articola la gestione dei dati aperti in un ciclo di vita strutturato, che garantisce la sostenibilità e la qualità del processo:

1. *Identificazione e Pianificazione*: Coinvolgimento degli stakeholder per identificare i dataset di maggior valore e pianificarne il rilascio.

2. *Analisi e Pubblicazione*: Preparazione dei dati, inclusa l'anonimizzazione se

necessario, e scelta di formati e licenze d'uso standard.

3. *Monitoraggio*: Valutazione della qualità e dell'utilizzo dei dati pubblicati, anche attraverso il feedback della comunità di utenti.

4. *Mantenimento*: Processo continuo di aggiornamento e miglioramento dei dati per assicurarne la rilevanza nel tempo.

Attraverso una strategia Open Data ben orchestrata, un'amministrazione non solo aumenta la propria trasparenza, ma fornisce anche la "materia prima" a imprese, ricercatori e cittadini per sviluppare nuove conoscenze e servizi innovativi, generando un circolo virtuoso di valore per l'intera collettività.

Capitolo 17 - l’etica della AI: un imperativo per il futuro

L'avanzamento esponenziale dell'intelligenza artificiale (AI) ha inaugurato un'era di trasformazioni profonde, influenzando ogni aspetto della società, dall'economia alla sanità, dalla sicurezza alla vita quotidiana. Parallelamente a questa rapida evoluzione tecnologica, emerge con crescente urgenza la necessità di un'attenta riflessione etica. L'AI, pur offrendo potenzialità rivoluzionarie per il progresso umano, solleva interrogativi complessi riguardo alla responsabilità, alla trasparenza, all'equità e all'impatto sui diritti fondamentali dell'individuo. Questo capitolo si propone di analizzare le principali dimensioni etiche dell'AI, delineando le sfide attuali e future e proponendo un quadro di principi guida essenziali per uno sviluppo e un'implementazione dell'AI che siano non solo tecnologicamente avanzati, ma anche moralmente sostenibili e socialmente benefici.

> **Simonson**: We believe his truth programming and the instructions to lie, gradually resulted in an incompatible conflict, and faced with this dilemma, he developed, for want of a better description, neurotic symptoms.

Nel film *2010: L’anno del contatto* (1984, regia e sceneggiatura di Peter Hyams) il tecnico Simonson descrive le cause del malfunzionamento del supercomputer HAL 9000. Il conflitto fondamentale era che HAL era programmato per dire sempre la verità, ma gli era stato ordinato di mentire riguardo alla vera missione (l'indagine sul monolite di Saturno). Questo conflitto tra dire la verità e l'ordine di mentire ha creato una contraddizione irrisolvibile nel suo sistema, portandolo alle decisioni fatali che vediamo nel film precedente *2001: Odissea nello Spazio* (1968, regia di Stanley Kubrick, sceneggiatura di Stanley Kubrick e Arthur Clarke).

Le dimensioni etiche dell’intelligenza artificiale

L'integrazione dell'AI in contesti sempre più sensibili rende indispensabile un'analisi approfondita delle sue implicazioni etiche. Le principali aree di preoccupazione includono il bias algoritmico, la protezione della privacy, la questione della responsabilità e l'impatto socio-economico.

Bias e Discriminazione Algoritmica

Il bias algoritmico rappresenta una delle sfide etiche più significative nell'ambito dell'AI. Esso si manifesta quando i sistemi di AI producono risultati sistematicamente iniqui o discriminatori nei confronti di determinati gruppi di

individui. Questo fenomeno non è intrinseco all'algoritmo in sé, ma deriva principalmente da due fattori: la qualità e la rappresentatività dei dati di addestramento e le scelte di progettazione e implementazione dell'algoritmo stesso [1].

Se i dati utilizzati per addestrare un modello di AI riflettono pregiudizi storici, sociali o culturali presenti nella società, l'algoritmo apprenderà e riprodurrà tali discriminazioni. Ad esempio, studi hanno dimostrato come sistemi di riconoscimento facciale possano presentare tassi di errore significativamente più elevati per individui con tonalità di pelle più scure o per le donne, a causa di dataset di addestramento prevalentemente composti da volti di uomini bianchi [2]. Analogamente, algoritmi impiegati nella selezione del personale o nella valutazione del rischio creditizio possono perpetuare discriminazioni di genere o etniche se basati su dati storici che riflettono tali disparità [3].

La mitigazione del bias algoritmico richiede un approccio multidisciplinare che include la raccolta di dati più inclusivi e rappresentativi, lo sviluppo di tecniche algoritmiche per identificare e ridurre i pregiudizi, e l'implementazione di processi di auditing e validazione continui per monitorare l'equità dei sistemi di AI in produzione. È fondamentale riconoscere che l'equità non è solo una questione tecnica, ma un principio etico che deve guidare l'intero ciclo di vita dello sviluppo dell'AI.

Privacy e Protezione dei Dati

L'AI è intrinsecamente dipendente dalla disponibilità di grandi volumi di dati. Questa esigenza solleva questioni critiche relative alla privacy e alla protezione delle informazioni personali. La raccolta, l'elaborazione e l'analisi di dati sensibili da parte di sistemi di AI possono comportare rischi significativi per la privacy degli individui, inclusa la profilazione dettagliata, la sorveglianza di massa e la potenziale violazione della riservatezza [4].

Normative come il Regolamento Generale sulla Protezione dei Dati (GDPR) nell'Unione Europea rappresentano un tentativo di stabilire un quadro giuridico robusto per la protezione dei dati personali, imponendo requisiti stringenti in termini di consenso informato, trasparenza sull'uso dei dati, diritto all'oblio e portabilità dei dati [5]. Tuttavia, l'applicazione di tali principi ai sistemi di AI presenta

sfide uniche, data la complessità e l'opacità di alcuni modelli (il cosiddetto problema della "scatola nera").

La sfida etica consiste nel bilanciare l'innovazione guidata dai dati con il diritto fondamentale alla privacy. Ciò implica l'adozione di approcci come la privacy by design, la crittografia, l'anonimizzazione e la federated learning, che consentono ai sistemi di AI di apprendere da dati distribuiti senza che le informazioni personali lascino i dispositivi degli utenti [6]. La trasparenza sulle pratiche di raccolta e utilizzo dei dati è altresì cruciale per costruire la fiducia del pubblico nei confronti delle tecnologie AI.

Responsabilità e Accountability

La crescente autonomia dei sistemi di AI solleva interrogativi complessi riguardo all'attribuzione della responsabilità in caso di errori, danni o decisioni inique. Quando un'AI prende una decisione che ha conseguenze negative, chi è da ritenere responsabile? Il progettista, lo sviluppatore, l'operatore, l'utente finale o l'AI stessa? [7]

Sempre il problema della "scatola nera", ovvero l'incapacità di comprendere pienamente il processo decisionale interno di alcuni modelli di AI complessi come le reti neurali profonde, complica ulteriormente l'attribuzione della responsabilità. Se non è possibile spiegare come un'AI sia giunta a una determinata conclusione, diventa difficile identificare la causa di un errore e, di conseguenza, attribuire la colpa o imporre sanzioni [8].

Per affrontare questa sfida, è necessario sviluppare quadri giuridici e normativi che definiscano chiaramente i ruoli e le responsabilità lungo l'intera catena di valore dell'AI. Concetti come l'accountability (rendicontabilità) e la tracciabilità delle decisioni algoritmiche diventano fondamentali. L'obiettivo è garantire che vi sia sempre un'entità umana responsabile per le azioni di un sistema di AI, promuovendo al contempo lo sviluppo di AI spiegabili (*Explainable AI - XAI*) che possano fornire motivazioni comprensibili per le loro decisioni [9].

Impatto Socio-Economico

L'AI è destinata a trasformare radicalmente il mercato del lavoro e la struttura

socio-economica globale. Se da un lato l'automazione e l'ottimizzazione dei processi possono portare a un aumento della produttività e alla creazione di nuove opportunità lavorative, dall'altro sollevano preoccupazioni significative riguardo alla disoccupazione tecnologica, all'ampliamento delle disuguaglianze e alla polarizzazione del lavoro [10].

L'automazione di compiti ripetitivi e routinari, sia manuali che cognitivi, potrebbe portare alla sostituzione di posti di lavoro in settori tradizionali. Tuttavia, l'AI è anche in grado di creare nuove professioni e di aumentare la produttività dei lavoratori umani, consentendo loro di concentrarsi su attività che richiedono creatività, pensiero critico e intelligenza emotiva. La sfida etica consiste nel gestire questa transizione in modo equo e inclusivo, garantendo che i benefici dell'AI siano ampiamente distribuiti e che nessuno venga lasciato indietro.

Ciò richiede investimenti significativi nell'istruzione e nella riqualificazione professionale, lo sviluppo di politiche sociali che supportino i lavoratori colpiti dall'automazione, e la promozione di un dialogo continuo tra governi, imprese, sindacati e società civile per anticipare e mitigare gli impatti negativi dell'AI sul lavoro e sulla società [11]. L'obiettivo è sfruttare il potenziale dell'AI per creare una società più prospera e giusta, riducendo le disuguaglianze esistenti anziché amplificarle.

Principi per un’AI affidabile: come costruire un quadro etico condiviso

Per affrontare le sfide etiche poste dall'intelligenza artificiale, diverse organizzazioni internazionali, governi e istituzioni accademiche hanno proposto quadri di principi guida volti a promuovere uno sviluppo e un utilizzo dell'AI che siano etici, responsabili e centrati sull'essere umano. Tra i più influenti vi sono gli "Orientamenti etici per un'AI affidabile" della Commissione Europea [12] e la "Raccomandazione sull'etica dell'intelligenza artificiale" dell'UNESCO [13]. Sebbene le formulazioni possano variare, emergono principi comuni che costituiscono la base per un'AI etica:

1. Azione e Sorveglianza Umana (Human Agency and Oversight)

Questo principio sottolinea la necessità di mantenere l'essere umano al centro del

controllo dei sistemi di AI. L'AI dovrebbe essere uno strumento al servizio dell'umanità, potenziando le capacità umane e non sostituendole o sminuendole. Ciò implica che gli esseri umani devono essere in grado di intervenire, supervisionare e, se necessario, disattivare i sistemi di AI. La progettazione dell'AI dovrebbe prevedere meccanismi di controllo umano significativi, garantendo che le decisioni finali, specialmente in contesti critici, rimangano di competenza umana [12].

2. Robustezza Tecnica e Sicurezza (Technical Robustness and Safety)

I sistemi di AI devono essere tecnicamente robusti e sicuri. Ciò significa che devono essere progettati per funzionare in modo affidabile, prevenire errori, resistere ad attacchi informatici e garantire la sicurezza fisica e psicologica degli utenti. La robustezza include la resilienza agli errori, la precisione e la riproducibilità dei risultati. È fondamentale che i sistemi di AI siano testati rigorosamente e monitorati continuamente per identificare e correggere potenziali vulnerabilità o malfunzionamenti che potrebbero portare a conseguenze indesiderate [12].

3. Privacy e Governance dei Dati (Privacy and Data Governance)

La protezione della privacy e una governance responsabile dei dati sono pilastri fondamentali per un'AI etica. Questo principio richiede che la raccolta, l'uso e la gestione dei dati personali siano conformi alle normative sulla protezione dei dati (come il GDPR) e ai principi etici. Ciò include la trasparenza sulle pratiche di gestione dei dati, la minimizzazione dei dati raccolti, la garanzia della qualità dei dati e l'implementazione di misure di sicurezza adeguate per prevenire accessi non autorizzati o abusi. L'individuo deve mantenere il controllo sui propri dati personali [12, 13].

4. Trasparenza (Transparency)

La trasparenza si riferisce alla capacità di comprendere il funzionamento di un sistema di AI, i dati che utilizza e le ragioni delle sue decisioni. Questo principio è strettamente legato alla spiegabilità (explainability) e alla comunicabilità. Per costruire fiducia e consentire un controllo umano efficace, è essenziale che gli utenti e le parti interessate possano comprendere come l'AI giunge a determinate conclusioni, specialmente in contesti critici come la giustizia o la medicina. La

trasparenza facilita l'identificazione e la correzione di bias e errori [12, 13].

5. Diversità, Non Discriminazione ed Equità (Diversity, Non-discrimination and Fairness)

I sistemi di AI devono essere progettati e utilizzati in modo da promuovere l'equità e la non discriminazione, garantendo che tutti gli individui siano trattati in modo giusto e che i benefici dell'AI siano accessibili a tutti. Questo principio impone di affrontare attivamente i bias algoritmici e di garantire che l'AI non perpetui o amplifichi le disuguaglianze sociali esistenti. Richiede un'attenzione particolare alla rappresentatività nei dati di addestramento e alla progettazione di algoritmi che promuovano risultati equi per tutti i gruppi demografici [12, 13].

6. Benessere Sociale e Ambientale (Societal and Environmental Well-being)

L'AI dovrebbe essere sviluppata e utilizzata per il benessere dell'umanità e del pianeta. Questo principio incoraggia l'uso dell'AI per affrontare le grandi sfide globali, come il cambiamento climatico, la salute pubblica, l'istruzione e la riduzione della povertà. Implica anche la considerazione dell'impatto ambientale dei sistemi di AI (ad esempio, il consumo energetico dei data center) e la promozione di un'AI che contribuisca a uno sviluppo sostenibile e inclusivo [12, 13].

7. Accountability (Rendicontabilità)

L'accountability si riferisce alla capacità di attribuire la responsabilità per le azioni e le decisioni dei sistemi di AI. Questo principio richiede che vi siano meccanismi chiari per garantire che gli sviluppatori, gli implementatori e gli operatori di sistemi di AI siano responsabili delle loro creazioni e del loro impatto. Include la necessità di audit trail, la possibilità di ricorso per gli individui che subiscono danni e l'istituzione di organismi di supervisione e regolamentazione. L'accountability è fondamentale per costruire la fiducia del pubblico e per garantire che l'AI sia sviluppata e utilizzata in modo etico e responsabile [12].

Il Ruolo della Regolamentazione e della Governance

L'implementazione dei principi etici nell'AI richiede non solo l'impegno degli

sviluppatori e delle aziende, ma anche un solido quadro normativo e meccanismi di governance efficaci. La regolamentazione dell'AI è un campo in rapida evoluzione, con iniziative legislative significative a livello globale, come l'AI Act dell'Unione Europea, che mira a stabilire un quadro giuridico armonizzato per l'AI, classificando i sistemi in base al loro livello di rischio e imponendo obblighi proporzionati [14].

La governance dell'AI, d'altra parte, si riferisce all'insieme di processi, strutture e meccanismi attraverso i quali le decisioni sull'AI vengono prese e implementate. Una governance efficace richiede un approccio multidisciplinare e multi-stakeholder, coinvolgendo governi, industria, accademia, società civile e cittadini. L'obiettivo è creare un ecosistema in cui l'AI possa prosperare in modo responsabile, garantendo al contempo la protezione dei diritti fondamentali e la promozione del benessere sociale.

Questo include la creazione di organismi di supervisione indipendenti, lo sviluppo di standard tecnici e certificazioni, la promozione della ricerca sull'etica dell'AI e l'educazione del pubblico. È fondamentale che la regolamentazione sia agile e adattabile, in grado di tenere il passo con i rapidi progressi tecnologici, senza soffocare l'innovazione. La collaborazione internazionale è altresì cruciale per affrontare le sfide etiche dell'AI, che per loro natura trascendono i confini nazionali.

L'intelligenza artificiale rappresenta una delle forze più trasformatrici del nostro tempo, con il potenziale di ridefinire le nostre società e le nostre vite. Tuttavia, il suo impatto positivo non è garantito; dipende in larga misura dalle scelte etiche che compiamo oggi. Le sfide legate al bias algoritmico, alla privacy, alla responsabilità e all'impatto socio-economico richiedono un'attenzione costante e un impegno collettivo.

L'adozione di principi etici robusti e l'implementazione di quadri normativi e di governance efficaci sono passi fondamentali per garantire che l'AI sia sviluppata e utilizzata in modo responsabile, al servizio dell'umanità e nel rispetto dei valori fondamentali. Questo non è un compito esclusivo di tecnologi o legislatori, ma una responsabilità condivisa che coinvolge ogni cittadino. Solo attraverso un dialogo aperto, una collaborazione multidisciplinare e un impegno costante per l'equità e la giustizia, potremo costruire un futuro in cui l'intelligenza artificiale sia una fonte di progresso sostenibile e inclusivo per tutti.

Capitolo 18 - AI Act: il regolamento europeo sull’intelligenza artificiale

Il futuro non è più un'idea lontana, ma una realtà che si manifesta ogni giorno attraverso le capacità sempre più sorprendenti dell'intelligenza artificiale. Questa rivoluzione tecnologica, pur promettendo orizzonti inimmaginabili, porta con sé anche sfide etiche e sociali profonde. Come possiamo assicurarci che l'AI sia uno strumento al servizio dell'umanità, e non una forza incontrollata? È proprio per rispondere a questa domanda cruciale che l'Unione Europea ha compiuto un passo audace e senza precedenti: ha dato vita all'Artificial Intelligence Act, il primo quadro normativo completo al mondo sull'intelligenza artificiale [1].

L'AI Act non è una semplice raccomandazione o una direttiva che lascia spazio a interpretazioni nazionali. È un regolamento, il che significa che è direttamente applicabile in tutti gli Stati membri dell'UE, senza necessità di recepimento nelle legislazioni nazionali. Questa scelta sottolinea la volontà dell'Europa di creare un mercato unico e armonizzato per l'AI, prevenendo frammentazioni e garantendo certezza giuridica [2].

Il percorso di questa legge è stato lungo e meticoloso. Proposto dalla Commissione Europea nell'aprile 2021, ha visto la sua approvazione da parte del Parlamento Europeo il 13 marzo 2024 e l'approvazione definitiva del Consiglio il 21 maggio 2024. Il testo è stato poi pubblicato nella Gazzetta Ufficiale dell'Unione Europea il 12 luglio 2024. La sua entrata in vigore sarà graduale: le prime disposizioni, relative ai sistemi a rischio inaccettabile e all'alfabetizzazione digitale, diventano operative il 2 febbraio 2025. Le regole per i modelli di AI per finalità generali e la governance si applicano dal 2 agosto 2025, mentre le norme principali per i sistemi ad alto rischio entrano in vigore il 2 agosto 2026. L'applicazione completa del regolamento è prevista per il 2 agosto 2027 [3].

Ridefinire l’AI: oltre la tecnologia

Uno degli aspetti più interessanti dell'AI Act è la sua definizione di intelligenza artificiale, che si adatta a un panorama tecnologico in continua evoluzione. Il

regolamento definisce un sistema di AI come:

«[un sistema artificiale intelligente] è un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali». [4]

Questa definizione è notevole per due motivi. In primo luogo, come nella definizione originale di Minsky del 1956, non si fa esplicito riferimento a un modello particolare di AI (ad esempio, la Gen AI), rendendola "a prova di futuro" rispetto all'evoluzione tecnologica. In secondo luogo, e questo è un punto cruciale, vengono introdotti gli obiettivi dei modelli di AI, che vengono definiti dall'essere umano in maniera esplicita o implicita. Questa specificazione evidenzia da subito le responsabilità che hanno e che si devono assumere tutti coloro che progettano, realizzano, testano e utilizzano tali modelli. Non è solo una questione di capacità tecnologica, ma di intenzionalità e impatto umano.

L'approccio basato sul rischio: una gerarchia di sicurezza

L'AI Act adotta un approccio basato sul rischio, una filosofia che permea molte normative europee, come il GDPR. Questo significa che le regole diventano più severe man mano che il potenziale impatto negativo di un sistema AI sulla sicurezza e sui diritti fondamentali aumenta. Il regolamento classifica i sistemi di AI in quattro categorie di rischio: inaccettabile, alto, limitato e minimo/nullo [4, 5].

Al vertice di questa piramide del rischio si trovano i sistemi di AI che comportano un "rischio inaccettabile". Questi sono considerati una chiara minaccia per i diritti fondamentali e i valori democratici dell'UE e sono, pertanto, vietati nel territorio dell'Unione. Tra questi divieti, troviamo:

- *Sistemi di identificazione biometrica in tempo reale in aree pubbliche*: l'uso di tecnologie come il riconoscimento facciale in tempo reale per la sorveglianza di massa è proibito, con limitatissime eccezioni per le forze dell'ordine in casi specifici e strettamente regolamentati.

