LUISS

Corso di laurea in Economa e Management

Cattedra di Informatica

Dal Passato al Futuro: L'Evoluzione dell'Intelligenza Artificiale e le Sfide del Presente

Prof. Luigi Laura

RELATORE

Ludovica De Biase 275671

CANDIDATA

Anno Accademico 2024/2025

A mio padre

SOMMARIO

INTRODUZIONE

L’intelligenza artificiale (IA) rappresenta una delle tecnologie più rivoluzionarie del nostro tempo. L'idea che una macchina possa replicare o addirittura superare l’intelligenza umana solleva interrogativi profondi non solo sul piano tecnologico, ma anche su quello etico, giuridico e filosofico. La diffusione dell’IA ha un impatto trasversale che tocca numerosi ambiti della vita umana, questa tesi si propone di esplorare l’evoluzione dell’intelligenza artificiale attraverso il tempo, analizzando in tre capitoli distinti il passato, il presente e il futuro di questa disciplina.

Il primo capitolo è dedicato alla nascita e allo sviluppo dell’intelligenza artificiale, alla distinzione tra intelligenza artificiale e intelligenza umana, concetto centrale nella riflessione sull'IA; ed esplora il passaggio dalle prime teorie logiche e computazionali agli approcci basati sulle reti neurali, evidenziando come la ricerca si sia progressivamente focalizzata sull'imitazione dei processi cognitivi umani attraverso l'automazione e i sistemi di apprendimento automatico.

Il secondo capitolo si concentra sullo stato attuale dell’intelligenza artificiale, analizzando le sue applicazioni pratiche e le principali sfide etiche e legali che emergono dall'uso crescente di tecnologie autonome. Viene proposto un esempio delle potenzialità dell’intelligenza artificiale, il progetto AlphaFold, sviluppato da DeepMind, che è riuscito a risolvere uno dei più complessi problemi scientifici: la predizione della struttura tridimensionale delle proteine. Nel contesto attuale, la fiducia nei sistemi di intelligenza artificiale è una questione centrale, un recente sondaggio ha infatti rilevato che molti utenti sono preoccupati dalla mancanza di trasparenza e dall'eventuale uso scorretto dei dati generati da sistemi di IA, evidenziando una diffusa sfiducia verso le tecnologie troppo complesse da interpretare. Il Regolamento Generale sulla Protezione dei Dati (GDPR) e normative come l'Artificial Intelligence Act, proposto dalla Commissione Europea, cercano di rispondere a queste sfide.

Infine, il terzo capitolo della tesi guarda al futuro dell’intelligenza artificiale, esplorando le prospettive e le sfide che questa tecnologia comporta per le generazioni future. Il dibattito sul futuro dell’IA è particolarmente acceso e polarizzato tra visioni ottimistiche e scenari distopici. Da un lato, l’intelligenza artificiale promette di risolvere alcuni dei problemi più urgenti dell’umanità, come la cura delle malattie, la lotta ai cambiamenti climatici e la gestione di risorse globali. Dall’altro, l’avvento di sistemi superintelligenti solleva preoccupazioni riguardanti l’occupazione, la disuguaglianza sociale e persino il rischio di perdere il controllo su tali tecnologie.

CAPITOLO I: PASSATO

1.1 Definizione e distinzione tra intelligenza umana e artificiale

L'intelligenza artificiale è un termine ampio particolarmente complesso: nel campo di ricerca dell'informatica, infatti, non esiste una definizione precisa di intelligenza artificiale; si va da qualsiasi sistema software che esegue compiti complessi, attraverso un'ampia gamma di algoritmi di apprendimento automatico, fino alle tecnologie digitali incarnate della robotica. In generale, la caratteristica distintiva dei sistemi di intelligenza artificiale è che presentano compiti cognitivi complessi, ovvero esibiscono un'elaborazione avanzata degli input che può essere paragonata alla comprensione umana delle immagini o della voce. Le sorprendenti prestazioni di classificazione e previsione degli algoritmi di apprendimento suggeriscono quindi l'idea che la macchina abbia davvero imparato e forse capito qualcosa, mentre in realtà ha semplicemente ottimizzato un (enorme) numero di parametri cercando in un dato (ricco) insieme di soluzioni. Inoltre, a differenza di altri sistemi software complessi, ma in modo simile alla comprensione umana, l'elaborazione dei dati dell'IA è in grado di adattarsi in modo flessibile, in una certa misura, ai cambiamenti ambientali, in modo che il suo comportamento possa essere adattato all'utente o all'agente che interagisce con la macchina. D'altra parte, il termine Intelligenza Artificiale è ormai andato oltre l'informatica e le tecnologie digitali per entrare in molti altri campi, come l'economia, la politica, il diritto, la filosofia, le scienze sociali, le neuroscienze, la psicologia, l'educazione, ed è un tema ricorrente anche in mass media e nel discorso pubblico. In ciascuno di questi ambiti il termine si arricchisce di numerose connotazioni specifiche che spesso rimangono implicite, anche se talvolta le implicazioni di tali significati o sfumature nascoste possono diventare molto esplicite, dando luogo ad esempio a una regolamentazione ufficiale o a investimenti[1]. In un simile contesto, un ruolo fondamentale può essere giocato dalla ricerca interdisciplinare, che si basa principalmente sul dialogo umano.