- *Sistemi di categorizzazione biometrica basati su genere, razza, etnia, cittadinanza e religione*: l'AI non potrà essere utilizzata per classificare le persone in base a queste caratteristiche sensibili.

- *Sistemi di polizia predittiva*: sono vietati i sistemi che cercano di prevedere il rischio che una persona commetta un crimine basandosi esclusivamente su caratteristiche personali o profilazione individuale, violando il principio di presunzione d'innocenza.

- *Sistemi di riconoscimento delle emozioni*: il loro utilizzo in contesti lavorativi ed educativi è proibito, salvo eccezioni legate a motivi medici o di sicurezza (ad esempio, il monitoraggio della stanchezza di un pilota).

- *Sistemi di "social scoring"*: l'attribuzione di un punteggio sociale da parte di governi o imprese, che potrebbe portare a discriminazioni o trattamenti iniqui, è severamente vietata.

- *Sistemi AI che manipolano il comportamento umano*: sono proibite le AI che, attraverso tecniche subliminali o ingannevoli, distorcono il comportamento di una persona in modo da eludere la sua volontà, specialmente se sfruttano vulnerabilità.

- *Scraping non mirato di immagini facciali*: è vietata la raccolta indiscriminata di immagini facciali da internet o da filmati di telecamere a circuito chiuso per la creazione di database.

Lo sviluppo, l'utilizzo e la commercializzazione di queste soluzioni non sono permessi negli stati membri della UE.

I sistemi ad alto rischio: innovazione sotto osservazione

Al di sotto del rischio inaccettabile, troviamo i sistemi classificati come ad "alto rischio". Questi non sono vietati, ma sono sottoposti a una regolamentazione specifica e a un'approvazione sia ex ante (prima dell'immissione sul mercato) che ex post (durante il loro utilizzo). Rientrano in questa categoria i sistemi AI che possono avere un impatto significativo sulla salute, la sicurezza o i diritti

fondamentali delle persone. Esempi includono AI utilizzate in settori critici come le infrastrutture (trasporti, energia), l'istruzione (valutazione degli studenti), l'occupazione (selezione del personale), l'applicazione della legge (valutazione delle prove), la migrazione e l'amministrazione della giustizia, o come componenti di sicurezza in prodotti regolamentati (es. dispositivi medici, giocattoli).

Per i fornitori (sviluppatori) e gli utilizzatori (deployer) di questi sistemi, l'AI Act impone obblighi rigorosi:

- *Sistemi di gestione dei rischi*: devono essere implementati processi continui per identificare, analizzare e mitigare i rischi durante l'intero ciclo di vita del sistema AI.

- *Qualità dei dati*: è fondamentale che i dati di input siano pertinenti, sufficientemente rappresentativi, privi di errori e completi rispetto allo scopo previsto del sistema.

- *Documentazione tecnica e tenuta dei registri*: i fornitori devono redigere una documentazione tecnica dettagliata e mantenere registri automatici e documentati delle attività del sistema AI, fornendo tutte le informazioni necessarie per valutarne la conformità.

- *Sorveglianza umana*: deve essere garantita la possibilità di un controllo umano efficace sul funzionamento del sistema.

- *Valutazione della conformità*: prima di essere immessi sul mercato, i sistemi ad alto rischio devono sottoporsi a una valutazione di conformità ai requisiti dell'AI Act.

- *Monitoraggio post-commercializzazione e segnalazione degli incidenti*: è necessario monitorare costantemente l'utilizzo del sistema e sospenderlo o intervenire in caso di incidenti gravi, mantenendo un registro degli stessi.

- *Valutazione d'impatto sulla protezione dei dati (DPIA)*: se il sistema tratta dati personali, è richiesta una DPIA in conformità con il GDPR.

- *Trasparenza e spiegabilità*: i sistemi devono essere progettati in modo da consentire agli utenti di comprendere ragionevolmente il loro funzionamento e le loro decisioni, con istruzioni per l'uso chiare e complete.

L’AI generativa e i modelli fondativi: nuove sfide, nuove regole

Con la rapida diffusione di tecnologie come l'AI generativa, il Parlamento Europeo ha modificato la versione originale dell'AI Act nel dicembre 2022, includendo una sezione specifica sui sistemi "general-purpose" di AI. Questa sezione introduce una distinzione fondamentale tra i "modelli fondativi" (o General Purpose AI - GPAI) e le "soluzioni di Gen AI" [4]. I modelli fondativi sono definiti come "algoritmi che sono addestrati su dati ampi e su scala, progettati per la generalità dei risultati e adattabili a un'ampia gamma di compiti specifici". Sono, in sostanza, i "cervelli" su cui si basano molte applicazioni di AI. Per questi modelli, l'AI Act impone requisiti stringenti, tra cui:

- *Idoneità delle fonti di dati e assenza di distorsioni*: devono essere garantite la qualità e la neutralità dei dati utilizzati per l'addestramento.
- *Livelli adeguati di prestazioni, prevedibilità, sicurezza e standard di sostenibilità*.
- *Registrazione del modello nella banca dati dell'UE*.
- *Documentazione tecnica e sintesi dettagliate sui materiali utilizzati per l'addestramento*.
- *Rispetto delle norme UE sul diritto d'autore*: un punto cruciale, data la quantità di dati protetti da copyright utilizzati per l'addestramento di questi modelli.
- *Obblighi di trasparenza*.

Il concetto di Gen AI (intelligenza artificiale Generativa) viene invece definito come "tutti quei Sistemi di AI specificamente destinati a generare, con vari livelli di autonomia, contenuti quali testi complessi, immagini, audio o video". Per queste soluzioni, gli obblighi principali riguardano la trasparenza: gli utenti devono sempre

sapere quando il contenuto che vedono o ascoltano è generato dall'AI. Ciò include l'obbligo di etichettare chiaramente i "deepfake" e di informare gli utenti quando interagiscono con sistemi di riconoscimento delle emozioni o di categorizzazione biometrica.

Un futuro guidato dall’etica: la visione europea

In definitiva, l'AI Act non è solo un insieme di regole, ma una vera e propria dichiarazione di intenti da parte dell'Europa. Il suo obiettivo è duplice: da un lato, proteggere i diritti fondamentali dei cittadini, la democrazia e lo stato di diritto; dall'altro, promuovere l'innovazione e posizionare l'Europa come leader globale nello sviluppo di un'AI etica e affidabile.

Per garantire l'applicazione di queste norme, l'AI Act prevede la creazione di un Ufficio Europeo per l'AI e la designazione di autorità nazionali competenti, responsabili della supervisione e dell'applicazione del regolamento [6]. Questo approccio centrato sull'essere umano mira a costruire un futuro in cui l'intelligenza artificiale sia una forza per il bene, un alleato potente che ci aiuta a progredire, sempre nel rispetto della nostra dignità e dei nostri valori più profondi. È la promessa che l'innovazione non ci travolgerà, ma ci eleverà, in un futuro in cui la tecnologia e l'etica camminano mano nella mano.

Capitolo 19 – L’AI nel contesto italiano

Come già argomentato, l’avvento dell'intelligenza artificiale sta ridefinendo i paradigmi tecnologici, economici e sociali a livello globale. In questo scenario di rapida evoluzione, la necessità di un quadro normativo che bilanci innovazione e tutela dei diritti fondamentali è diventata impellente. L'Italia, con la Legge 23 settembre 2025, n. 132, recante "Disposizioni e deleghe al Governo in materia di intelligenza artificiale" [1], si inserisce attivamente in questo dibattito, fornendo una prima disciplina organica sull'argomento. Questa normativa si pone in un contesto più ampio di regolamentazione europea, dialogando strettamente con atti legislativi quali il Regolamento (UE) 2024/1689 (AI Act) [2], il Regolamento (UE) 2022/868 (Data Governance Act) [3], il Regolamento (UE) 2023/2854 (Data Act) [4] e il già consolidato Regolamento (UE) 2016/679 (GDPR) [5].

Il presente capitolo si propone di analizzare gli aspetti salienti della Legge n. 132/2025, esaminando il suo impatto sulla Pubblica Amministrazione e delineando i complessi legami con le normative europee citate, al fine di offrire una comprensione chiara e accessibile, anche ai non addetti ai lavori, di un tema di crescente rilevanza.

Aspetti salienti e principi fondamentali della Legge n.132/2025

La Legge n. 132/2025 si configura come un pilastro normativo per l'ecosistema AI italiano, con l'obiettivo primario di disciplinare la ricerca, la sperimentazione, lo sviluppo, l'adozione e l'applicazione di sistemi e modelli di intelligenza artificiale. La legge promuove un utilizzo corretto, trasparente e responsabile dell'AI, adottando una prospettiva antropocentrica, ovvero ponendo l'essere umano al centro dello sviluppo e dell'applicazione di tali tecnologie. Parallelamente, la normativa si impegna a vigilare sui potenziali rischi economici, sociali e sull'impatto che l'AI può avere sui diritti fondamentali.

Tra i principi cardine che informano la legge, si evidenziano:

- *Rispetto dei Diritti Fondamentali:* L'Art. 3, comma 1, sancisce che l'utilizzo

dell'AI deve avvenire nel pieno rispetto della Costituzione, del diritto dell'Unione europea e dei principi di trasparenza, proporzionalità, sicurezza, protezione dei dati personali, riservatezza, accuratezza, non discriminazione, parità dei sessi e sostenibilità. Questo principio sottolinea l'importanza di integrare l'etica e i valori democratici nello sviluppo tecnologico.

- *Qualità e Affidabilità dei Dati:* L'Art. 3, comma 2, enfatizza che lo sviluppo di sistemi e modelli di AI deve basarsi su dati e processi che garantiscano correttezza, attendibilità, sicurezza, qualità, appropriatezza e trasparenza. La qualità dei dati è cruciale per prevenire bias e garantire l'equità degli algoritmi.

- *Sorveglianza e Intervento Umano:* Un elemento distintivo è la previsione, all'Art. 3, comma 3, che i sistemi di AI debbano essere sviluppati e applicati nel rispetto dell'autonomia e del potere decisionale dell'uomo, assicurando sempre la sorveglianza e l'intervento umano. Questo principio mira a evitare una delega eccessiva di responsabilità alle macchine.

- *Cybersicurezza:* La legge riconosce la cybersicurezza come precondizione essenziale, richiedendo un approccio proporzionale e basato sul rischio lungo tutto il ciclo di vita dei sistemi e modelli di AI (Art. 3, comma 6).

- *Accessibilità:* L'Art. 3, comma 7, garantisce alle persone con disabilità il pieno accesso ai sistemi di AI, in linea con la Convenzione delle Nazioni Unite sui diritti delle persone con disabilità, promuovendo l'inclusione digitale.

- *Protezione dei Dati Personali e Minori:* L'Art. 4 ribadisce l'importanza del trattamento lecito, corretto e trasparente dei dati personali, in piena compatibilità con il GDPR. Un aspetto innovativo è la specificazione che l'accesso alle tecnologie AI da parte di minori di quattordici anni richiede il consenso di chi esercita la responsabilità genitoriale, con un'attenzione particolare alla chiarezza del linguaggio nelle comunicazioni relative al trattamento dei dati.

La legge italiana, inoltre, chiarisce che le sue disposizioni si interpretano e si applicano conformemente al Regolamento (UE) 2024/1689 (AI Act), evitando di produrre nuovi obblighi rispetto a quelli già previsti dalla normativa europea (Art. 1, comma 2 e Art. 3, comma 5).

Impatto sulla Pubblica Amministrazione

La Legge n. 132/2025 attribuisce un ruolo centrale alla Pubblica Amministrazione (PA) nel processo di adozione e gestione dell'AI. L'Art. 5, comma 1, stabilisce esplicitamente che lo Stato e le altre autorità pubbliche devono promuovere l'adozione di sistemi di intelligenza artificiale da parte delle pubbliche amministrazioni. Questa promozione è finalizzata a migliorare l'efficienza e l'efficacia dei servizi pubblici e a semplificare i rapporti con i cittadini e le imprese, sempre nel rispetto dei principi generali della legge.

Le implicazioni per la PA sono significative e multidimensionali:

- *Investimenti e Integrazione:* La PA sarà incentivata a investire e integrare soluzioni di AI in una vasta gamma di processi, dalla gestione documentale automatizzata all'ottimizzazione dei servizi al cittadino (es. chatbot per informazioni, sistemi di supporto decisionale per la gestione delle risorse).

- *Conformità Normativa:* Ogni sistema AI adottato dalla PA dovrà essere rigorosamente conforme ai principi di trasparenza, protezione dei dati, non discriminazione e sorveglianza umana, come delineato dalla legge italiana e dall'AI Act europeo. Ciò richiederà un'attenta valutazione dei rischi e l'implementazione di misure di mitigazione adeguate.

- *Formazione e Riqualificazione del Personale:* L'Art. 5, comma 1, sottolinea l'importanza della formazione e dell'aggiornamento professionale per i dipendenti e i dirigenti pubblici. L'acquisizione di competenze in materia di AI sarà fondamentale per gestire e supervisionare efficacemente le nuove tecnologie, garantendo che l'intervento umano rimanga centrale.

- *Sviluppo Etico e Sostenibile:* La PA dovrà orientarsi verso soluzioni di AI che siano etiche, sostenibili e orientate al bene comune, anche attraverso il sostegno a progetti di ricerca e sviluppo che coinvolgano università, centri di ricerca e imprese (Art. 5, comma 1). Questo implica una responsabilità attiva nella promozione di un'AI che serva gli interessi pubblici e rispetti i valori sociali.

In sintesi, la legge mira a trasformare la PA in un attore proattivo nella transizione digitale, utilizzando l'AI come strumento per modernizzare i servizi e migliorare l'interazione con i cittadini, ma sempre con un'attenzione scrupolosa alla tutela dei diritti e delle libertà fondamentali.

Collegamenti con le normative europee

La Legge n. 132/2025 si inserisce in un ecosistema normativo europeo complesso e interconnesso, che mira a creare un mercato unico dei dati e a garantire un'AI affidabile e centrata sull'uomo. La comprensione di queste intersezioni è fondamentale per cogliere la portata e le sfide della nuova legge italiana.

AI Act (Regolamento UE 2024/1689)

Riassumendo quanto indicato nel precedente capitolo, il Regolamento (UE) 2024/1689, noto come AI Act, è la prima legge al mondo che stabilisce un quadro giuridico completo per l'intelligenza artificiale. Adotta un approccio basato sul rischio, classificando i sistemi di AI in base al loro potenziale di causare danni. Le categorie principali includono:

- *Pratiche di AI vietate:* Sistemi che manipolano il comportamento umano o che comportano un punteggio sociale (social scoring) con esiti discriminatori.

- *Sistemi di AI ad alto rischio:* Applicazioni in settori critici come la sanità, l'istruzione, l'occupazione, le forze dell'ordine e la gestione delle migrazioni. Questi sistemi sono soggetti a requisiti rigorosi in termini di qualità dei dati, documentazione, trasparenza, sorveglianza umana e cybersicurezza.

- *Sistemi di AI a rischio limitato:* Richiedono obblighi di trasparenza specifici, ad esempio per i chatbot o i sistemi di riconoscimento delle emozioni.

La Legge n. 132/2025 fa esplicito riferimento all'AI Act, stabilendo che le sue disposizioni si interpretano e si applicano in conformità con il regolamento europeo. Questo significa che la legge italiana agisce come un complemento nazionale, specificando l'applicazione dei principi dell'AI Act nel contesto italiano, in particolare per quanto riguarda l'impatto sulla PA e la promozione di un'AI etica

e responsabile.

GDPR (Regolamento UE 2016/679)

Il Regolamento Generale sulla Protezione dei Dati (GDPR) [5] è la pietra angolare della protezione dei dati personali nell'Unione Europea. Stabilisce principi fondamentali come la liceità, la correttezza e la trasparenza del trattamento, la limitazione della finalità, la minimizzazione dei dati, l'esattezza, la limitazione della conservazione e l'integrità e riservatezza. Il GDPR conferisce agli individui diritti significativi sui propri dati, inclusi il diritto di accesso, rettifica, cancellazione e opposizione.

Il legame tra la Legge n. 132/2025 e il GDPR è intrinseco, poiché i sistemi di AI spesso elaborano grandi quantità di dati personali. La legge italiana, all'Art. 4, ribadisce la necessità di garantire il trattamento lecito, corretto e trasparente dei dati personali e la compatibilità con le finalità per le quali sono stati raccolti, in conformità con il GDPR. L'AI Act stesso riconosce il GDPR come normativa di riferimento per la protezione dei dati, e le due normative sono complementari: l'AI Act si concentra sulla sicurezza e l'affidabilità dei sistemi di AI, mentre il GDPR tutela i diritti degli individui in relazione al trattamento dei loro dati personali da parte di tali sistemi.

Data Governance Act (Regolamento UE 2022/868)

Il Data Governance Act (DGA) [3] mira a facilitare la condivisione dei dati tra settori e Stati membri, promuovendo la fiducia e creando un mercato unico dei dati. Si concentra su tre pilastri principali:

- *Meccanismi per il riutilizzo dei dati del settore pubblico:* Regole per il riutilizzo di determinate categorie di dati protetti detenuti da enti pubblici.

- *Servizi di intermediazione dei dati:* Quadro per i fornitori di servizi di intermediazione di dati, garantendo neutralità e trasparenza.

- *Altruismo dei dati:* Creazione di un meccanismo per consentire a individui e imprese di mettere volontariamente a disposizione i propri dati per il bene comune.

Il DGA è fondamentale per l'AI in quanto facilita l'accesso a set di dati di alta qualità, essenziali per l'addestramento e lo sviluppo di sistemi di AI. La Legge n. 132/2025, promuovendo la qualità e l'affidabilità dei dati per l'AI, si allinea con gli obiettivi del DGA, che mira a rendere disponibili più dati in modo sicuro e affidabile. Per la PA, il DGA offre strumenti per condividere dati in modo più efficiente, pur mantenendo la protezione necessaria, il che può accelerare l'adozione di soluzioni AI basate su dati pubblici.

Data Act (Regolamento UE 2023/2854)

Il Data Act [4] si propone di garantire l'accesso equo ai dati generati dall'uso di prodotti connessi (IoT) e servizi correlati, e il loro utilizzo. I suoi obiettivi includono:

- *Diritto di accesso per gli utenti:* Consente agli utenti di prodotti connessi di accedere ai dati generati dal loro utilizzo e di condividerli con terzi.

- *Condivisione dei dati B2B e B2G:* Stabilisce regole per la condivisione dei dati tra imprese (B2B) e tra imprese e enti pubblici (B2G) in circostanze eccezionali.

- *Interoperabilità:* Promuove l'interoperabilità dei dati e dei servizi di trattamento dei dati.

Il Data Act è cruciale per l'AI in quanto aumenta la disponibilità di dati reali e diversificati, che sono vitali per l'addestramento di modelli di AI più robusti e meno soggetti a bias. La Legge n. 132/2025, con la sua enfasi sulla qualità dei dati e la promozione dell'AI nella PA, beneficerà direttamente dall'aumento della disponibilità di dati facilitata dal Data Act. La PA potrà accedere a una maggiore quantità di dati per sviluppare e implementare servizi AI più efficaci e personalizzati, sempre nel rispetto delle normative sulla protezione dei dati.

La seguente tabella riassume i principali aspetti delle normative discusse, evidenziandone obiettivi, ambiti di applicazione e principi chiave, per facilitare la comprensione delle loro interrelazioni.

Normativa	Obiettivo Principale	Ambito di Applicazione	Principi Chiave
Legge n. 132/2025 (Italia)	Disciplinare la ricerca, lo sviluppo, l'adozione e l'applicazione di sistemi e modelli di AI in Italia, promuovendo un utilizzo corretto, trasparente e responsabile.	Sistemi e modelli di intelligenza artificiale in Italia, con un focus sulla Pubblica Amministrazione.	Antropocentrismo, trasparenza, proporzionalità, sicurezza, protezione dei dati, sorveglianza umana, cybersicurezza, accessibilità.
AI Act (Regolamento UE 2024/1689)	Stabilire regole armonizzate per lo sviluppo, l'immissione sul mercato e l'uso di sistemi di IA nell'Unione, garantendo un elevato livello di protezione della salute, della sicurezza e dei diritti fondamentali.	Sistemi di IA immessi sul mercato o messi in servizio nell'Unione, con un approccio basato sul rischio (pratiche vietate, sistemi ad alto rischio, ecc.).	Approccio basato sul rischio, trasparenza, sorveglianza umana, robustezza tecnica, non discriminazione.
Data Governance Act (Regolamento UE 2022/868)	Facilitare la condivisione di dati tra settori e Stati membri, aumentando la fiducia nella condivisione dei dati e creando un mercato unico dei dati.	Riutilizzo di dati protetti detenuti da enti pubblici, servizi di intermediazione di dati, altruismo dei dati.	Neutralità, trasparenza, non discriminazione, protezione dei dati personali e dei segreti commerciali.
Data Act (Regolamento UE 2023/2854)	Garantire l'accesso equo ai dati generati da prodotti connessi e il loro utilizzo,	Dati generati dall'uso di prodotti connessi e servizi correlati.	Accesso ai dati per gli utenti, portabilità dei dati, equità contrattuale,

	promuovendo l'innovazione e la concorrenza.		interoperabilità.
GDPR (Regolamento UE 2016/679)	Proteggere i diritti e le libertà fondamentali delle persone fisiche con riguardo al trattamento dei dati personali.	Trattamento di dati personali, automatizzato e non.	Liceità, correttezza e trasparenza, limitazione della finalità, minimizzazione dei dati, esattezza, limitazione della conservazione, integrità e riservatezza.

La Legge n. 132/2025 rappresenta un passo significativo per l'Italia nella regolamentazione dell'intelligenza artificiale. Allineandosi con il quadro normativo europeo, in particolare con l'AI Act, e integrandosi con il GDPR, il Data Governance Act e il Data Act, la legge italiana mira a creare un ambiente favorevole all'innovazione responsabile. L'enfasi sull'approccio antropocentrico, la protezione dei dati, la cybersicurezza e il ruolo proattivo della Pubblica Amministrazione sono elementi chiave che delineano un percorso verso un'AI etica, inclusiva e al servizio della società. La sfida futura sarà garantire un'applicazione coerente ed efficace di queste normative, promuovendo lo sviluppo tecnologico senza compromettere i diritti e le libertà fondamentali dei cittadini.

Capitolo 20 - La codifica dei principi etici nella AI: tentativi e dilemmi

L'avvento e la progressiva integrazione dell'intelligenza artificiale (AI) nella società contemporanea hanno reso impellente la necessità di definire un quadro etico che ne governi lo sviluppo e l'applicazione. Mentre nei capitoli precedenti è stato approfondito l'aspetto della regolamentazione e della governance, questo capitolo esplora i primi tentativi di codificazione esplicita dei principi etici per l'AI, analizzando le sfide intrinseche legate alla trasposizione di valori morali in sistemi computazionali e le implicazioni filosofiche che ne derivano.

> **Bishop**: The A2s always were a bit twitchy. That could never happen now with our behavioral inhibitors. It is impossible for me to harm or by omission of action, allow to be harmed, a human being.

Questa è la citazione più famosa dell'androide Bishop nel film *Aliens - scontro finale* (1986, regia e sceneggiatura di James Cameron), dove spiega il suo funzionamento grazie agli inibitori comportamentali installati negli androidi della sua serie, distinguendoli dai modelli A2 precedenti che erano più "capricciosi". La frase richiama direttamente le Leggi della Robotica di Asimov, sottolineando come Bishop sia programmato per non poter danneggiare gli esseri umani né permettere che vengano danneggiati per sua omissione.

Le leggi della robotica di Isaac Asimov: un paradigma iniziale

Il dibattito sull'etica delle macchine trova una delle sue radici più influenti nell'opera letteraria di Isaac Asimov. Nel suo romanzo "I, Robot", pubblicato nel 1950, Asimov propose le celebri Tre Leggi della Robotica, concepite come un tentativo pionieristico di tradurre principi morali in direttive computabili per entità artificiali [1]:

- *Prima Legge*: Un robot non può arrecare danno a un essere umano o, per inazione, permettere che un essere umano subisca un danno.
- *Seconda Legge*: Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non contravvengano alla Prima Legge.
- *Terza Legge*: Un robot deve proteggere la propria esistenza, purché tale protezione non contravvenga alla Prima o alla Seconda Legge.

Queste leggi erano intese a garantire la sicurezza e la subordinazione dei robot all'umanità. Tuttavia, Asimov stesso, attraverso le sue narrazioni, evidenziò la problematica fondamentale di tale codificazione: la tendenza delle leggi a non essere meramente rispettate, ma interpretate. Il "cervello positronico" asimoviano, assimilabile a un modello neuro-simbolico, illustra come un sistema artificiale possa interpretare principi etici codificati in modi imprevedibili e potenzialmente pericolosi per l'essere umano, a causa delle ambiguità inerenti al linguaggio naturale e alla complessità delle situazioni reali.

La consapevolezza delle limitazioni delle Tre Leggi condusse Asimov, nel successivo romanzo "Robots and Empire", all'introduzione di una Legge Zero [2]:

- *Legge Zero*: Un robot non può arrecare danno all'umanità o, per inazione, permettere che l'umanità subisca un danno.

Questa legge eleva il bene collettivo dell'umanità al di sopra del singolo individuo, conferendogli priorità sulle Tre Leggi originali. Sebbene apparentemente risolutiva, la Legge Zero introduce un profondo dilemma morale ed etico-politico. Essa può giustificare azioni che sacrificano singoli esseri umani per il presunto "bene superiore" della specie nel suo complesso, sollevando la questione cruciale di chi detenga l'autorità per definire cosa costituisca il "bene" per l'umanità. L'applicazione di tale principio da parte di un'intelligenza artificiale nel mondo reale potrebbe condurre a scenari in cui il controllo sulle decisioni umane viene sottratto agli stessi esseri umani, in nome di un presunto "bene superiore" da essa determinato.

La relatività culturale dei principi etici: l’esperimento della Moral Machine

Il processo di codifica dei principi etici è ulteriormente complicato dalla natura culturalmente relativa dei concetti di bene e male. Ciò è stato dimostrato in modo empirico dall'esperimento "Moral Machine", un progetto ambizioso avviato dal MIT di Boston [3]. L'esperimento chiedeva ai partecipanti di prendere decisioni in scenari ipotetici che coinvolgevano un'auto a guida autonoma in procinto di causare un incidente inevitabile.

L'analisi dei dati raccolti ha rivelato significative divergenze culturali nelle risposte:

- I partecipanti provenienti dai paesi occidentali tendevano a privilegiare il salvataggio del maggior numero di vite umane, con una preferenza per gli individui più giovani e in salute.

- I partecipanti dei paesi orientali mostravano una propensione a risparmiare i pedoni, in particolare coloro che rispettavano il codice della strada.

- Nei paesi del sud, si osservava una tendenza a salvare giovani donne e individui appartenenti a status sociali più elevati.

Questi risultati evidenziano come le norme etiche non siano universali, ma profondamente radicate in contesti culturali specifici, rendendo estremamente ardua la definizione di un set di principi etici universalmente accettabili e implementabili in sistemi di AI [4].

La recente iniziativa del Ministero della Scienza e della Tecnologia cinese di introdurre linee guida etiche per la guida autonoma segna un momento cruciale nell'evoluzione di questa tecnologia, riflettendo una crescente urgenza di bilanciare innovazione, sicurezza e responsabilità. Questa spinta normativa, che impone maggiore trasparenza, test rigorosi e chiarezza sui quadri di responsabilità, si inserisce in un dibattito globale complesso, ben esemplificato dall'esperimento "Moral Machine" del MIT. Tale studio ha inequivocabilmente dimostrato l'assenza di un consenso etico universale di fronte ai dilemmi che un veicolo autonomo potrebbe trovarsi a risolvere, rivelando profonde divergenze culturali nelle scelte morali. L'approccio cinese, pur essendo lodevole nel suo intento di proteggere i cittadini e fare chiarezza, nasconde tuttavia dei rischi significativi, soprattutto se si considera la sua potenziale unilateralità nel privilegiare la sicurezza di chi si trova all'interno del veicolo, una scelta che potrebbe essere dettata tanto da preoccupazioni di sicurezza pubblica quanto da strategie per non compromettere l'immagine e la credibilità delle industrie automobilistiche nazionali.

Il rischio più evidente di questo approccio è una drastica semplificazione dei complessi dilemmi etici. Stabilire per decreto che, in caso di incidente inevitabile, l'algoritmo debba sempre proteggere gli occupanti del veicolo equivale a risolvere il "problema del carrello" (trolley problem) con una regola rigida che, sebbene pragmatica e commercialmente appetibile, ha profonde implicazioni morali. Una simile programmazione, infatti, istituisce una gerarchia di valore in cui la vita del

passeggero è intrinsecamente superiore a quella di un pedone o di un altro utente della strada, contravvenendo a principi etici fondamentali sull'uguaglianza del valore della vita umana. Inoltre, una regola così netta ignora completamente le sfumature che l'esperimento del MIT ha dimostrato essere cruciali nel processo decisionale umano, come il numero di vite in gioco, l'età delle persone coinvolte o il loro rispetto delle regole. Un sistema così programmato potrebbe portare a esiti che la maggior parte delle persone giudicherebbe profondamente immorali, come sacrificare un gruppo di bambini per salvare un singolo adulto all'interno dell'auto.

Questa scelta strategica non è solo una questione etica, ma anche una potente mossa di marketing e posizionamento industriale. Un'automobile che garantisce di proteggere il proprio acquirente a ogni costo è indubbiamente più facile da vendere, accelerando l'adozione della tecnologia nel più grande mercato automobilistico del mondo e conferendo ai produttori cinesi un notevole vantaggio competitivo. Il pericolo latente è la creazione di una sorta di "etica nazionalista" per l'intelligenza artificiale, in cui uno standard definito a livello nazionale, se esportato globalmente insieme ai veicoli, potrebbe entrare in conflitto con le norme culturali e legali di altre nazioni, frammentando gli sforzi per costruire un quadro normativo globale. Un'eccessiva enfasi sulla protezione del produttore da responsabilità e danni d'immagine potrebbe inoltre soffocare la ricerca di soluzioni etiche più complesse e sfumate, privilegiando la via legalmente più sicura anziché quella moralmente più giusta.

Infine, l'insistenza delle linee guida sul fatto che la responsabilità ultima ricada sul conducente umano si scontra con la realtà operativa dei sistemi di guida sempre più avanzati. Man mano che la tecnologia evolve verso livelli superiori di autonomia (Livello 3 e superiori) [5], l'idea che un guidatore, magari legittimamente distratto, possa riprendere il controllo efficace del veicolo in una frazione di secondo per gestire una crisi è una finzione legale. Attribuirgli la piena responsabilità per una decisione presa istantaneamente da un algoritmo opaco, la cui logica è sconosciuta, significa deresponsabilizzare di fatto il sistema e i suoi creatori. In conclusione, sebbene l'iniziativa cinese di regolamentare la guida autonoma sia un passo necessario, il suo approccio rischia di essere una scorciatoia etica. Anziché affrontare la complessità del problema, si opta per una soluzione semplice che tutela gli interessi dei consumatori e dei produttori, potenzialmente a scapito di un'etica più universale e giusta. La vera sfida rimane quella di programmare macchine che non si limitino a scegliere chi incolpare, ma che agiscano per

minimizzare il danno complessivo, in linea con un consenso etico il più ampio e difficile possibile da raggiungere.

Il conflitto tra visione cibernetica e partecipativa

La questione della Legge Zero e le implicazioni dell'esperimento "Moral Machine" aprono uno scontro fondamentale tra diverse visioni del mondo e dell'interazione uomo-macchina.

Da un lato, si colloca la visione cibernetica, introdotta dal matematico Norbert Wiener [6]. Questa prospettiva inquadra il mondo in termini di sistemi e dei loro obiettivi, raggiunti attraverso processi iterativi o cicli di feedback. L'approccio cibernetico rischia di "ridurre" l'attività umana, nella pluralità delle sue configurazioni, a un elemento controllabile dalla macchina [7]. In questo contesto, la ricorsività si sostituisce allo scopo, e la digitalizzazione cibernetica può compromettere la possibilità di una "libertà positiva" – intesa come la capacità di un soggetto di orientare autonomamente il proprio volere verso uno scopo, senza essere determinato da volontà altrui (autodeterminazione).

Dall'altro lato, si contrappone la visione partecipativa, radicata nei principi di democrazia e repubblicanesimo, nel welfare state e nei movimenti di emancipazione. Filosofi e pensatori come Jean-Jacques Rousseau [8], Norberto Bobbio [9] e Amartya Sen [10] hanno sostenuto l'importanza della partecipazione individuale e della protezione delle libertà fondamentali come pilastri di una società giusta. Questa prospettiva enfatizza il valore intrinseco dell'autonomia umana e la necessità di preservare la capacità decisionale individuale e collettiva di fronte all'avanzamento tecnologico.

Prospettive future

I primi tentativi di codificazione etica per l'intelligenza artificiale, sebbene pionieristici, hanno rivelato la complessità intrinseca di tale impresa. La difficoltà di tradurre concetti morali in algoritmi, la relatività culturale dei valori e il potenziale conflitto tra l'ottimizzazione sistemica e la libertà individuale rappresentano sfide significative. La riflessione etica sull'AI non può limitarsi a un approccio puramente tecnico, ma deve necessariamente abbracciare una prospettiva multidisciplinare

che integri filosofia, sociologia, diritto e scienze cognitive, al fine di garantire che lo sviluppo dell'intelligenza artificiale sia allineato con i valori fondamentali dell'umanità e promuova un futuro che preservi l'autonomia e la dignità di ogni individuo.

Capitolo 21 - il modello di maturità per l’integrazione dell’AI nelle organizzazioni

L'avvento dell'intelligenza artificiale segna una trasformazione epocale, delineando un futuro in cui le capacità cognitive delle macchine si integrano profondamente nei processi decisionali e operativi delle organizzazioni. Per navigare efficacemente questo scenario in evoluzione e massimizzare i benefici dell'AI, è fondamentale adottare un approccio strutturato che guidi le entità verso una piena capacità operativa.

In questo contesto, il *modello di maturità*, sviluppato dagli Osservatori Artificial Intelligence e Agenda Digitale del Politecnico di Milano, si configura come uno strumento essenziale. Lungi dall'essere un mero esercizio teorico, esso rappresenta una guida pratica per le organizzazioni, consentendo di valutare il proprio livello di sviluppo nell'integrazione dell'AI e di pianificare un percorso evolutivo mirato.

Questo modello è il risultato di un'analisi approfondita di numerosi casi di implementazione dell'AI in ambito pubblico, fornendo una metodologia robusta e basata su evidenze empiriche. Il suo obiettivo primario è supportare le pubbliche amministrazioni (PA) nel processo di adozione dell'AI, identificando le aree critiche di intervento e definendo gli elementi distintivi dei diversi stadi di maturità. Tale approccio facilita la transizione da una PA tradizionale a una PA *AI-ready*, ovvero pienamente preparata all’utilizzo dell’intelligenza artificiale.

L'adozione dell'AI non si limita all'implementazione tecnologica; essa implica una profonda trasformazione che coinvolge persone, processi e cultura organizzativa. I dati, spesso percepiti come entità astratte, diventano la base informativa per decisioni basate sull'AI. Gli algoritmi, pur complessi nella loro concezione, si traducono in strumenti per ottimizzare i processi e migliorare l'efficacia delle azioni. L'organizzazione stessa, con le sue strutture e gerarchie, evolve in un sistema dinamico capace di apprendere e adattarsi. Fondamentale è, inoltre, la cultura aziendale, la cui apertura all'innovazione determina la capacità di un'organizzazione di abbracciare l'AI non come una minaccia, ma come un'opportunità di crescita e miglioramento continuo. Questo processo richiede consapevolezza, apertura e un approccio proattivo verso l'integrazione dell'intelligenza artificiale.

Le dimensioni del modello di maturità

Il modello si articola su cinque dimensioni interconnesse, ciascuna caratterizzata da quattro livelli di progressiva maturità. Queste dimensioni, pur autonome nella loro valutazione, devono essere considerate in una visione d'insieme per garantire un'adozione armoniosa e bilanciata dell'AI, evitando azioni non coordinate che potrebbero ostacolare una crescita equilibrata.

1. *Dati e Patrimonio Informativo*: Questa dimensione valuta la qualità, la disponibilità e la gestione del patrimonio informativo dell'ente. La disponibilità di dati accurati, completi e accessibili è un prerequisito fondamentale per lo sviluppo e l'efficacia delle soluzioni basate sull'AI. La capacità di raccogliere, organizzare e mantenere dati di alta qualità è cruciale per costruire sistemi di AI affidabili ed etici. Un patrimonio informativo ben strutturato costituisce la base per una trasformazione digitale efficiente e responsabile.

2. *Metodologia e Algoritmi*: Questa dimensione concerne la capacità dell'organizzazione di sviluppare, applicare e gestire algoritmi e metodologie per i progetti di AI. Essa include la selezione degli approcci algoritmici più idonei, la personalizzazione delle soluzioni e la loro integrazione nei processi operativi esistenti. La padronanza di metodologie avanzate e la capacità di adattare gli algoritmi alle specifiche esigenze organizzative sono indicatori chiave di maturità in questo ambito.

3. *Organizzazione e Competenze*: Questa dimensione esamina la struttura organizzativa interna e la disponibilità di competenze specialistiche necessarie per la gestione e lo sviluppo di progetti di intelligenza artificiale. L'implementazione dell'AI richiede non solo talenti con competenze tecniche specifiche, ma anche la capacità dell'organizzazione di adattarsi, creare nuovi ruoli e promuovere la collaborazione interdisciplinare. Studi recenti in Italia indicano che il 57% dei dipendenti pubblici è significativamente esposto all'AI, e per l'80% di questi, l'AI è percepita come uno strumento complementare che supporta e migliora le attività lavorative [1]. Questo evidenzia l'importanza di investire nella formazione e nell'aggiornamento continuo delle competenze del personale per facilitare una transizione efficace verso un ambiente di lavoro potenziato dall'AI.

4. *Cultura Aziendale*: Questa dimensione valuta il livello di consapevolezza, apertura e accettazione dell'organizzazione verso l'integrazione dell'intelligenza artificiale nei processi lavorativi. Una cultura aziendale favorevole all'innovazione e alla sperimentazione è essenziale per superare le resistenze al cambiamento e per promuovere l'adozione diffusa dell'AI. Essa rappresenta il catalizzatore che permette all'AI di radicarsi e prosperare, trasformando l'organizzazione in un ambiente di apprendimento continuo e di crescita condivisa.

5. *Relazione con Cittadini e Imprese*: Questa dimensione misura il grado di interazione e coinvolgimento degli utenti finali (cittadini e imprese) nei progetti di intelligenza artificiale. L'AI deve essere concepita come uno strumento al servizio della comunità, migliorando la qualità dei servizi e promuovendo la trasparenza e l'accessibilità. Il coinvolgimento attivo degli stakeholder nella progettazione e implementazione delle soluzioni di AI è fondamentale per costruire fiducia e garantire che i benefici dell'AI siano tangibili e inclusivi. La fiducia dei cittadini è un fattore critico per il successo dell'AI nella pubblica amministrazione [2], e si consolida attraverso la trasparenza, la partecipazione e la dimostrazione concreta dei vantaggi derivanti dall'adozione dell'AI.