In riferimento al confronto interdisciplinare, una distinzione principale tra le discipline accademiche è tra discipline scientifiche e discipline umanistiche. Oltre agli argomenti di interesse, differiscono nella metodologia di ricerca: mentre il metodo scientifico si fonda su un metodo sperimentale manipolativo e su una rigorosa formalizzazione basata sulla matematica, le discipline umanistiche acquisiscono nuove conoscenze attraverso metodi critici, speculativi e comparativi. Man mano che l'intelligenza artificiale entra in settori come la medicina, la giustizia penale, il lavoro e i mercati finanziari, un dialogo tra diverse competenze e un'adeguata combinazione di diversi metodi di ricerca sono di fondamentale importanza per fertilizzare le idee, fornire approfondimenti e prevenire fallimenti. In questa sezione ci concentriamo sul problema della combinazione di metodi di indagine

[1] Powles, J. 2017. *New York City's bold, flawed attempt to make algorithms accountable.* The New Yorker.

provenienti da diverse discipline, esplorando una serie di questioni nel contesto specifico della ricerca sull'IA.

L'approccio scientifico degli informatici richiede che lo sviluppo di una teoria o di un artefatto (un algoritmo, un software, uno strumento, una macchina) si basi su definizioni ben fondate. Questo approccio garantisce che, indipendentemente dai nomi scelti dai ricercatori per identificare un concetto, il suo significato e i risultati che logicamente e matematicamente ne derivano possano sempre essere disambiguati guardando la sua definizione. L'intento di una definizione formale è anche quello di rimuovere connotazioni e significati alternativi che solitamente sono attribuiti a un termine di un linguaggio umano: è una questione filosofica capire in che misura le definizioni formali, cioè matematiche e logiche, abbiano il significato univoco voluto. D'altra parte, ci sono argomenti che per loro stessa natura hanno un significato complesso e ampio; ad esempio, i concetti di salute, giustizia, giusto processo, discriminazione, etica e anche intelligenza difficilmente possono essere ridotti a una definizione matematica o ad una metrica[2].

Qui emerge una differenza fondamentale tra intelligenza umana e artificiale. L'ambiguità è un elemento costitutivo dell'intelligenza umana: le persone si impegnano efficacemente e creativamente in molte attività sociali facendo affidamento su termini e concetti ambigui, la cui semantica viene "definita" pragmaticamente durante le loro azioni. A questo proposito, occorre citare la "teoria delle intelligenze multiple"[3], elaborata da Howard Gardner, secondo cui non esiste una capacità universale chiamata intelligenza, che possa essere misurata e catalogata in modo assoluto, bensì una vasta gamma di dimensioni, ciascuna con caratteristiche differenti, presenti in misura diversa in ciascuno di noi, il cui sviluppo varia da soggetto a soggetto. Nella prima stesura del suo studio, Howard Gardner nominava sette diverse intelligenze, alle quali ne ha aggiunto altre due: linguistica, logico-matematica, spaziale, cinestetica, musicale, interpersonale, intrapersonale, naturalistica, filosofico-esistenziale. Le macchine richiedono invece rappresentazioni esplicite, come ontologie o reti semantiche, per gestire concetti imprecisi. È interessante notare che la tecnica dell'intelligenza artificiale con apprendimento per rinforzo può essere vista come l'utilizzo di una sorta di semantica pragmatica. In questo caso, l'agente autonomo valuta e interagisce con l'ambiente in cui opera, reagendo ai cambiamenti ambientali (eventualmente causati dalle sue azioni precedenti) in modo da massimizzare un determinato obiettivo. In particolare, l'agente riceve una ricompensa o una penalità per la reazione eseguita, in modo da calibrare il suo algoritmo di apprendimento e scegliere l'azione successiva adeguata. L'apprendimento per rinforzo è una soluzione efficace nelle applicazioni pratiche in cui l'ambiente dell'agente non è completamente definito, quindi in qualche modo ambiguo (tecnica model-free), tuttavia è lontano da un'"intelligenza" che lavora con concetti ambigui. Pertanto,