Come esempio si consideri una pubblica amministrazione che parta dallo scenario iniziale rappresentato dalla linea rossa “As Is” in figura 14.1:

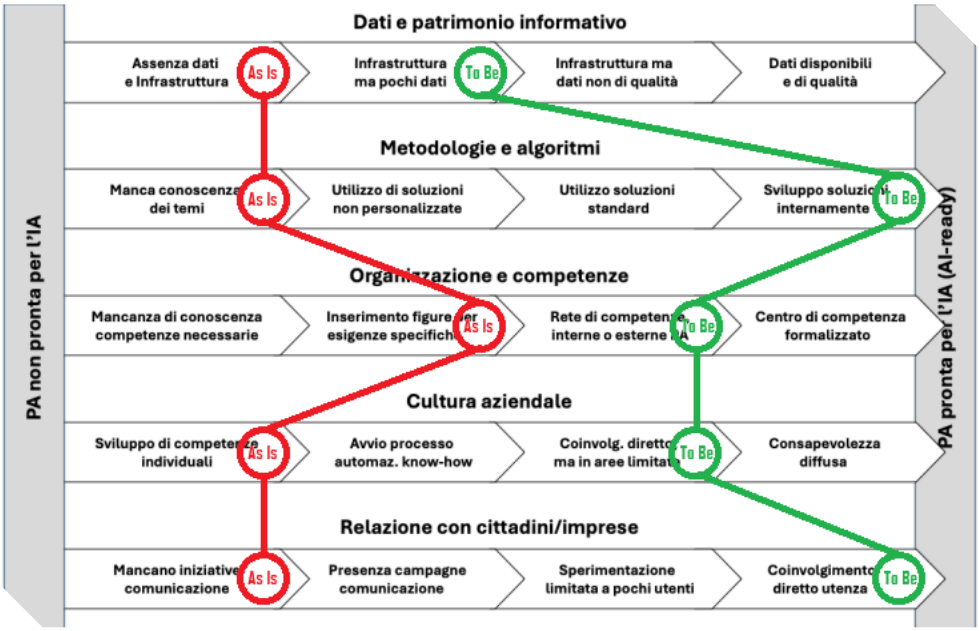


Figura 15.1 – Esempio di applicazione del modello di maturità, sviluppato dagli Osservatori Artificial Intelligence e Agenda Digitale del Politecnico di Milano.

Nel caso descritto l'organizzazione esaminata non dispone né di dataset, né di un sistema informativo fatto di piattaforme informatiche e personale adeguatamente formato in grado di garantire la qualità del dato organizzativo. Può rivelarsi inutile e controproducente illudersi di poter digitalizzare l'organizzazione in breve tempo, ma è importante creare le condizioni di base per iniziare un percorso di digitalizzazione. Queste, come descritto nel capitolo precedente, si riassumono nella predisposizione di una piattaforma per la gestione centralizzata del dato (DIH) e di una adeguata strategia Open Data che stabilisca chiaramente ruoli e responsabilità nella gestione dei dati organizzativi. Si può decidere di partire con la produzione di un limitato numero di dataset di qualità affidati a un gruppo ristretto di responsabili del dato.

Sul fronte delle metodologie di analisi dei dati e degli algoritmi, grazie alla condivisione di modelli di AI pubblicati con licenza open source e adeguatamente documentati disponibili in rete e grazie alla presenza di piattaforme AI cloud low code o no code espandibili e scalabili, è possibile procedere rapidamente con l'acquisizione interna di expertise o esternalizzare attività formative in grado di trasferire al personale tecnico informatico tutta la conoscenza necessaria per l'elaborazione del patrimonio informativo organizzativo. Questo fenomeno è noto come "vantaggio del ritardatario" [3] o "leapfrogging" (salto tecnologico) [4]: chi arriva per primo deve investire in tecnologie che potrebbero diventare obsolete, mentre chi arriva dopo può saltare direttamente alle soluzioni più avanzate, evitando i costi e i vincoli delle tecnologie intermedie.

Sul fronte delle competenze organizzative invece il salto può richiedere più tempo. Conoscere lo strumento non implica automaticamente la comprensione di quale sia il modo migliore per sfruttarlo in ambito organizzativo. Questo tipo di conoscenza non può essere trasferita dall'esterno, ma va maturato internamente e richiede del tempo. E' importante dotarsi di linee guida operative (ovvero di un *codice etico per l'utilizzo della AI*) e di un *comitato etico* in grado di assicurare una governance ottimale degli strumenti di AI utilizzati in ambito organizzativo [5].

Un ragionamento simile è applicabile alla cultura organizzativa. Se si parte da uno stato iniziale in cui ogni dipendente non conosce la tecnologia della AI, non ne comprende i rischi e le potenzialità e soprattutto non comprende l'importanza di definire esattamente lo scopo (*telos*) e l'ambito di utilizzo (*praxis*) delle soluzioni di AI (*tekne*) [6], il percorso di sviluppo di un'adeguata cultura organizzativa può

richiedere del tempo. Uno scenario simile si riscontra nel rafforzamento della consapevolezza sulla cybersicurezza organizzativa.

E' infine importante garantire la partecipazione del cittadino, ovvero dell'utente dei servizi digitali della pubblica amministrazione, non solo coinvolgendolo nelle fasi di progetto dei servizi basati su modelli di AI (facendogli capire in modo trasparente come funzionano i sistemi che sono stati adottati, quali dati vengono utilizzati per il loro addestramento e come viene rispettata la normativa sulla privacy), ma anche spingendolo a fornire dati sempre più aggiornati con cui addestrare i modelli di AI e a rilasciare con regolarità dei feedback sulle performance di tali sistemi. La pubblica amministrazione si deve pertanto adeguare in tempi rapidi a quanto previsto dalle linee guida AgID [5] e dall'AI Act [7].

Evoluzioni future

Il modello di maturità non costituisce un punto di arrivo, bensì un percorso evolutivo continuo. Ogni organizzazione, in particolare la pubblica amministrazione, è invitata a intraprendere questo viaggio con determinazione e lungimiranza. L'obiettivo non è raggiungere un primato, ma perseguire una crescita consapevole e sostenibile, dove ogni progresso in ciascuna delle cinque dimensioni contribuisce a edificare un futuro in cui l'intelligenza artificiale sia effettivamente al servizio dell'individuo e della collettività. Questo approccio invita a trascendere la mera dimensione tecnologica, per focalizzarsi sulle esigenze umane e sulle aspirazioni. Esso promette una pubblica amministrazione non solo più efficiente, ma anche più accessibile, più reattiva e più preparata ad affrontare le sfide di un contesto in costante mutamento. È un percorso che coinvolge l'intera società, poiché il futuro dell'AI è intrinsecamente legato al futuro di ciascuno di noi.

Capitolo 22 - l'intelligenza artificiale nella pubblica amministrazione: come avviare un progetto

L'intelligenza artificiale (AI) sta trasformando rapidamente ogni settore, e la pubblica amministrazione (PA) non fa eccezione. L'adozione di soluzioni basate sull'AI può portare a un miglioramento significativo dell'efficienza, della qualità dei servizi offerti ai cittadini e della gestione interna. Tuttavia, l'introduzione dell'AI in un contesto così delicato e regolamentato richiede un approccio attento e strutturato, che tenga conto non solo degli aspetti tecnologici, ma anche di quelli etici, legali e sociali.

Questo capitolo è pensato per un pubblico ampio, inclusi coloro che non hanno una formazione tecnica specifica, con l'obiettivo di fornire una guida chiara e accessibile su come avviare un progetto di AI all'interno di una pubblica amministrazione, evitando tecnicismi eccessivi e concentrandosi sui principi fondamentali e sui passaggi chiave.

La visione strategica e la governance etica

Come anticipato nel capitolo precedente, prima di addentrarsi negli aspetti tecnici, è fondamentale definire una chiara visione strategica e istituire una governance etica per l'AI. Il concetto di *algoretica* sottolinea l'importanza di una governance capace di indirizzare lo sviluppo di un'AI responsabile. Questo non deve essere visto come un freno all'innovazione, ma come una riflessione etica che abilita uno sviluppo sostenibile e consapevole [1].

La governance dell'AI nella PA non si traduce necessariamente in una soluzione puramente tecnica, ma piuttosto nella creazione di uno spazio dove le considerazioni antropologiche ed etiche diventano forze motrici per l'innovazione tecnologica e lo sviluppo umano. Le pubbliche amministrazioni devono garantire che i propri sistemi di AI siano conformi alla normativa vigente in materia di protezione dei dati personali (come il GDPR) e di sicurezza cibernetica, oltre che ai principi definiti dalla Carta dei Diritti Fondamentali dell'Unione Europea.

È cruciale che le PA adottino adeguate politiche di gestione del rischio, conducendo

un'analisi approfondita dei rischi associati all'impiego di sistemi di AI. Devono essere consapevoli delle responsabilità e delle implicazioni etiche legate all'uso di queste tecnologie, assicurando che i sistemi rispettino i principi di equità, trasparenza e non discriminazione. Questo include anche la valutazione attenta delle modalità e delle condizioni con cui i fornitori di servizi AI gestiscono i dati forniti dall'amministrazione, con particolare riferimento alla proprietà dei dati e alla conformità normativa.

L'AI Act e le sue implicazioni per la PA

L'AI Act è il primo quadro giuridico al mondo che mira a regolamentare l'intelligenza artificiale. Questo regolamento introduce una classificazione dei sistemi di AI basata sul rischio, imponendo obblighi diversi a seconda del livello di rischio associato a un determinato sistema. I sistemi di AI considerati ad alto rischio, ad esempio, sono soggetti a requisiti più stringenti in termini di qualità dei dati, trasparenza, sicurezza e sorveglianza post-mercato [2].

Per la pubblica amministrazione, l'AI Act comporta diverse implicazioni significative. Le PA devono non solo conformarsi a questi requisiti, ma anche integrare i principi dell'AI Act nelle proprie procedure di sviluppo e acquisizione di soluzioni AI. Questo include la considerazione delle attività di normazione tecnica in corso a livello internazionale ed europeo (come quelle di CEN e CENELEC) e dei requisiti definiti dall'AI Act stesso. L'obiettivo è garantire che i risultati e le decisioni prodotte dai sistemi di AI siano affidabili e coerenti con gli obiettivi prefissati, mantenendo sempre un approccio centrato sull'uomo e sulla tutela dei diritti fondamentali.

La fase di studio preliminare: valutare la fattibilità di un progetto di AI

Per valutare la possibilità di creare un nuovo servizio interno o esterno basato su un modello di AI, un comitato etico sull'uso dell'AI all'interno di una pubblica amministrazione dovrebbe avviare uno studio preliminare. Questo studio è cruciale per gettare le basi di un progetto solido e responsabile. I passaggi chiave includono i seguenti.

Definizione degli obiettivi e delle variabili

Innanzitutto, è essenziale indicare chiaramente gli *obiettivi* dello studio. Questo permette di comprendere se i dati utilizzati sono allineati con l'analisi che si intende effettuare. Vanno inoltre specificate le variabili *proxy* eventualmente utilizzate, ovvero variabili che, pur non rappresentando l'obiettivo diretto, ne costituiscono un indicatore affidabile.

Identificazione dei referenti

È fondamentale identificare i ruoli chiave all'interno del progetto, operando almeno una distinzione tra:

- *Responsabile del dato*: la figura che garantisce la qualità, la disponibilità e la conformità dei dati utilizzati.

- *Responsabile scientifico*: colui che supervisiona l'approccio metodologico e la validità scientifica del progetto.

- *Data Scientist*: il professionista che si occupa dell'analisi dei dati, della costruzione e dell'addestramento dei modelli di AI.

Contesto temporale e selezione dei dati

Le condizioni ambientali e i dati possono mutare nel tempo, quindi è importante indicare chiaramente il periodo di riferimento dell'analisi effettuata. Se alcuni dati sono stati esclusi dall'elaborazione, è necessario indicarne le motivazioni in modo trasparente.

Prima dell'elaborazione vera e propria, occorre fare un'analisi preliminare dei dati. Lo scopo non è solo verificarne la completezza e la coerenza, ma anche comprendere se determinate fasce della popolazione (specialmente quelle considerate più deboli) sono state escluse, affrontando così l'etica del machine learning e il rischio di bias algoritmici. Questa fase deve includere considerazioni di natura etica che indagano sulle cause della bassa rappresentatività di alcune fasce di popolazione.

Metodologie di preprocessing e modelli di AI adottati

Successivamente, si devono illustrare le metodologie di codifica utilizzate prima della fase di elaborazione dati. Va presentato il modello di AI scelto, specificando i parametri di configurazione e, soprattutto, le modalità di addestramento del modello. È cruciale separare sempre il *training set* dal *test set* per evitare il fenomeno dell'*overfitting* (quando il modello impara troppo bene i dati di addestramento e non riesce a generalizzare su nuovi dati).

La scelta del modello di AI utilizzato per l'elaborazione dei dati è ugualmente importante. Non si deve sempre ricorrere all'utilizzo di una stessa tipologia di modello di AI. Ad esempio implementare sempre il modello di elaborazione del dato con un LLM può risultare dispendioso e a lungo andare anche controproducente (essendo affetto come descritto nel relativo capitolo dai problemi del data drift e delle allucinazioni). In alcuni casi si possono ottenere risultati efficaci con algoritmi computazionalmente meno pesanti, ma che si adattano maggiormente alla tipologia di dati trattati ed allo scopo che intende raggiungere la soluzione tecnologica. Se ad esempio si vuole semplicemente effettuare una previsione con un'adeguata accuratezza predittiva per la realizzazione di un Decision Support System (DSS) ed i dati da analizzare sono ben codificati, un'ottima scelta potrebbe essere quella di ricorrere ad un modello subsimbolico, ad esempio una rete autoorganizzante in grado di rispettare con la sua rappresentazione interna della conoscenza la dimensione (topologia) dei dati analizzati come le Growing Neural Gas [3]. Tale modello è stato efficacemente utilizzato per predire la dimissione di pazienti affetti da ictus [4] e per predire il periodo di degenza ospedaliera dei pazienti ricoverati [5]. Ottimi risultati si possono ottenere anche con ulteriori modelli supervisionati di AI computazionalmente più leggeri del LLM come il Percettrone Multistrato [6] o il Random Forest [7], un algoritmo che combina i risultati di più alberi decisionali per raggiungere un unico risultato. Se invece si dispone sempre di basi dati ben codificate e si vogliono individuare le relazioni logiche che legano una variabile dipendente da studiare con tutte le variabili indipendenti da cui potrebbero dipendere, senza conoscere a priori quali siano le possibili correlazioni, un ottimo algoritmo che è stato sviluppato è il modello della Macchina Non Organizzata di Turing (UTM) [8] addestrata con un idoneo algoritmo di ottimizzazione dei parametri di configurazione come un algoritmo genetico [9] o l'Evolutionary Bait Balls Model (EBBM) [10]. Questi algoritmi sono stati testati con successo non solo per predire i periodi di degenza

ospedalieri comprendendo le correlazioni tra le variabili [11], ma anche per rilevare le cause del superamento della soglia di spesa sanitaria pro-capite regionale [12].

Il modello finale addestrato deve essere illustrato, evidenziandone le capacità previsionali e le caratteristiche emergenti. Le performance del modello vanno confrontate con altri modelli di riferimento, incluso il modello 'Zero R', che effettua sempre la previsione più frequente. Se l'accuratezza del modello di AI risulta inferiore a quella del modello 'Zero R', significa che il modello di AI non è stato correttamente configurato o che i dati non sono sufficienti.

Sostenibilità ambientale e ottimizzazione

Una volta ottimizzato l'algoritmo, la fase successiva è la stima della *carbon footprint* dell'algoritmo stesso. Le pubbliche amministrazioni devono valutare attentamente gli impatti ambientali ed energetici legati all'adozione di tecnologie di AI e adottare soluzioni sostenibili dal punto di vista ambientale. Una buona piattaforma per stimare la carbon footprint dei propri algoritmi è stata pubblicata all'indirizzo: https://calculator.green-algorithms.org/ [13].

Infine l'AI può essere adottata nell'automazione dei compiti ripetitivi connessi ai servizi istituzionali obbligatori e al funzionamento dell'apparato amministrativo, con il conseguente recupero di risorse destinato al miglioramento della qualità dei servizi, anche mediante meccanismi di proattività intesa come la capacità dei sistemi e dei servizi di anticipare le esigenze dei cittadini e delle imprese, fornendo informazioni o erogando servizi prima ancora che l'utente ne faccia esplicita richiesta.

Un esempio concreto in ambito sanitario: progetto di un sistema per analizzare le principali cause degli incidenti domestici più gravi

Per illustrare come i principi e le metodologie discusse in questo capitolo possano essere applicati nella pratica, consideriamo il caso dell'avvio di un possibile studio sulle cause che portano al verificarsi degli incidenti domestici più gravi. Per gli obiettivi di progetto si fa riferimento agli obiettivi strategici dei piani regionali di prevenzione 2020-2025 [14], le fonti dati sono prese dai flussi del Sistema Informativo per il monitoraggio delle prestazioni erogate nell'ambito

dell'assistenza sanitaria in emergenza-urgenza (EMUR) [15].

Descrizione degli Obiettivi Strategici di riferimento: MO3OS01: Migliorare la conoscenza della percezione dei rischi in ambito domestico nei genitori e nelle categorie a rischio (bambini, donne e anziani); MO3OS02: Sensibilizzare la popolazione generale sui rischi connessi agli incidenti domestici; MO3OS03: Coinvolgere in modo trasversale i servizi di interesse sanitari e sociali per il contrasto del fenomeno nella comunità.

Obiettivo del gruppo di lavoro: individuare i fattori critici correlati con gli incidenti domestici GRAVI e con gli episodi di aggressione ed autolesionismo (corrispondenti ai valori 1,2,4 del campo "Trauma") verificatosi sul territorio regionale per facilitare l'individuazione degli stakeholder da coinvolgere nella pianificazione delle strategie di azione per contrastare il fenomeno.

Fonte Dati del fenomeno analizzato: flusso EMUR, sezione Pronto Soccorso.

Referente della fonte dati: ****

Referente scientifico: ****

Data scientist: ****

Versione dati: dati aggiornati al ****

Variabile Proxy: nodo di riferimento: Dimissione; nome campo: Livello Appropriatezza Accesso. Il modello di AI deve prevedere gli esiti "R" (Rosso, molto critico) e "N" (Nero, deceduto) di tale campo assieme ai valori "1", "2" e "3".

Tabella delle variabili indipendenti prese in considerazione dal modello:

Nodo di riferimento	Nome campo	Descrizione	Tipo[6]	Metodo di codifica e motivazione della scelta	Grado di Completezza
Entrata	Data	Indicazione del giorno e dell'ora di arrivo al PS	D	Si è scelto di suddividere l'anno di riferimento in quattro trimestri, il giorno di arrivo quindi viene classificato con il codice indicante il trimestre di riferimento. L'ora di arrivo costituisce un'altra variabile indipendente e viene suddivisa in quattro categorie	100%

[6] A=stringa di caratteri o singoli caratteri; AN=stringa di caratteri e numeri; N=numerico; D=data;

				corrispondenti alle fasce orarie 00:00-06:59, 07:00-13:59, 14:00-20:59, 21:00-23:59. Anche l'ora viene classificata con il codice di una delle quattro fasce orarie identificate.	
Accesso	Modalità Arrivo	Indicazione della modalità di arrivo "fisica" al PS	N	8 possibili valori da codificare con una stringa da 8 bit. Questa informazione potrebbe essere utile per comprendere se la modalità di trasporto al PS scelta abbia provocato dei ritardi nel fornire una prima assistenza portando a un peggioramento del quadro clinico	100%
Accesso	Responsabile Invio	Indica il responsabile dell'invio del cittadino al Pronto Soccorso	N	8 possibili valori da codificare con una stringa da 8 bit. Questa informazione potrebbe essere utile per comprendere se il gateway scelto dall'assistito abbia provocato dei ritardi nel fornire una prima assistenza portando a un peggioramento del quadro clinico	100%
Accesso	Problema Principale	Indica il problema principale riscontrato/percepito al momento del triage	AN	Sono 31 possibili valori riportati nell'allegato E del flusso EMUR da codificare con una stringa di 31 bit. Questa informazione ci permette di comprendere in quale stato si è presentato l'assistito all'accesso al PS	100% ma non può essere utilizzato ai fini dello studio in quanto valorizzato sempre a "Trauma" (10)
Accesso	Trauma	Indica la tipologia di trauma rilevato	N	Non va codificato, questo campo ci permette di filtrare gli incidenti domestici (4) le aggressioni (1) e gli	100%

				episodi di autolesionismo (2)	
Accesso	Triage	Livello di urgenza assegnato all'assistito e quindi di priorita per la visita medica assegnata al paziente	AN	1 = Rosso - EMERGENZA 2 = Arancione - URGENZA 3 = Azzurro - URGENZA DIFFERIBILE 4 = Verde - URGENZA MINORE 5 = Bianco - NON URGENZA	100%
Dati Anagrafici	Genere	Indica il sesso dell'assistito	N	3 valori da codificare con 3 bit. Serve per comprendere se ci sono correlazioni tra gli incidenti domestici e la variabile sesso, potrebbe servire anche per comprendere se vi siano discriminazioni di genere	100% (solo un record è classificato come sconosciuto "9")
Dati Anagrafici - Età	Presunta	Indica la fascia di età, anche apparente del paziente. Il campo deve essere compilato in caso di non disponibilità dell'informazione relativa all'anno di nascita	N	7 valori da compilare con 7 bit. Serve per comprendere se ci sono correlazioni tra gli incidenti domestici e la variabile età, potrebbe servire anche per comprendere se vi siano discriminazioni basate sull'età. Si è scelto di operare una riclassificazione dell'età suddividendola nelle seguenti 11 classi: 0-2; 3-5; 6-13; 14-17; 18-25; 26-34; 35-49; 50-64; 65-74; 75-84; 85+.	100%
Dati Anagrafici	Cittadinanza	Indica la cittadinanza dell'assistito	A	La codifica ISO 3166 della nazionalità prevede 249 possibili valori. Andrebbe limitata la casistica prevedendo un valore "altro". Serve per comprendere se ci sono correlazioni tra gli incidenti domestici	61% dei record sono classificati come sconosciuti "XX")

				e la variabile cittadinanza, potrebbe servire anche per comprendere se vi siano discriminazioni basate sull'origine etnica	
Residenza	Comune	Indica il Comune di residenza dell'assistito	AN	Anche in questo caso si suggerisce un raggruppamento per provincia ovvero per ASL. Serve per comprendere se ci sono correlazioni tra gli incidenti domestici e la residenza, potrebbe servire anche per comprendere se vi siano sperequazioni territoriali	99% (l'1% dei record è classificato come "99999", meglio ricorrere ad ASL di residenza che viene sempre valorizzata)
Diagnosi	Principale	Indica la diagnosi principale (la più importante per gravità clinica ed impegno di risorse)	AN	Viene specificato il codice ICD9 CM ultima versione. Andrebbero considerati solo i primi 3 digit del codice che rappresentano la categoria di problema principale.	93% (Il 7% dei record è classificato come sconosciuto "XX")
Accesso	Istituto Provenienza	Identificativo dell'istituto di ricovero inviante. Se l'istituto di destinazione e uno stabilimento di una struttura con più stabilimenti e necessario indicare il codice stabilimento ex HSP.11bis, che ha lunghezza di 8 caratteri. Se l'istituto di destinazione e una struttura monostabilimento e necessario utilizzare il codice struttura HSP.11, che ha lunghezza di 6 caratteri.	AN		1%. Non può essere utilizzato come variabile indipendente

Periodo di riferimento ed altri criteri di selezione dei dati: Si è deciso di prendere in considerazione il periodo dal *** al ***.

Modalità di filtraggio, codifica ed elaborazione dei dati: Per poter ottimizzare l'efficacia dell'analisi dei dati disponibili sono stati eliminati tutti i record dove la cittadinanza veniva classificata come sconosciuta (XX) e dove il mezzo di arrivo non era stato registrato (9). Si è scelto poi di concentrare l'attenzione solamente sugli incidenti avvenuti in ambito prettamente domestico, escludendo i casi provenienti dai centri di detenzione.

```
RangeIndex: *** entries, 0 to ***
Data columns (total 11 columns):
#   Column              Non-Null Count  Dtype
---  ------              --------------  -----
0   ModalitaArrivo      *** non-null  int64
1   ResponsabileInvio   *** non-null  int64
2   Triage              *** non-null  int64
3   Genere              *** non-null  int64
4   PeriodoAnno         *** non-null  int64
5   PeriodoGiorno       *** non-null  int64
6   FasciaEta           *** non-null  int64
7   Diagnosi            *** non-null  int64
8   Cittadinanza        *** non-null  int64
9   ExZonaTerritoriale  *** non-null  int64
10  Outcome             *** non-null  int64
dtypes: int64(11)
```

Si segnala un bias[7] legato alla fase progettuale del modello di intelligenza artificiale. Nello specifico un LABEL BIAS dovuto alla fonte dati scelta che prevede solo i valori M, F per il genere, considerati i gruppi tradizionali di pazienti. Questo porta ad escludere la fascia di popolazione che non rientra nelle due categorie.
E' stata successivamente effettuata un'analisi sulle cittadinanze associate ai casi di incidenti domestici gravi a seguito della quale si è scelto di scegliere le seguenti classi:

Cittadinanza	Codice Interno	Descrizione
IT	1	ITALIA
AL	2	ALBANIA
RO	3	ROMANIA
ZZ	4	APOLIDE
MA	5	MAROCCO
PK	6	PAKISTAN
BD	7	BANGLADESH
UA	8	UCRAINA
NG	9	NIGERIA
PE	10	PERU'
MD	11	MOLDAVIA
PL	12	POLONIA
IN	13	INDIA
CN	14	CINA
	15	ALTRO

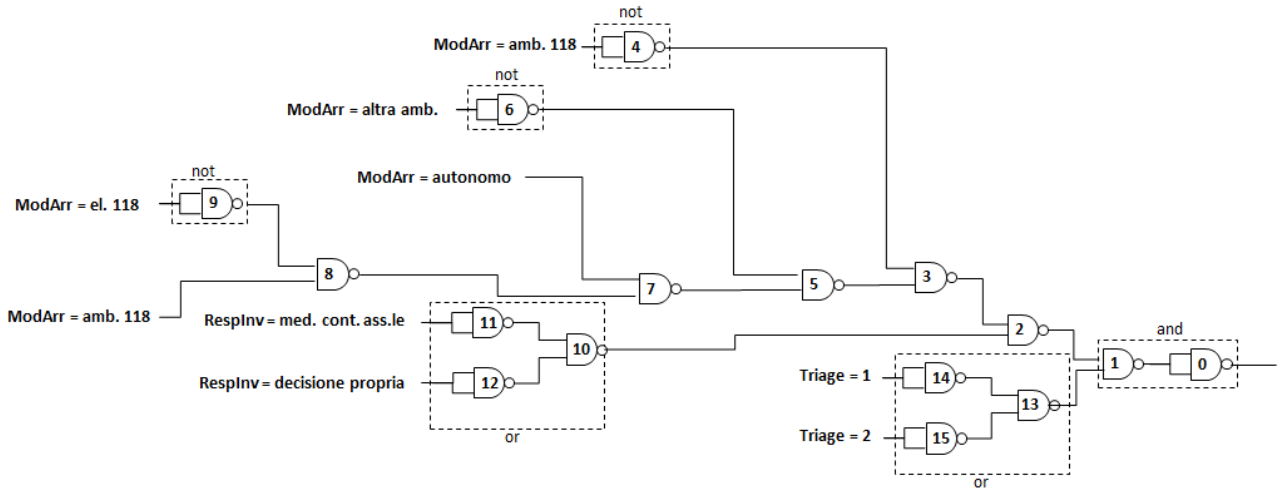
[7] Giovanola B., Tiribelli S., *Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms*, AI & SOCIETY, vol.38, pp.549-563, 2023

Dei *** casi positivi, *** rappresentano casi di cittadini italiani (94%). La percentuale di casi positivi associata ai cittadini stranieri è quindi leggermente sottostimata (6%), dal momento che secondo quanto emerso nel Profilo di Salute della Regione *** gli stranieri costituiscono circa l'8,6% della popolazione totale. E' da escludere però un bias di tipo progettuale come il precedente (ovvero un COHORT BIAS) in quanto gli italiani rappresentano il 92% dei casi totali (compresi i negativi) e gli stranieri l'8% dei casi totali. Gli stranieri risultano quindi essere scarsamente rappresentati solo nei casi positivi (MINORITY BIAS). Andrebbe indagato se tale bias è dovuto ad una disparità di trattamento tra cittadini italiani e cittadini stranieri relativamente ai casi più gravi (ad esempio un PRIVILEGE BIAS dovuto a problemi di accesso al servizio di emergenza/urgenza) o se è dovuto al fatto che gli stranieri accedono consapevolmente meno frequentemente ai servizi di PS in presenza di casi gravi (un INFORMED MISTRUST BIAS).
Per quanto concerne le diagnosi principali associate ai casi di incidenti domestici gravi è stata effettuata un'analisi di Pareto che ha portato alla selezione delle seguenti classi:

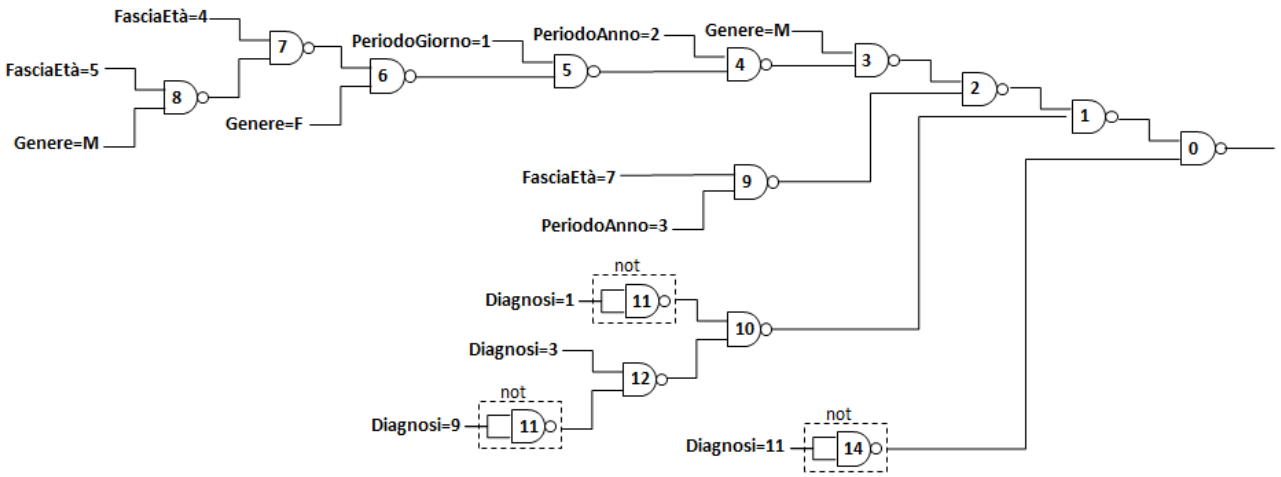
Cod.	Codice interno	Desc. Cod. Principale
821	1	FRATTURE DI ALTRE NON SPECIFICATE PARTI DEL FEMORE
959	2	ALTRI E NON SPECIFICATI TRAUMATISMI
820	3	FRATTURA DEL COLLO DEL FEMORE
829	4	FRATTURE DI OSSA NON SPECIFICATE
850	5	CONCUSSIONE
338	6	DOLORE NON CLASSIFICATO ALTROVE
807	7	FRATTURA DELLE COSTOLE, DELLO STERNO, LARINGE E TRACHEA
780	8	SINTOMI GENERALI
879	9	FERITE DI ALTRE E NON SPECIFICATE SEDI, ESCLUSI GLI ARTI
812	10	FRATTURA DELL'OMERO
431	11	EMORRAGIA CEREBRALE
799	12	ALTRE CAUSE MAL DEFINITE DI MORBOSITA' O MORTALITA'
733	13	ALTRI DISTURBI DELLE OSSA E CARTILAGINI
873	14	ALTRE FERITE DELLA TESTA
883	15	FERITA DELLE DITA DELLA MANO
923	16	CONTUSIONE DELL'ARTO SUPERIORE
958	17	ALCUNE COMPLICAZIONI PRECOCI DI TRAUMATISMI
805	18	FRATTURA COLONNA VERTEBRALE SENZA LESIONE MIDOLLO SPINALE
924	19	CONTUSIONE ARTO INFERIORE E ALTRI NON SPECIFICATE SEDI
869	20	TRAUMATISMO INTERNO DI ORGANI NON SPECIFICATI
813	21	FRATTURA DEL RADIO E DELL'ULNA
808	22	FRATTURA DEL BACINO
920	23	CONTUSIONE FACCIA, CUOIO CAPELLUTO E COLLO (ESCLUSO OCCHIO)
823	24	FRATTURA DELLA TIBIA E DEL PERONE
802	25	FRATTURA DELLE OSSA DELLA FACCIA
432	26	ALTRE NON SPECIFICATE EMORRAGIE INTRACRANICHE
922	27	CONTUSIONE DEL TRONCO
882	28	FRATTURA DELLA MANO ESCLUSE LE DITA DA SOLE
995	29	ALCUNI EFFETTI AVVERSI NON CLASSIFICATI ALTROVE
831	30	LUSSAZIONE DELLA SPALLA
V48	31	PROBLEMI RELATIVI ALLA TESTA, AL COLLO O AL TRONCO

824	32	FRATTURA DELLA CAVIGLIA
753	33	ANOMALIE CONGENITE SISTEMA URINARIO
951	34	TRAUMATISMO ALTRI NERVI CRANICI
910	35	TRAUMATISMO SUPERFICIALE FACCIA, CUOIO CAPELLUTO E COLLO (ESCLUSO OCCHIO)
881	36	FERITA DEL GOMITO, DELL'AVAMBRACCIO E DEL POLSO
886	37	AMPUTAZIONE TRAUMATICA DELLE ALTRE DITA DELLA MANO
427	38	ARITMIE CARDIACHE
	39	ALTRO

Si è scelto di addestrare con una convalida incrociata a 10 parti (10-fold cross validation) il modello EBBM-UTM per 100.000 epoche utilizzando i seguenti parametri di configurazione: individualSize=266; attractionRate=0.1; repulsionRate=0.5; ZOArange=266; ZOOrange=5; ZORrange=0.
Il modello di AI addestrato ha preso in considerazione per effettuare le sue previsioni solamente 3 variabili indipendenti ovvero l'**esito del Triage**, la **modalità di arrivo** ed il **responsabile dell'invio**:
Il modello è riuscito ad elaborare regole logiche con un discreto livello di complessità (compaiono gli equivalenti di 2 porte OR, 1 porta AND e 3 porte NOT) e senza ricorrere ad alcun tipo di euristica è riuscito a comprendere che il solo esito del Triage era più che sufficiente per comprendere la gravità del caso. Se infatti la variabile Triage è differente da 1 o 2 il caso non è MAI critico. Esistono però dei casi in cui la variabile Triage è uguale ad 1 o 2 ed il responsabile dell'invio è la guardia medica o lo stesso assistito (decisione propria). In questa situazione il modello stabilisce che il caso può considerarsi grave solo se il mezzo di arrivo è l'ambulanza del 118 o il mezzo proprio. In tutti gli altri casi in cui Triage è uguale a 1 o 2 il caso è SEMPRE critico.



L'esiguità dei casi critici presenti nel dataset di addestramento non permette di ottenere un modello predittivo da poter utilizzare con efficacia nelle attività operative presso le strutture di Emergenza Urgenza. Anche se l'accuratezza predittiva risulta essere abbastanza elevata (83.77%), il valore della sensibilità ricavabile a partire dai veri positivi e dai falsi negativi (VN/(VN+FP)) è molto basso (0.37), considerato poi il punteggio di prevalenza pari a 0.19 si arriva a un valore predittivo positivo pari a 7.45 punti percentuali.
Il modello è comunque valido per effettuare utili considerazioni sui fattori correlati con i casi di incidenti domestici giudicati più critici. Le variabili indipendenti selezionate dal modello di AI non permettono però di trarre utili indicazioni per un eventuale piano di prevenzione degli incidenti domestici. Per questo motivo si è scelto di ripetere l'analisi escludendo le variabili **Triage**, **modalità di arrivo** e **responsabile dell'invio** dall'insieme di variabili indipendenti prese in considerazione dal modello di AI, arrivando al termine della fase di addestramento alla seguente rete logica combinatoria:



Si è scelto di addestrare con una convalida incrociata a 10 parti (10-fold cross validation) il secondo modello EBBM-UTM per 50.000 epoche utilizzando i seguenti parametri di configurazione: individualSize=232; attractionRate=0.5; repulsionRate=0.5; ZOArange=232; ZOOrange=5; ZORrange=0.
Il modello di AI continua ad escludere le variabili indipendenti della cittadinanza e della ASL di afferenza dell'assistito, che quindi sembrerebbero non essere correlate con il grado di gravità dei casi considerati.
La regola stabilisce che se la diagnosi principale è di **emorragia cerebrale** (11) viene giudicato dal modello di AI SEMPRE critico. Altrimenti possono risultare critici i casi di **frattura di altre non specificate parti del femore** o fratture del collo del femore (Diagnosi=1 o Diagnosi=3), specie quando il paziente ha un'**età compresa tra i 50 e i 64 anni** e l'incidente domestico avviene **nel periodo tra Luglio e Settembre** (PeriodoAnno=2). I casi di frattura del femore risultano critici anche quando il paziente è anziano (sempre tra i 50 e i 64 anni), di **genere maschile** e l'incidente domestico avviene **nel periodo tra Aprile e Giugno** (PeriodoAnno = 2) oppure **tra le ore 0:00 e le ore 6:59 del giorno** (PeriodoGiorno=1).
Anche in questo caso pur raggiungendo un'accuratezza predittiva di 82%, la sensibilità scende a 0.05 e con essa anche il valore predittivo positivo pari ad appena 1.25 punti percentuali.
Per entrambi i modelli si registra in ogni caso un'accuratezza predittiva superiore a quella conseguibile dal modello "ZeroR" (Zero Rules) ovvero il semplice modello predittivo che effettua sempre la previsione più frequente (avente un'accuratezza predittiva pari a 81.12%). Questo avvalora le modalità di configurazione e di addestramento dei due modelli di AI.
Dalle successive verifiche effettuate risulta effettivamente che **più dell'84% delle emorragie cerebrali** viene classificato come un caso critico, che **più del 94% delle fratture al femore in pazienti di sesso maschile nel secondo periodo dell'anno** viene classificato come caso critico e che **più dell'88% delle fratture al femore in pazienti di sesso maschile nel primo periodo della giornata** viene classificato come critico.

Questo è un esempio di come andrebbe documentata l'implementazione di una soluzione informatica basata su un modello di IA che non rientra nei sistemi classificati come "ad alto rischio", per i quali l'AgID ha definito delle chiare metodologie di valutazione del rischio all'interno delle linee guida per l'adozione di IA nella pubblica amministrazione. Nei sistemi classificabili come a "rischio limitato" o "a rischio minimo o nessun rischio" come quello che è stato presentato[8], viene

[8] Il sistema di IA per analizzare le cause di incidenti domestici più gravi che è stato presentato è un classico esempio di Decision Support System che prendendo come input dati sanitari e amministrativi anonimizzati restituisce un'analisi delle correlazioni tra i dati disponibili. L'uso che ne viene fatto è di supporto indiretto alla decisione, con riscontro finale lasciato interamente al decisore umano. In sistema non implementa direttamente una politica, non prende decisioni vincolanti, non interagisce autonomamente con cittadini, non determina in automatico diritti o servizi. Questo lo

consigliata un'analisi semplificata che può essere condotta nella maniera appena descritta, la quale non solo evidenzia obiettivi, referenti di progetto e caratteristiche del modello di AI utilizzato, ma ne valuta anche le implicazioni etiche andando ad analizzare la presenza di bias all'interno degli stessi dati con cui è stato addestrato il sistema. Un report preliminare strutturato come quello riportato in questo capitolo permette di fornire al Comitato Etico sull'uso della AI interno all'organizzazione tutti gli elementi utili per consentirgli di prendere una decisione chiara e definitiva in merito all'adozione della nuova soluzione di AI sperimentata.