[2] Russell, S. & Norvig, P. 2010. *Artificial Intelligence: a modern approach*. New Jersey: Pearson.
[3] Gardner, H. (1983). "Frames of Mind: The Theory of Multiple Intelligences."

un utile esito di un dialogo interdisciplinare sarebbe l'identificazione nelle tecniche di IA sia di ambiguità che dovrebbero essere meglio determinate, sia di definizioni e assunzioni la cui formalizzazione non è del tutto corretta rispetto al concetto a cui si riferiscono.

Lipton spiega chiaramente che spesso le formulazioni dei problemi di machine learning corrispondono in modo imperfetto ai compiti della vita reale che dovrebbero risolvere. Ciò può accadere quando obiettivi complessi della vita reale sono difficili da codificare come parametri o semplici funzioni numeriche da ottimizzare[4]. Ad esempio, l'etica e la legalità non possono essere direttamente ridotte a obiettivi di ottimizzazione numerica di un algoritmo decisionale. Quando gli algoritmi si occupano di obiettivi che riteniamo importanti ma che facciamo fatica a modellare formalmente, sono richiesti requisiti problematici come interpretabilità, spiegazione, trasparenza. Pensiamo che le conoscenze sviluppate dalle discipline umanistiche possano essere utili qui, poiché il metodo di ricerca di queste discipline ha una chiara idea di cosa significhi essere precisi senza essere matematicamente formali[5].

Un dialogo efficace tra diverse discipline richiede una comprensione condivisa delle principali nozioni dell'argomento in discussione. Nel caso dell'IA si possono elencare ad esempio i termini intelligenza, comportamento, volontà, somiglianza, causalità, neurone, azione, accuratezza, verità, equità, precisione, ma se ne potrebbero proporre molti altri. Ciascuno di questi termini appartiene al vocabolario di molteplici discipline diverse, dove acquisisce connotazioni specifiche, sfumature e forse riferimenti nascosti che sono diventati impliciti in specifiche comunità di ricerca. Ad esempio, i termini intelligenza o causalità sono usati dagli informatici in un senso molto più ristretto che dai filosofi o dai neuroscienziati[6]. D'altra parte, i termini accuratezza, precisione ed equità sono utilizzati negli algoritmi di machine learning con riferimento a definizioni matematiche molto specifiche, che consentono preziosamente di confrontare adeguatamente le prestazioni di diversi algoritmi. Inoltre, i ricercatori di intelligenza artificiale a volte utilizzano termini antropomorfi e definizioni colloquiali suggestive (come algoritmo di comprensione della lettura o vettore di pensiero) che potrebbero essere un'utile fonte di ispirazione se mantenuti all'interno della comunità di ricerca insieme alla loro adeguata qualifica tecnica, ma che possono creare confusione e dare un'idea in senso fuorviante delle capacità dell'intelligenza artificiale quando comunicate al di fuori del loro contesto originale. Pertanto, il difficile processo di instaurazione di un dialogo interdisciplinare basato su un linguaggio comune ha il vantaggio di portare alla luce presupposti e connotazioni che potrebbero essere diventati impliciti, e forse trascurati o dimenticati, nel gergo quotidiano della ricerca[7].

---

[4] Lipton, Z. 2016. *The mythos of model interpretability*. ICML Workshop on Human Interpretability, and Communications of the ACM 61(10):36-43, 2018.

[5] Lipton, Z. & Steinhardt, J. 2018. *Troubling trends in machine learning scholarship.* CoRR, abs/1807.03341.

[6] Russo, F. 2018. "Digital technologies, ethical questions, and the need of an informational framework". Philosophy & Technology 31:655-677.

[7] Ibidem.