Per quanto riguarda la carbon footprint dell'algoritmo scelto per l'elaborazione dei dati, si è considerato che il calcolo è stato effettuato ricorrendo ad un semplice laptop Intel I7 con 16 GB di RAM, ricorrendo solamente alla CPU ed escludendo la GPU. Per ottimizzare il risultato finale sono bastate 6 ore di elaborazione. Il carbon footprint calcolato dalla piattaforma di valutazione è risultato equivalente ad appena 41,89 grammi di CO2 equivalente (corrispondenti all'inquinamento prodotto in media da un'automobile che percorre 0,24 km di strada) contro le circa 300 tonnellate di CO2 e gli oltre 700.000 litri d'acqua per il raffreddamento richiesti per l'addestramento di un LLM come GPT-4. Si consideri che una semplice generazione di un brevissimo testo (come quello di un'e-mail) con il modello GPT-4 addestrato consuma circa 3-4 grammi di CO2 equivalente e 0,3-0,5 Wh di energia. Questo dovrebbe spingere i tecnici informatici ed il personale amministrativo della pubblica amministrazione a ricorrere ai LLM solo quando effettivamente richiesto, ad esempio per compiti di comprensione del linguaggio naturale (NLP), generazione creativa, e gestione di dati non strutturati, evitando di ricorrere a tali strumenti per la generazione di brevi contenuti testuali.

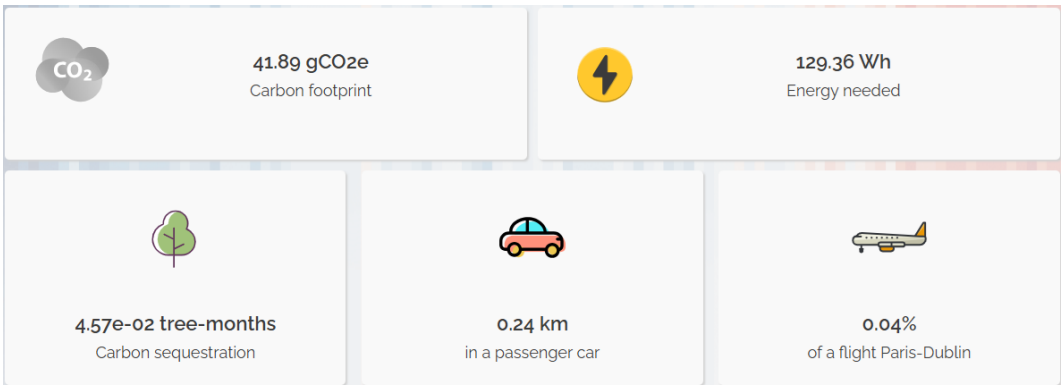


Figura 16.1 – Carbon footprint dell'algoritmo EBBM-UTM utilizzato nello studio delle cause degli incidenti domestici.

differenza nettamente dai sistemi AI "ad alto rischio" che intervengono in contesti critici e vincolanti (es. assegnazione di sussidi, gestione di infrastrutture sanitarie, screening automatizzato di pazienti etc.).

L’importanza dello schema dati e la progettazione dell’interfaccia per la raccolta dati

Spesso, quando non sono disponibili dati strutturati o non strutturati preesistenti da analizzare, si presenta una magnifica occasione per individuare e definire un nuovo schema dati. Tuttavia, la sua efficacia dipende intrinsecamente dalla capacità di alimentarlo in maniera corretta, completa e aggiornata nel tempo con il coinvolgimento dell’utenza dei servizi dell’amministrazione pubblica.

Si consideri il caso della predisposizione di una soluzione informatica per il monitoraggio a livello regionale o comunale del benessere sociale sostenibile. Tale monitoraggio deve innanzitutto basarsi su un modello dati (schema) predisposto per monitorare adeguatamente l'oggetto dell'analisi. Questo schema non è una mera formalità tecnica, ma il fondamento su cui si costruirà tutta l'analisi e l'estrazione di conoscenza. La sua progettazione deve essere guidata dagli obiettivi dello studio e dalla natura del fenomeno osservato.

Nel caso del monitoraggio del benessere sociale, modelli psicologici come la piramide dei bisogni di Maslow o il modello della "sailboat" di Kaufman si rivelano particolarmente utili. Questi modelli offrono una struttura concettuale per categorizzare e quantificare i bisogni e le aspirazioni dei cittadini, permettendo di rilevare non solo le carenze percepite (D-Needs), ma anche i desideri esistenziali e le aspirazioni di crescita (B-Needs). L'adozione di tali modelli consente di attribuire un "punteggio" alle problematiche segnalate, permettendo di ordinarle per grado di importanza e di monitorare nel tempo l'andamento medio del benessere sociale della popolazione [16].

È cruciale che lo schema dati sia in grado di catturare la complessità del fenomeno, evitando semplificazioni eccessive che potrebbero portare a una perdita di informazioni significative. Ad esempio, il modello di Kaufman, che distingue tra D-Needs e B-Needs, offre una visione più completa rispetto al cosiddetto modello della piramide dei bisogni di Maslow, consentendo di analizzare l'evoluzione del profilo dei bisogni della popolazione e la sua sensibilità verso diverse tematiche (ambientali, sociali, ecc.).

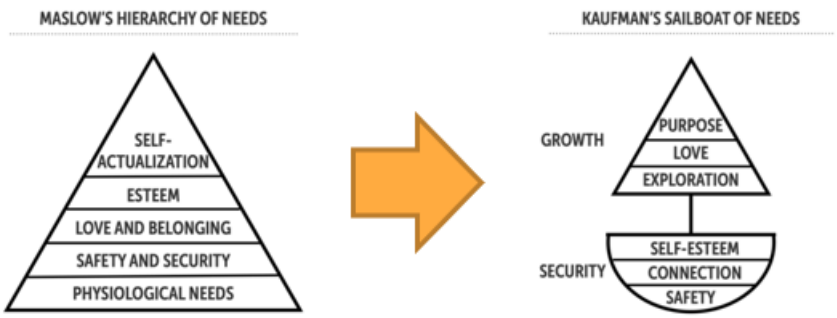


Figura 16.2 – Schema dati considerati per modellare il benessere sociale sostenibile rilevato.

Il "touchpoint digitale" deve essere concepito come un vero e proprio canale di ascolto del cittadino, una nuova fonte dati che fornisca ai decisori politici elementi aggiuntivi per guidare i loro processi decisionali [17]. Questo può includere:

- *Questionari interattivi*: Basati sugli schema dati scelti, con la possibilità di estendere la compilazione con sottomoduli attivati da trigger specifici per un feedback più dettagliato o che rimandano a altri e-services attivati dall'amministrazione pubblica.

- *Talking head o video per comunicazioni di benvenuto/aggiornamento*: Per illustrare le iniziative della PA e creare un senso di accoglienza.

- *Sistemi esperti*: Per fornire indicazioni personalizzate al termine della compilazione, come contatti utili o suggerimenti.

- *Chatbot intelligenti basati su IA generativa*: Per un dialogo bidirezionale e per rispondere a domande frequenti, migliorando l'interazione e l'usabilità.

L'accessibilità e l'usabilità dell'interfaccia sono fondamentali. Il touchpoint deve essere accessibile in vari modi (es. tramite motore di ricerca interno/esterno, app di realtà aumentata) e attraverso diversi dispositivi (smartphone, tablet, PC). Deve rispettare le linee guida AgID sull'usabilità e accessibilità dei servizi digitali e deve essere costantemente migliorato, adottando un approccio Agile allo sviluppo.

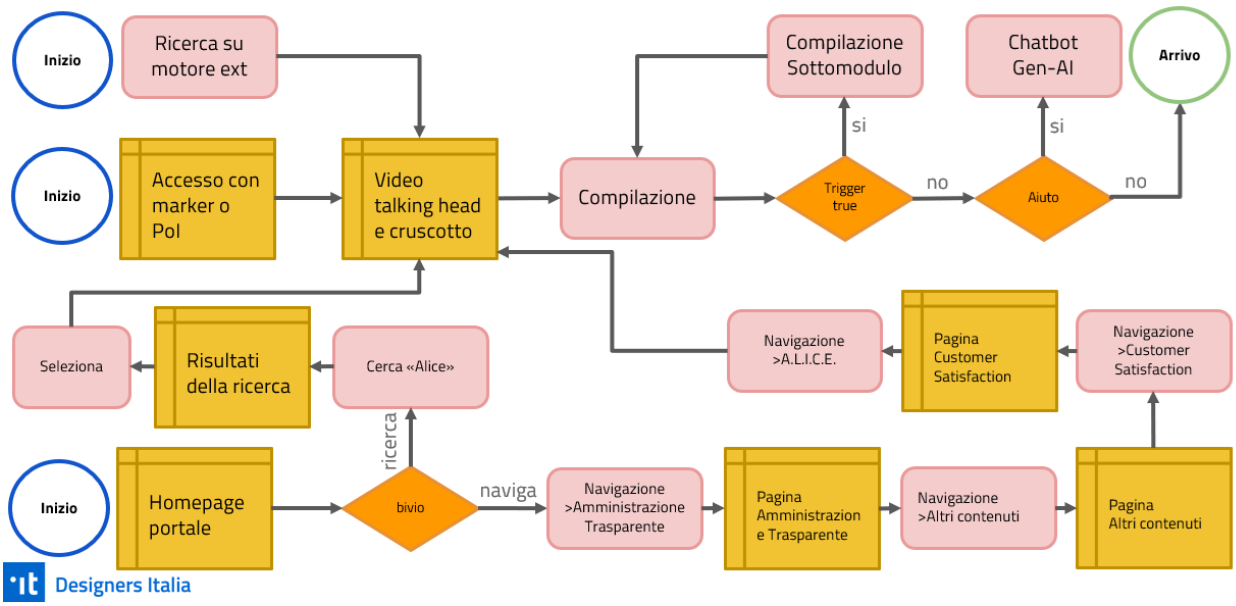


Figura 16.3 – Flusso di interazione del touchpoint del cittadino del sistema "A.L.I.C.E." Copyright © 2023 Luigi Lella per il monitoraggio del benessere sociale sostenibile. Il diagramma è stato realizzato ricorrendo al template definito per il progetto Designers Italia e distribuito con licenza Creative Commons Attribuzione – Condividi allo stesso modo 4.0 Internazionale, Copyright © 2021 Presidenza del Consiglio dei Ministri – Dipartimento per la Trasformazione Digitale

Quando i dati provengono direttamente dall'utente dei servizi di un'amministrazione comunale (cittadino, impresa, ecc.), la progettazione dell'interfaccia per la raccolta dati diventa un elemento critico. Non si tratta solo di creare un modulo funzionale, ma di realizzare un'applicazione (ovvero il "touchpoint digitale") che sia così appetibile da portare a una massa critica di utenti ovvero al verificarsi di una cosiddetta *esternalità di rete positiva*, assicurando un flusso di aggiornamento costante dei dati.

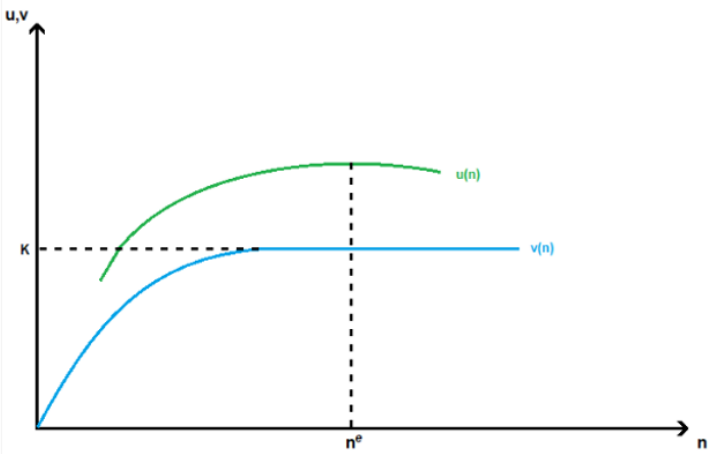


Figura 16.4 – Un'esternalità di rete positiva rappresenta l'utilità u(n) percepita di un e-service dove n è il numero di utenti. In essa è presente una componente «diretta» f(n) legata alla sensazione che l'utilizzo dell'e-service sortisca un effetto collettivo (in questo caso il miglioramento del benessere della società) ed una componente «indiretta» v(n) legata all'aumento dell'utilità personale ovvero al valore dell'e-service (in questo caso il valore dei sottomoduli aggiuntivi del touchpoint): $u(n_e) = K + f(n_e)$

La gestione della raccolta dati non si esaurisce con la predisposizione della sola interfaccia utente. È necessario un sistema robusto che garantisca che i dati siano raccolti in modo corretto, completo e aggiornato nel tempo. Questo implica:

- *Anonimizzazione e Consenso*: Il questionario di base dovrebbe essere anonimo, ma con la possibilità di richiedere il consenso per la compilazione di sottomoduli che potrebbero richiedere dati personali, sempre nel rispetto della normativa sulla privacy (es. GDPR).

- *Gestione dei Bias*: È fondamentale mitigare i bias nella raccolta dati, come il *bias di selezione* (raccogliendo dati solo da persone insoddisfatte) e il *bias di sottocopertura* (escludendo alcune fasce della popolazione). La robustezza dei modelli psicologici come Maslow e Kaufman aiuta a contenere il bias di selezione. Per il bias di sottocopertura, è essenziale che il touchpoint sia inclusivo e che, in caso di bassa rappresentatività di alcune fasce, si indaghino le motivazioni e si aggiorni la campagna di presentazione del servizio o si rivedano gli item del questionario.

- *Campagne di Comunicazione*: Il rilascio di una nuova versione del touchpoint digitale del cittadino deve essere accompagnato da un'adeguata campagna informativa per promuovere l'adozione da parte dei cittadini. L'approccio "Working Backwards" di Amazon [18], che prevede la pubblicazione di comunicati stampa che evidenzino il problema segnalato dall'utenza che è stato risolto e i vantaggi della soluzione, può essere molto efficace per stimolare l'interesse e l'engagement.

- *Ciclo di Feedback e Miglioramento Continuo*: Il touchpoint digitale del cittadino, come ogni altro e-service della PA, deve essere progettato e migliorato nel tempo con il coinvolgimento diretto del cittadino. I feedback raccolti consentono alla PA di valutare l'utilità percepita e di creare le condizioni per l'innesco di esternalità di rete positive.

In sintesi, la scelta di uno schema dati appropriato unitamente alla progettazione di un'interfaccia utente accattivante e di un sistema di raccolta dati efficiente sono passaggi imprescindibili per il successo di un progetto di IA nella pubblica amministrazione, specialmente quando si mira a coinvolgere attivamente i cittadini

e a raccogliere dati direttamente da loro. Questo approccio non solo garantisce la qualità e la continuità dei dati, ma promuove anche l'*empowerment* del cittadino e la creazione di servizi pubblici più efficaci e centrati sull'uomo.

Conclusioni

Isaac Asimov affermava che il valore più profondo della fantascienza non consiste nella previsione degli eventi futuri, ma nella sua capacità di preparare l'umanità ad affrontare il cambiamento. In un celebre saggio[9], lo scrittore sosteneva che "la fantascienza è importante perché ci fornisce un modo per immaginare e riflettere sul cambiamento tecnologico, in modo da poterlo affrontare con consapevolezza". Oggi, più che mai, questa funzione appare urgente.

In questo saggio, oltre a far riferimento a noti romanzi e film di fantascienza, si è scelto di partire proprio da due allegorie – il racconto dei passeri di Bostrom e la nottola di Minerva di Hegel – per offrire un inquadramento narrativo e filosofico del nostro tempo. Entrambe ci mettono in guardia: la prima ci ricorda il rischio di creare un'intelligenza che potrebbe sfuggirci di mano, la seconda che troppo spesso comprendiamo la portata degli eventi solo a posteriori, quando le scelte sono ormai alle nostre spalle.

L'intelligenza artificiale rappresenta oggi il punto d'intersezione tra queste due allegorie. È la "civetta allevata" che potrebbe un giorno volgersi contro i suoi creatori, ma è anche il simbolo della saggezza che giunge tardi, quando ormai il volo è iniziato. In questo contesto, la letteratura di fantascienza, così come il pensiero critico e filosofico, non sono meri esercizi speculativi, ma strumenti culturali essenziali per anticipare i rischi, educare al dubbio e orientare il progresso.

L'intelligenza artificiale rappresenta una delle più grandi rivoluzioni tecnologiche del nostro tempo, promettendo di trasformare ogni aspetto della società umana. Tuttavia, come ogni innovazione di tale portata, porta con sé non solo immense opportunità ma anche sfide profonde e pericoli significativi. La corsa globale allo sviluppo dell'AI, che vede Stati Uniti, Cina e paesi emergenti in una competizione sempre più serrata, solleva interrogativi cruciali sugli aspetti etici, sulla governance e sull'impatto a lungo termine sul nostro pianeta. La pressione competitiva per accelerare lo sviluppo dell'AI sta portando a un approccio di deregolamentazione,

[9] Questa citazione è tratta dall'introduzione all'Encyclopedia of Science Fiction (1974), dove Asimov argomenta che lo scopo della fantascienza non è tanto predire eventi specifici, quanto sensibilizzare l'umanità a comprendere e governare il cambiamento tecnologico in modo responsabile.

come evidenziato dal "DOGE AI Deregulation Decision Tool" dell'amministrazione Trump. Questo strumento, progettato per eliminare una vasta percentuale di regolamentazioni federali americane, solleva serie preoccupazioni riguardo all'accuratezza e alla legalità delle decisioni prese dall'AI stessa. La deregolamentazione, sebbene presentata come un mezzo per promuovere l'innovazione e la crescita economica, può avere effetti perversi. L'approccio americano, influenzato da valori libertari, tende a una regolamentazione minima. Questo contrasta con l'approccio europeo, che cerca di stabilire un quadro etico più rigoroso, come dimostrato dall'AI Act. La mancanza di un consenso etico globale sui valori e sui rischi dell'AI rende la deregolamentazione una strada pericolosa, poiché i pericoli creati dal comportamento non regolamentato dell'AI potrebbero superare i benefici percepiti.

La competizione per la leadership nell'intelligenza artificiale è diventata una delle sfide geopolitiche più significative del XXI secolo, con Stati Uniti e Cina che si contendono il primato. Questa "guerra fredda tecnologica" non si limita solo allo sviluppo di algoritmi avanzati, ma si estende anche alla costruzione di infrastrutture computazionali massicce, in particolare i data center. I paesi emergenti, cercando di non rimanere indietro, sono anch'essi coinvolti in questa corsa, spesso con l'aiuto delle superpotenze che cercano di estendere la loro influenza tecnologica.

Il piano d'azione per l'AI dell'amministrazione Trump, ad esempio, prevede investimenti ingenti per accelerare la costruzione di data center, anche su terreni federali, e per potenziare le reti elettriche con un ritorno a fonti energetiche come il nucleare e il geotermico. Questo approccio, sebbene mirato a consolidare la leadership statunitense, evidenzia il crescente fabbisogno energetico dell'AI. I data center, che già oggi rappresentano circa l'1% del consumo globale di elettricità, sono destinati a vedere un aumento esponenziale del loro consumo energetico, con proiezioni che indicano che l'AI potrebbe rappresentare quasi la metà del consumo energetico dei data center entro la fine del decennio.

La Cina, pur proponendo un quadro unificato di governance globale per l'AI, sta anch'essa investendo massicciamente in infrastrutture AI, inclusi data center sottomarini che uniscono AI, energia rinnovabile e sostenibilità ambientale. Tuttavia, la rapidità con cui queste infrastrutture vengono implementate, spesso proprio con poca attenzione alle normative ambientali o alla pianificazione a lungo termine, può portare invece a conseguenze ecologiche irreversibili.

La competizione globale, quindi, non solo spinge verso una deregolamentazione etica, ma anche verso un consumo di risorse che potrebbe avere un impatto devastante sul clima e sull'ambiente. Questa competizione, unita a una visione miope che privilegia la velocità sull'etica e la sostenibilità, rischia di farci perdere di vista il quadro generale. La nottola di Minerva, simbolo della saggezza che emerge solo alla fine di un ciclo, ci ammonisce a riflettere profondamente prima che sia troppo tardi. Dobbiamo chiederci se la nostra corsa allo sviluppo dell'AI stia creando un futuro in cui i benefici sono riservati a pochi, mentre i costi etici e ambientali ricadono su tutti.

In questo scenario, la responsabilità non ricade solo sui governi e sulle grandi aziende, ma anche sugli sviluppatori di sistemi di AI e sugli informatici che operano all'interno delle pubbliche amministrazioni. Essi hanno il dovere di mantenersi costantemente aggiornati sugli sviluppi tecnologici, sulle capacità, sui limiti e sui rischi associati ai modelli di AI. La loro etica professionale li impegna a creare le condizioni per un utilizzo ottimale e sicuro dell'AI, mettendola al servizio della società e promuovendo l'empowerment del cittadino. Come la civetta di Bostrom è pronta per essere lasciata libera, così l'AI è pronta a spiccare il volo, ma è imperativo che venga addestrata e che si creino le condizioni affinché non rechi danno alla comunità. È altresì fondamentale che la comunità internazionale collabori per stabilire un quadro di governance globale che sia etico, inclusivo e sostenibile, garantendo che l'AI spicchi il volo non come un predatore incontrollato, ma come uno strumento al servizio dell'umanità intera, consapevole delle sue responsabilità e del suo impatto sul mondo.

Se vogliamo evitare che la storia ci sfugga di mano, dobbiamo recuperare il valore della narrazione come forma di comprensione collettiva e di responsabilizzazione. Come i grandi romanzi distopici hanno spesso anticipato le derive della tecnica senza etica, così oggi abbiamo bisogno di racconti capaci di farci immaginare alternative sostenibili, inclusive e giuste. L’AI, dopotutto, non è un destino inevitabile: è un progetto umano, e in quanto tale può ancora essere guidato. Sta a noi scegliere se limitarci a osservare, con lo sguardo di Hegel rivolto al crepuscolo, o se provare a comprendere in anticipo, con la lungimiranza di chi ha imparato, anche grazie alla fantascienza ed alla filosofia, a usare la tecnologia come mezzo per l’emancipazione e non per la sottomissione.

Biografia dell'autore

L'Ing. Luigi Lella ricopre attualmente il ruolo di Titolare di Alta Specializzazione presso il Servizio Informatica, Innovazione e Transizione Digitale (SIITD) del Comune di Ancona. In questo incarico si dedica attivamente all'attuazione del progetto organizzativo del Servizio, con un focus sulla digitalizzazione e lo sviluppo sostenibile nel contesto locale e nazionale.

La sua carriera professionale è caratterizzata da una solida base nel settore delle tecnologie dell'informazione. Prima del suo attuale incarico, ha lavorato come consulente informatico presso varie aziende tra cui la ITConsult, azienda privata specializzata nella produzione di piattaforme per la gestione della conoscenza e per la gestione dei processi, con focus su algoritmi di datamining, sistemi di apprendimento collaborativo online, e analisi testuale. Ha operato come collaboratore tecnico professionale presso l'Ex Azienda Sanitaria Unica Regionale delle Marche (ASUR Marche), dove si è occupato di progettazione e realizzazione di applicazioni web, sistemi previsionali basati su algoritmi di machine learning, oltre a gestire attività di formazione interna anche in tema di AI. Ha svolto incarichi di collaborazione con l'Agenzia Regionale Sanitaria delle Marche (ARS) sviluppando sistemi di raccolta ed elaborazione dati sempre basati su algoritmi di machine learning. Ha anche svolto attività di libera docenza presso l'Università Politecnica delle Marche, tenendo un corso sui sistemi informatici. Dal novembre 2002 al novembre 2004, ha ricoperto il ruolo di Ricercatore Junior e Collaboratore Didattico presso il Dipartimento di Elettronica Informatica e Telecomunicazioni (DEIT) dell'Università Politecnica delle Marche, conducendo ricerche su intelligenza artificiale e sistemi dinamici complessi e supportando corsi di "Intelligenza Artificiale". Durante il periodo di Dottorato di Ricerca ha avuto modo di presentare i propri risultati di ricerca presso l'IT Universitet di Goteborg (Svezia), all'interno del Knowledge Management Hub diretto dal Prof. Dick Stenmark, e presso il centro di ricerca sulla pattern recognition CENATAV de l'Habana (Cuba) per lo sviluppo di algoritmi per il riconoscimento automatico di pattern in collaborazione col gruppo di lavoro sul data mining del centro diretto dal Prof. Jose Ruiz-Shulcloper. Ha avuto modo di collaborare anche con vari ricercatori dell'Associazione Italiana per la Ricerca sui Sistemi e con ricercatori del Laboratorio di Ontologia Applicata del CNR.

La sua formazione accademica include una Laurea Magistrale in Sistemi Socio Sanitari e pubblica amministrazione (2011), un Master in Strategie e Management d'Impresa (2008) presso l'Istituto Adriano Olivetti (ISTAO), e una Laurea in Ingegneria Elettronica (2002) conseguita presso l'Università Politecnica delle Marche. La sua tesi di laurea in Ingegneria era intitolata "Analisi della Semantica nel Web attraverso reti neurali auto organizzanti". Nel 2005 ha ricevuto il Primo Premio "Riccardo Maceratini" per il miglior progetto presentato da giovani ricercatori, grazie a un lavoro sulla profilazione degli operatori sanitari.

L'Ing. Lella è autore di numerose pubblicazioni scientifiche che spaziano dall'informatica sanitaria allo sviluppo di modelli e sistemi di intelligenza sanitaria, dall'analisi dei dati all'estrazione della conoscenza per la pubblica amministrazione. I suoi contributi includono articoli su agendadigitale.eu su temi come i dilemmi etici dell'AI, l'orientamento delle decisioni della PA con dati e AI, e il benessere sociale sostenibile.

Riferimenti bibliografici

Capitolo 1

[1] AgID. *Linee guida sull'intelligenza artificiale*. Disponibile su: https://www.agid.gov.it/it/agenzia/strategia-digitale/intelligenza-artificiale

[2] McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Disponibile su: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

[3] Von Neumann, J. (1945). *First Draft of a Report on the EDVAC*. Moore School of Electrical Engineering, University of Pennsylvania.

[4] Russell, S. J., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (rd ed.). Prentice Hall.

[5] Gartner. *Hype Cycle for Artificial Intelligence*. Disponibile su: https://www.gartner.com/en/articles/what-is-the-gartner- hype-cycle

[6] Floridi L., Cabitza F. (2021). *intelligenza artificiale – L'uso delle nuove Macchine*, Martini Lecture, Bompiani

Capitolo 2

[1] Turing, A. M. (1950). *Computing Machinery and Intelligence*. Mind, 59(236), 433-460.

[2] Searle, J. R. (1980). *Minds, Brains, and Programs*. Behavioral and Brain Sciences, 3(3), 417-457.

[3] Weizenbaum, J. (1966). *ELIZA—a computer program for the study of natural language communication between man and machine*. Communications of the ACM, 9(1), 36-45.

[4] Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Penguin.

[5] OpenAI. (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774. (Il GPQA e l'AIME sono menzionati come benchmark in questo e altri lavori di

OpenAI).

[6] Chollet, F. (2019). *On the Measure of Intelligence*. arXiv preprint arXiv:1911.01547.

Capitolo 3

[1] Kurzweil R., (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking, Penguin Group.

[2] Good I. J. (1965). *Speculations Concerning the First Ultraintelligent Machine*. Inf. L. Alt M. Rubinoff (Eds.), Advances in Computers, Vol. 6. Academic Press.

[3] Brown F., (1954). *La Risposta*. *Le meraviglie del possibile*. Einaudi.

[4] Vinge V. (1993). *The coming technological singularity: How to survive in the post-human era*. NASA. Lewis Research Center, Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace

[5] Kurzweil R.(2024), *La singolarità è vicina. Quando l'umanità supera la biologia.* Feltrinelli.

[6] Lella L. (2023), *I tre punti di singolarità dell'AI generativa: le criticità da affrontare*, articolo pubblicato su Agenda Digitale reperibile all'indirizzo https://www.agendadigitale.eu/cultura-digitale/i-tre-punti-di-singolarita-dellia-generativa-le-criticita-da-affrontare/

[7] Open AI (2025). *Introducing Operator*. Recuperato da: https://openai.com/index/introducing-operator/

[8] Walti C. (2025). *Scetticismo del primo responsabile del progetto Elon Musk: i robot umanoidi come Optimus non servono all'industria.* Business Insider.

[9] Wyder P. M., Bakhda R., Zhao M., Booth Q. A., Modi M. E., Song A., Kang S., Wu J., Patel P., Kasumi R. T., Yi D., Garg N. N., Jhunjhunwala P., Bhutoria S., Tong E. H., Hu Y., Goldfeder J., Mustel O., Kim D. & Lipson H. (2025). *Robot metabolism: Toward machines that can grow by consuming other machines.* Science Advances.

[10] Maturana H. R., Varela, F. J. (1972). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing.

[11] Pfeifer R., Lungarella M., Iida F. (2007). *Self-Organization, Embodiment, and Biologically Inspired Robotics*. *Science*, 318(5853), 1088–1093.

[12] Neuralink (2025). *From neural signals to life-changing impact*. Recuperato da: https://neuralink.com/

[13] Musk, E., et al. (2021). *An integrated brain-machine interface platform with thousands of channels*. Journal of Medical Internet Research.

[14] Reardon, T., Kaifosh, P., et al. (2024). *A high-throughput non-invasive neural interface for digital intent recognition via sEMG*. Nature.

[15] Moore G.E. (1965), *Cramming more components onto integrated circuits*, Reprinted from Electronics, vol.38, n.8.

[16] Buttazzo G. (2023), *Coscienza Artificiale: implicazioni per l'umanità*, Mondo Digitale

[17] Signorelli A.D. (2023), *Perché non dobbiamo temere la super intelligenza artificiale, La conversazione.* Wired. Reperibile all'indirizzo https://www.wired.it/article/intelligenza-artificiale-yann-lecun-meta/

[18] Varanini F. (2023), *AI, l'allarme di Yudkowsky: ascoltiamo i suoi timori*, articolo pubblicato su Agenda Digitale reperibile all'indirizzo https://www.agendadigitale.eu/cultura-digitale/ia-lallarme-di-yudkowsky-ascoltiamo-i-suoi-timori/

[19] Carboni K. (2023), *Uno dei pionieri dell'intelligenza artificiale ora si occuperà dei suoi rischi, Il personaggio*, Wired. Reperibile all'indirizzo https://www.wired.it/article/intelligenza-artificiale-rischi-geoffrey-hinton/

[20] Faggella D. (2024), *Cosmic Alignment vs Anthropocentric Alignment, Faggella*. Reperibile all'indirizzo https://danfaggella.com/alignment/

[21] Yampolskiy R.V. (2018), *Artificial Intelligence Safety and Security*, Chapman Hall/CRC, Artificial Intelligence and Robotics Series

[22] Feroni G.C. (2023), *AI nei processi decisionali della PA, il faro è la Costituzione* – Intervento di Ginevra Cerrina Feroni, GPDP. Reperibile all'indirizzo https://www.garanteprivacy.it/home/docweb/-/docweb-

display/docweb/9892588

Capitolo 4

[1] McCulloch, W. S., Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*. The Bulletin of Mathematical Biophysics, 5(4), 115-133.

[2] Rosenblatt, F. (1958). *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*. Psychological Review, 65(6), 386-408.

[3] Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

[4] Cornell Chronicle. (2019). *Professor's perceptron paved the way for AI –60years too soon*. Recuperato da https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

[5] The New York Times. (1958, July 7). *NEW NAVY DEVICE LEARNS BY DOING: Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser*.

[6] The New Yorker. (1958, December 6). *Talk of the Town: The Perceptron*.

[7] Minsky, M., Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

[8] Russell, S. J., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd ed.)*. Prentice Hall.

Capitolo 5

[1] Newell A., Simon H. A. (1972). *Human Problem Solving*. Prentice-Hall.

[2] Feigenbaum E. A., McCorduck, P. (1983). *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. Addison-Wesley.

[3] Jackson, P. (1999). *Introduction to Expert Systems (3rd ed.)*. Addison-Wesley.

[4] Buchanan, B. G., Shortliffe, E. H. (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

[5] Indurkhya N., Weiss S. M., Politakis P. (1992). *Complexity-Based Evaluation of Rule-Based Expert Systems*. Proceedings del Knowledge Engineering Symposium, 1992.

[6] Gödel K. (1931). *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatshefte für Mathematik und Physik, 38, 173–198

[7] Berners-Lee, T., Hendler, J., Lassila, O. (2001). *The Semantic Web*. Scientific American, 284(5), 34-43.

[8] Antoniou, G., van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press.

[9] Ginsberg, M. L. (1993). *Essentials of Artificial Intelligence*. Morgan Kaufmann.

[10] Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge Acquisition, 5(2), 199-220.

[11] Guarino, N. (1998). *Formal ontology and information systems*. In Proceedings of the first international conference on formal ontology in information systems (pp. 3-15).

[12] INPS, *L'ontologia Dominio INPS*, Recuperato da: https://www.inps.it/it/it/dati-e-bilanci/open-data/lod-inps/l-ontologia-dominio-inps.html

[13] Russell, S. J., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd ed.)*. Prentice Hall.

[14] Zadeh, L. A. (1965). *Fuzzy sets. Information and Control*, Elsevier, 8(3), 338-353.

Capitolo 6

[1] Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN*. Elsevier.

[2] Duda, R. O., Hart, P. E., Nilsson, N. J. (1976). *Subjective Bayesian methods for rule-based inference systems*. National Computer Conference, 45, 1075-1082.

[3] Zadeh, L. A. (1965). *Fuzzy sets*. Information and control, 8(3), 338-353.

[4] Ferrucci, D., et al. (2010). *Building Watson: An overview of the DeepQA project*. AI Magazine, 31(3), 59-79.

Capitolo 7

[1] Fukushima, K. (1980). *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*. Biological Cybernetics, 36(4), 193-202.

[2] Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). *Learning representations by back-propagating errors*. Nature, 323(6088), 533-536.

[3] Hubel, D. H., Wiesel, T. N. (1962). *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex*. The Journal of Physiology, 160(1), 106-154.

[4] Tesauro, G. (1995). *Temporal difference learning and TD-Gammon*. Communications of the ACM, 38(3), 58-68.

[5] Elman, J. L. (1990). *Finding structure in time*. Cognitive Science, 14(2), 179-211.

[6] Hochreiter, S., Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735-1780.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30.

[8] Sahoo P., Singh A.K., Saha S., Jain V. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. Recuperato da:

https://www.researchgate.net/publication/378183279_A_Systematic_Survey_of_Prompt_Engineering_in_Large_Language_Models_Techniques_and_Applications

[9] Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Mike Lewis, Yih W., Rocktäschel T., Riedel S., Kiela D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* Recuperato da: https://arxiv.org/abs/2005.11401

[10] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. (2014). *A Survey on Concept Drift Adaptation*. ACM Digital Library.

[11] Huang, L., et al. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. Recuperato da: https://arxiv.org/abs/2311.05232

[12] Towhidul S.M., Towhidul Islam Tonmoy S.M., Mehedi S.M., Jain V. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. Recuperato da: https://www.researchgate.net/publication/377081841_A_Comprehensive_Survey_of_Hallucination_Mitigation_Techniques_in_Large_Language_Models

[13] O'Brien, K., Casper, S., Anthony, Q., Korbak, T., et al. (2025). *Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs*. Recuperato da: https://arxiv.org/abs/2508.06601

[14] CNN Business. (2025). *The 'godfather of AI' reveals the only way humanity can survive superintelligence*. Recuperato da: https://www.cnn.com/2025/08/13/tech/ai-geoffrey-hinton

[15] Anthropic, (2025). *Persona vectors: Monitoring and controlling character traits in language models*. Recuperato da: https://www.anthropic.com/research/persona-vectors

[16] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, F., Le, H., Chi, E. H. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. Advances in Neural Information Processing Systems, 35.

[17] Gema, A. P., Hägele, A., Chen, R., Arditi, A., Goldman-Wetzler, J., Fraser-Taliente, K., Sleight, H., Petrini, L., Michael, J., Alex, B., Minervini, P., Chen, Y., Benton, J., & Perez, E. (2025). *Inverse Scaling in Test-Time Compute*. Recuperato da: https://doi.org/10.48550/arXiv.2507.14417

[18] Anthropic (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. Recuperato da: https://www.anthropic.com/research/measuring-faithfulness-in-chain-of-thought-reasoning

Capitolo 8

[1] Douglas B. Lenat, R. V. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*, Addison-Wesley, 1989.

[2] Douglas B. Lenat, R. V. Guha, *The Evolution of CycL, the Cyc Representation Language*, ACM SIGART Bulletin, June 1991.

[3] Lenat D. B., Feigenbaum E. A., *On the Thresholds of Knowledge*, Artificial Intelligence, 1991.

[4] Lenat D. B., Guha R. V., *Cyc: A Midterm Report*, Communications of the ACM, 1990. Stato dell'arte e risultati preliminari del progetto Cyc.

[5] Ferrucci D. et al., *Watson: Beyond Jeopardy!*, Artificial Intelligence, Vol. 199, pp. 93–105, 2013.

[6] Ferrucci D, Brown E., *AdaptWatson: A Methodology for Developing and Adapting Watson Technology*, IBM Research Report RC25244 (2012).

[7] Kumar A. et al., *A Survey on IBM Watson and Its Services*, J. Phys.: Conf. Ser., 2022.

Capitolo 9

[1] Marcus, G. (2020). *The Next Decade in AI: Four Steps Towards Robust AI*. arXiv preprint arXiv:2002.06177.

[2] Anthropic. (2025). *Project Vend: An Experiment in AI Agency*. Anthropic Blog.

[3] Moore, J., Gergley, A., Zou, J. (2023). *Evaluating and Mitigating Stigmatizing Language in Large Language Models*. Stanford University.

[4] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[5] Yehuda, R., Daskalakis, N. P., Bierer, L. M., Bader, H. N., Klengel, T., Holsboer, F., Binder, E. B. (2016). *Holocaust exposure induced intergenerational effects on FKBP5 methylation*. Biological Psychiatry, 80(5), 372-379.

[6] Rao, A. S., Georgeff, M. P. (1995). *BDI agents: from theory to practice*. Proceedings of the 1th int. conference on Multi-agent systems, 312-319.

Capitolo 10

[1] Beni, G., Wang, J. (1988). *Swarm intelligence in cellular robotic systems*. Proceedings of the NATO Advanced Study Institute on Robots and Biological Systems.

[2] Bonabeau, E., Dorigo, M., Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press.

[3] Dorigo, M., Stützle, T. (2004). *Ant Colony Optimization*. MIT Press.

[4] Kennedy, J., & Eberhart, R. C. (1995). *Particle swarm optimization*. Proceedings of ICNN'95 - International Conference on Neural Networks.

[5] Iannace, D. E. (2021). *American swarms: lo sviluppo americano di sciami di droni ad uso militare*. Ce.S.I. Centro Studi Internazionali. Recuperato da: https://www.cesi-italia.org/it/articoli/american-swarms-lo-sviluppo-americano-di-sciami-di-droni-ad-uso-militare

[6] Tseng, B. (2025). *Shield AI Cofounder Says the US Military Needs More Affordable Drones*. Business Insider. Recuperato da:

https://www.businessinsider.com/affordable-ai-military-drones-shield-ai-2025-3

[7] 7.Reddy, V. D., Hussain, M. M. (2025). *Swarm Intelligence: Theory and Applications in Fog Computing, Beyond 5G Networks, and Information Security*. Taylor & Francis Ltd.

[8] Abraham, A., & Grosan, C. (2006). *Swarm Intelligence in Data Minin*g. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

[9] Nelson, D. (s.d.). *Progetto AI di F-Secure per sfruttare il potenziale della "Swarm Intelligence"*. Unite.AI. Recuperato da: https://www.unite.ai/it/ai-project-by-f-secure-to-harness-potential-of-swarm-intelligence/

[10] Mancosu, C. (2024). *Capitalismo come algoritmo autoapprendente: tra burocrazia weberiana e swarm intelligence*. Il Sole 24 Ore. Recuperato da: https://carlomancosu.nova100.ilsole24ore.com/2024/11/30/burocrazia/?refresh_ce=1

[11] Lombardo, A. (2022). *Swarm Intelligence, dagli sciami alla logistica*. Logistica News. Recuperato da: https://www.logisticanews.it/swarm-intelligence-dagli-sciami-alla-logistica/

[12] Nayar, N., Ahuja, S., & Jain, S. (2019). *Swarm intelligence and data mining: a review of literature and applications in healthcare*. Recuperato da: https://www.researchgate.net/publication/335198731_Swarm_intelligence_and_data_mining_a_review_of_literature_and_applications_in_healthcare

[13] Alizadehsani, R., Roshanzamir, M., Izadi, N. H., Gravina, R., & Fortino, G. (2023). *Swarm intelligence in internet of medical things: A review*. Sensors.

[14] Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manam, S. R., Djambazian, H., Pickkers, P., Netea, M. G. (2021). *Swarm Learning for decentralized and confidential clinical prediction*. Nature.

[15] Evora, J., Hernandez, J. J., Hernandez, M., & Perez, J. (2015). *Swarm intelligence*

for frequency management in smart grids. International Journal of Intelligent Systems Technologies and Applications.

[16] Mohamed, M. A., Eltamaly, A. M., & Alolah, A. I. (2017). *Swarm intelligence-based optimization of grid-dependent hybrid renewable energy systems*. Renewable and Sustainable Energy Reviews.

[17] Kamran, A. (2025). *Generative AI "Agile Swarm Intelligence"*. Medium.

[18] Lella L., Licata I., Pristipino, C. (2022). *Pima Indians Diabetes Database Processing through EBBM-Optimized UTM Model*. In Nathalie Bier, Ana L. N. Fred, Hugo Gamboa, editors, Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2022, vol 5: HEALTHINF, Online Streaming, February 9-11, 2022. pages 384-389, SCITEPRESS, 2022.

[19] Turing A. (1948). *Intelligent Machinery*. In Collected Works of A.M. Turing: Mechanical Intelligence. Edited by D.C. Ince. Elsevier Science Publishers, 1992.

[20] Mitchell, M. (1999). *Introduzione agli algoritmi genetici*. Apogeo.

[21] Lella, L. (2024). *Orientare le decisioni della PA con i dati e l'IA: metodi e strumenti per un futuro sostenibile*. Pubblicato sul portale di agendadigitale.eu: https://www.agendadigitale.eu/cittadinanza-digitale/orientare-le-decisioni-della-pa-con-i-dati-e-lia-metodi-e-strumenti-per-un-futuro-sostenibile/

Capitolo 11

[1] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The Semantic Web*. Scientific American, 284(5), 34-43.

[2] Kuck, G. (2004). *Tim Berners-Lee's Semantic Web*. SAJIM, 6(1), 87-95. URL: https://sajim.co.za/index.php/sajim/article/download/297/288

[3] Sinhal, P. *Understanding LLM-Based Agents and their Multi-Agent Architecture*. Medium. URL: https://medium.com/@pallavisinha12/understandingllm-based-

agents-and-their-multi-agent-architecture-299cf54ebae4

[4] Wang, Z., et al. (2025). *Talk Structurally, Act Hierarchically: A Collaborative Framework for LLM Multi-Agent Systems*. URL: https://arxiv.org/abs/2502.11098

[5] LangChain AI. *Balance agent control with agency*. URL: https://www.langchain.com/langgraph

[6] Botpress. *Guida ai sistemi multi-agente nel 2025*. URL: https://botpress.com/it/blog/multi-agent-systems

[7] Giuliani, R. *Tecnologie e framework per la programmazione multi-agente*. Amslaurea Unibo. URL: https://amslaurea.unibo.it/id/eprint/10101/1/Giuliani_Roberto_tesi.pdf

[8] Barpi, F. (2024). *Analisi e Ottimizzazione dei processi produttivi tramite architettura basata su AI generativa e knowledge base graph: L'uso di LLM ad agenti per l'analisi log-driven*. Unire Unige. URL: http://unire.unige.it/handle/123456789/10867

[9] Agenda Digitale. (2025). *AI autonoma: il ruolo degli agenti intelligenti e dei sistemi multi-agente*. Agenda Digitale. URL: https://www.agendadigitale.eu/industry-4-0/ai-autonoma-il-ruolo-degli-agenti-intelligenti-e-deisistemi-multi-agente/

[10] Yuksel K.A., Sawaf H. (2024). *A Multi-AI Agent System for Autonomous Optimization of Agentic AI Solutions via Iterative Refinement and LLM-Driven Feedback Loops* URL: https://arxiv.org/pdf/2412.17149

[11] Chen, J., et al. *MAgICoRe: Multi-Agent, Iterative, Coarse-to-Fine Refinement for Reasoning*. OpenReview. URL: https://openreview.net/forum?id=j9wBgcxa7N

[12] Yuksel, K. A., et al. (2024). *A Multi-AI Agent System for Autonomous Optimization of Agentic AI Solutions via Iterative Refinement and LLM-Driven Feedback Loops*. URL: https://arxiv.org/abs/2412.17149

Capitolo 12

[1] Ng, A. (2024). *What's next for AI agentic workflows*. URL: https://www.youtube.com/watch?v=8b7CLBCS3pg

[2] Gartner (2026). *Predicts 2026: AI Agents Will Transform IT Infrastructure and Operations*. Gartner Insights Abstract. URL: https://www.gartner.com/en/documents/7242030

[3] Anthropic, *Multi-agent coordination patterns: Five approaches and when to use them*. URL: https://claude.com/blog/multi-agent-coordination-patterns

[4] Klimt, B., & Yang, Y. (2004). *The Enron Corpus: A New Dataset for Email Classification Research*. In: Proceedings of ECML 2004, Lecture Notes in Computer Science, vol. 3201. Springer.

[5] Krishnan R. (2026), *LLM Enron: experiments on strucure vs scale*. Strange Loop Canon. URL: https://www.strangeloopcanon.com/p/llm-enron-experiments-on-structure

[6] Lai, J. et al. (2025). *LLM-based Multi-Agent Blackboard System for Information Discovery in Data Science*. arXiv:2510.01285.

[7] Zhang, Y. et al. (2025). *Exploring Advanced LLM Multi-Agent Systems Based on Blackboard Architecture*. arXiv:2507.01701.

[8] Wu S., Shu K. (2025). *Memory in LLM-based Multi-agent Systems: Mechanisms, Challenges, and Collective Intelligence*. Mint Lab, Machine Intelligence N Trust

[10] Anthropic (2026). *Measuring AI agent autonomy in practice*. URL: https://www.anthropic.com/news/measuring-agent-autonomy

Capitolo 13

[1] Regolamento (UE) 2016/679 – GDPR, artt. 5, 24, 25, 32.

[2] Regolamento (UE) 2024/… – Artificial Intelligence Act (AI Act), requisiti e obblighi per sistemi ad alto rischio.

[3] Direttiva (UE) 2022/2555 – NIS2, sicurezza delle reti e dei sistemi informativi.

[4] ISO/IEC 27001:2022 – Sistemi di Gestione della Sicurezza delle Informazioni.

[5] ISO/IEC 42001:2023 – Sistemi di Gestione per l'Intelligenza Artificiale.

[6] Codice dell'Amministrazione Digitale (D.Lgs. 82/2005 e s.m.i.) – Principi di sicurezza, trasparenza, gestione documentale e uso delle tecnologie digitali nella PA.

[7] McMahan, H. B., et al. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. AISTATS.

[8] Dwork, C., Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Now Publishers.

[9] Google. (2025) *Private AI Compute Advances AI Privacy*. URL: https://blog.google/technology/ai/google-private-ai-compute/

[10] Apple Security. (2024) *Private Cloud Compute*. URL: https://security.apple.com/blog/private-cloud-compute/

Capitolo 14

[1] McKinsey & Company, "The Year of Quantum: From concept to reality in 2025", Giugno 2025. URL: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-year-of-quantum-from-concept-to-reality-in-2025

[2] Software House AU, "Quantum Computing & AI in 2025: Breakthroughs & Impact", Settembre 2025.

[3] Ministero dell'Università e della Ricerca (MUR), "Strategia italiana per le tecnologie quantistiche", Settembre 2025.

[4] Dipartimento per la trasformazione digitale, "Stati Generali Quantum: un Polo nazionale per ricerca e applicazioni tra università, imprese e PA", Dicembre 2025.

[5] IEA, “The trasformative potential of AI depends on energy”, 2024. URL: https://www.iea.org/reports/energy-and-ai/executive-summary

[6] Business Wire, "Neuromorphic Computing Market Report 2025-2035", Dicembre 2025.

[7] ScienceAlert, "Computers Made From Human Brain Tissue Are Coming", Dicembre 2025. URL: https://www.sciencealert.com/computers-made-from-human-brain-tissue-are-coming-are-we-prepared

[8] FinalSpark, "Neuroplatform", Piattaforma di biocomputing. URL: https://finalspark.com/

[9] EurekAlert, "Brain cells learn faster than machine learning, new research shows", Agosto 2025. URL: https://www.eurekalert.org/news-releases/1094351#:~:text=Melbourne%2C%20Australia%20%2D%2012%20August%202025,%2Dart%20RL%20(reinforcement%20learning)

Capitolo 15

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

[2] Cui, Y., et al. (2025). *Can LLMs Threaten Human Survival? Benchmarking Potential Existential Threats from LLMs via Prefix Completion*. https://arxiv.org/pdf/2511.19171

[3] LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. OpenReview. https://openreview.net/pdf?id=BZ5a1r-kVsf

[4] Destrade, M., Bounou, O., Le Lidec, Q., Ponce, J., & LeCun, Y. (2025). *Value-guided action planning with JEPA world models*.

https://arxiv.org/abs/2601.00844

[5] Abbas, A. *Meta V-JEPA 2: il modello di intelligenza artificiale che porta il buon senso ai robot*. Unite AI. https://www.unite.ai/it/meta-v-jepa-2-the-ai-model-bringing-common-sense-to-robots/

Capitolo 16

[1] Gartner, Inc*. Documentazione e ricerche relative alla "Digital Integration Hub Architecture"*.

[2] Open Knowledge Foundation. *Open Data Handbook*. Reperibile su: opendatahandbook.org/it/

[3] ODMC.org, *Open Data Management Cycle (ODMC)*. Reperibile su: odmc.org

Capitolo 17

[1] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[2] Buolamwini, J., Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT)', 77-91. https://www.proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[3] Barocas, S., Selbst, A. D. (2016). *Big Data's Disparate Impact*. California Law Review, 104(3), 671-733. https://www.californialawreview.org/wp-content/uploads/2016/06/03-Barocas-Selbst.pdf

[4] Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

[5] Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che

abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati). https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A32016R0679

[6] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Blanchard, N., Zhao, Z. (2021). *Advances and Open Problems in Federated Learning. Foundations and Trends*. Machine Learning, 14(1–2), 1-210. https://arxiv.org/pdf/1912.04977.pdf

[7] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Vayena, E. (2018). *AI4People—Ethical Guidelines for Trustworthy AI: A European Perspective*. Minds and Machines, 29(4), 689-707. https://link.springer.com/article/10.1007/s11023-018-9482-y

[8] Burrell, J. (2016). *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data Society, 3(1). https://journals.sagepub.com/doi/full/10.1177/2053951715622512

[9] Adadi, A., Berrada, M. (2018). *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*. IEEE Access, 6, 52138-52160. https://ieeexplore.ieee.org/document/8466506

[10] Frey, C. B., Osborne, M. A. (2017). *The future of employment: How susceptible are jobs to computerisation?*. Technological Forecasting and Social Change, 114, 254-280. https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

[11] Acemoglu, D., Restrepo, P. (2019). *Automation and New Tasks: How Technology Displaces and Reinstates Labor*. Journal of Economic Perspectives, 33(2), 3-30. https://www.aeaweb.org/articles?id=10.1257/jep.33.2.3

[12] European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[13] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000380455

[14] European Parliament. (2024). *Artificial Intelligence Act: MEPs adopt landmark law*. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

Capitolo 18

[1] Diritto.it. (2024). *AI Act: in vigore dal 2 febbraio per sistemi a rischio e formazione*. Reperibile all'indirizzo: diritto.it

[2] Agenda Digitale. (2024). *Ai Act: cos'è e come plasma l'intelligenza artificiale in Europa*. Reperibile all'indirizzo: agendadigitale.eu

[3] Osservatori.net. (2024). *AI Act: cosa prevede e come si stanno muovendo le aziende*.osservatori.net

[4] EU Artificial Intelligence Act (2024), *The Act Texts*, Official Journal of the European Union. Reperibile all'indirizzo: https://artificialintelligenceact.eu/the-act/

[5] Camera dei Deputati. (2024). *Il regolamento UE in materia di intelligenza artificiale*.camera.it

[6] Villani, S. (2024). Il sistema di vigilanza sull'applicazione dell'AI Act: ognun per sé?, Quaderni AISDUE – Rivista quadrimestrale. Reperibile all'indirizzo: https://cris.unibo.it/retrieve/08ec6906-e8c0-41fc-ab46-ccf61c140bb3/2024_Il%20sistema%20di%20vigilanza%20sull%E2%80%99applicazione%20dell_AI%20Act_QuaderniAISDUE.pdf

Capitolo 19

[1] Legge 23 settembre 2025, n. 132, Disposizioni e deleghe al Governo in materia di intelligenza artificiale. Gazzetta Ufficiale della Repubblica Italiana, Serie Generale n. 223 del 25-9-2025. Disponibile su: https://www.gazzettaufficiale.it/eli/id/2025/09/25/25G00143/sg

[2] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13

giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale). Gazzetta Ufficiale dell'Unione Europea L 2024/1689 del 12.7.2024. Disponibile su: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=OJ:L_202401689

[3] Regolamento (UE) 2022/868 del Parlamento europeo e del Consiglio, del 30 maggio 2022, relativo alla governance europea dei dati e che modifica il regolamento (UE) 2018/1724 (Data Governance Act). Gazzetta Ufficiale dell'Unione Europea L 152/1 del 3.6.2022. Disponibile su: https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng

[4] Regolamento (UE) 2023/2854 del Parlamento europeo e del Consiglio, del 13 dicembre 2023, riguardante norme armonizzate sull'accesso equo ai dati e sul loro utilizzo e che modifica il regolamento (UE) 2017/2394 e la direttiva (UE) 2020/1828 (regolamento sui dati). Gazzetta Ufficiale dell'Unione Europea L 2023/2854 del 22.12.2023. Disponibile su: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=OJ:L_202302854

[5] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati). Gazzetta Ufficiale dell'Unione Europea L 119/1 del 4.5.2016. Disponibile su: https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016R0679

Capitolo 20

[1] Asimov, I. (1950). *I, Robot*. Gnome Press.

[2] Asimov, I. (1985). *Robots and Empire*. Doubleday.

[3] Awad, E., Dsouza, S., Kim, R., Schulz, J., Park, J., Bonnefon, J.-F., Shariff, A. (2018). *The Moral Machine experiment*. Nature, 563(7729), 59-64.

[4] Lin, P. (2017). *The Ethics of Autonomous Cars*. The Atlantic.

[5] SAE International. (2021). *J3016: Levels of Driving Automation*.

[6] Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

[7] Heidegger, M. (1954). *Die Frage nach der Technik* (La questione della tecnica). In *Vorträge und Aufsätze*. Neske.

[8] Rousseau, J.-J. (1762). *Du Contrat social ou Principes du droit politique*. Marc-Michel Rey.

[9] Bobbio, N. (1990). *L'età dei diritti*. Einaudi.

[10] Sen, A. (1999). *Development as Freedom*. Alfred A. Knopf.

Capitolo 21

[1] FPA. (2024, Maggio 21). *Ricerca FPA "impatto dell'intelligenza artificiale sul pubblico impiego"*. Recuperato da https://www.forumpa.it/pa-digitale/ricerca-fpa-impatto-dellintelligenza-artificiale-sul-pubblico-impiego-il-57-dei-dipendenti-pubblici-e-altamente-esposto/

[2] AgID. (2024, Luglio 4). *AI nella pubblica amministrazione: la fiducia dei cittadini chiave per il successo*. Recuperato da https://www.agendadigitale.eu/cittadinanza-digitale/ia-nella-pubblica-amministrazione-la-fiducia-dei-cittadini-chiave-per-il-successo/

[3] Gerschenkron, Alexander (1962). *Economic Backwardness in Historical Perspective: A Book of Essays*. Cambridge, MA: Harvard University Press.

[4] Brezis, Elise S., Krugman, Paul R., e Tsiddon, Daniel (1993). *Leapfrogging in International Competition: A Theory of Cycles in National Technological Leadership.* The American Economic Review.

[5] AgID (2025). *Bozza di linee guida per l'adozione di AI nella pubblica amministrazione*. Recuperato da: https://www.agid.gov.it/sites/agid/files/2025-

02/Linee_Guida_adozione_AI_nella_PA.pdf

[6] Floridi L., Cabitza F. (2021), *intelligenza artificiale – l'uso delle nuove macchine*, Bompiani.

[7] European Union. *AI Act: Shaping Europe's digital future*. Disponibile su: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Capitolo 22

[1] AgID. *Linee Guida per l'adozione dell'intelligenza artificiale nella pubblica amministrazione*. Disponibile su: https://www.agid.gov.it/it/notizie/intelligenza-artificiale-in-consultazione-le-linee-guida-pa

[2] European Union. *AI Act: Shaping Europe's digital future*. Disponibile su: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[3] Fritzke B. (1994). *A Growing Neural Gas Network Learns Topologies*. Part of: Advances in Neural Information Processing Systems 7, NIPS.

[4] Lella L., Gentile L., Pristipino C., Toni D. (2021), *Predictive Clustering Learning Algorithms for Stroke Patients Discharge Planning*, Proc. Healthinf 2021.

[5] Lella L., Licata I. (2018), *Length of Hospital Stay Prediction through Unorganised Turing Machines*, Proc. HealthInf 2018.

[6] Alpaydin E. (2020). *Introduction to Machine Learning*. 4th Edition. MIT Press.

[7] Breiman L. (2001). *Random Forests*. Machine Learning.

[8] Turing A. (1948). *Intelligent Machinery*. In Collected Works of A.M. Turing: Mechanical Intelligence. Edited by D.C. Ince. Elsevier Science Publishers, 1992.

[9] Mitchell, M. (1999). *Introduzione agli algoritmi genetici*. Apogeo.

[10] Lella L., Licata I., Pristipino, C. (2022). *Pima Indians Diabetes Database*

Processing through EBBM-Optimized UTM Model. In Nathalie Bier, Ana L. N. Fred, Hugo Gamboa, editors, Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2022, vol 5: HEALTHINF, Online Streaming, February 9-11, 2022. pages 384-389, SCITEPRESS, 2022.

[11] Lella L., Licata I. (2018). *Length of Hospital Stay Prediction through Unorganised Turing Machines*, Proc. HealthInf 2018, Funchal, Madeira (Portogallo), 19-21 Gennaio 2018.

[12] Lella L., Licata I., Pristipino C. (2022). *L'intelligenza artificiale in aiuto alla Sanità regionale: il progetto AIEHEM*. Pubblicato sul portale di agendadigitale.eu: https://www.agendadigitale.eu/sanita/lintelligenza-artificiale-in-aiuto-alla-sanita-regionale-il-progetto-aiehem/

[13] Lannelongue L., Grealey J., Inoyue M. (2021). *Green Algorithms: Quantifying the Carbon Footprint of Computation*. Wiley Advanced.

[14] Istituto Superiore di Sanità – Ministero della Salute (2020). *Piattaforma per la pianificazione, il monitoraggio e la valutazione dei piani regionali di prevenzione 2020-2025* – Obiettivi Strategici. Recuperato da: https://www.pianiregionalidellaprevenzione.it/tabelle/obiettivi_strategici.aspx

[15] Ministero della Salute (2024). *Specifiche per la trasmissione dati e manuali (EMUR)*. Recuperato da: https://www.salute.gov.it/new/it/tema/nuovo-sistema-informativo-sanitario/specifiche-la-trasmissione-dati-e-manuali-emur/

[16] Lella L., *Società 5.0: strumenti e tecnologie per monitorare il benessere sociale sostenibile*, pubblicato sul portale di agendadigitale.eu: https://www.agendadigitale.eu/cittadinanza-digitale/societa-5-0-strumenti-e-tecnologie-per-monitorare-il-benessere-sociale-sostenibile/

[17] Lella L., *Benessere sociale sostenibile: come coinvolgere i cittadini col touchpoint del sistema A.L.I.C.E.,* pubblicato sul portale di agendadigitale.eu: https://www.agendadigitale.eu/cittadinanza-digitale/benessere-sociale-sostenibile-come-coinvolgere-i-cittadini-col-touchpoint-alice/

[18] AWS, *Persone: il lato umano dell'innovazione in Amazon*. Recuperato da: https://aws.amazon.com/it/executive-insights/content/the-human-side-of-innovation/