Tuttavia, per molti concetti potrebbe essere impossibile per diversi esperti concordare completamente su un significato comune senza sacrificare l'espressività intesa di un termine specifico. Pertanto, invece di un unico linguaggio comune, un utile dialogo interdisciplinare potrebbe basarsi su più linguaggi disciplinari che interagiscono produttivamente. Ad esempio, consideriamo il caso dei processi di apprendimento disparati (DLP), che è una classe di algoritmi di apprendimento automatico che è stata proposta per affrontare il problema della discriminazione degli algoritmi di classificazione. Gli informatici hanno fatto ricorso ad una nota terminologia giuridica per definire i criteri tecnici che quantificano la discriminazione dell'algoritmo. Più precisamente, la nozione giuridica di trattamento disparato è una forma di differenza intenzionale nel trattamento di sottogruppi protetti, mentre un impatto disparato si riferisce a pratiche faccialmente neutre che hanno esiti disuguali a causa di correlazioni implicite tra caratteristiche protette e non protette degli individui. Allo stesso modo, si dice che un algoritmo di classificazione eviti un trattamento disparato se è cieco rispetto alle caratteristiche protette dei dati di input, mentre la sua disparità di impatto viene misurata controllando se la proporzione assegnata alla decisione positiva è uguale tra diversi gruppi di individui[8]. Si scopre che gli algoritmi DLP soddisfano entrambi i criteri tecnici, ma un giudice di un tribunale non assegnerebbe agli algoritmi DLP uno status giuridico migliore dell'esplicita disparità di trattamento, poiché sostanzialmente raggiungono la parità di gruppo a scapito dell'ingiustizia individuale. Pertanto, mentre i termini tecnici si ispirano a concetti giuridici, la semplice ottimizzazione dei criteri tecnici può non riuscire a soddisfare i desiderata giuridici ed etici sottesi ai criteri giuridici. Questo esempio illustra la difficoltà di comunicare i desiderata tra diverse discipline e mostra un caso in cui è importante mantenere la distinzione tra terminologia tecnica e giuridica, trovando invece un modo per far interagire fruttuosamente i due linguaggi[9].

L'informatica, e l'intelligenza artificiale in particolare, è un ambito in cui scienza e tecnologia spesso si intrecciano, come testimonia anche la ricca letteratura di filosofia della scienza dedicata all'IA. Osserviamo che la ricerca scientifica è indirizzata alla conoscenza mentre lo sviluppo tecnologico è dedicato alla costruzione di applicazioni. Questi due percorsi si basano su metodologie diverse, che possono interoperare produttivamente se, come nel caso del dialogo interdisciplinare, vengono combinate in modo tale da ottenere cooperazione e integrazione senza perdere la loro specificità. Un aspetto distintivo della scienza è che i suoi risultati sono sempre aperti a essere confutati, invalidati o inglobati in nuovi risultati. Lo scienziato indaga i limiti della conoscenza, cercando di trovare qualcosa di nuovo mettendo in discussione la comprensione consolidata e testandone la robustezza e la replicabilità. Lo sviluppo tecnologico è piuttosto votato a trarre il massimo da un'idea, da una teoria,

[8] Hoffman, A.J. 2016. "Reflections: academia's emerging crisis of relevance and the consequent role of the engaged scholar". *Journal of Change Management* 16(2):77–96.
[9] Crawford, K. 2018. *You and AI - just an engineer: the politics of AI*. Distinguished lecture, Royal Society, London, https://www.youtube.com/watch?v=HPopJb5aDyA

da un risultato scientifico, con lo scopo di produrre un artefatto che sia conveniente rispetto a qualche scopo prefissato. Vale la pena osservare che spetta alla discussione sociale e alla politica definire un quadro giuridico che segni i limiti che gli artefatti tecnologici dovrebbero rispettare. Il rapporto tra scienza e tecnologia è fondamentale anche per affrontare adeguatamente le questioni etiche sollevate dalle tecnologie digitali. Russo invita a ripensare i rapporti tra la conoscenza e le sue applicazioni per evitare il cosiddetto determinismo tecnologico, cioè un percorso predefinito o utopico o distopico. Propone il quadro dell'etica dell'informazione, che affonda le sue radici nella filosofia dell'informazione e si basa sull'idea che non siamo vittime delle tecnologie: non solo costruiamo artefatti digitali discutibili, creiamo anche gli ambienti, le possibilità o le opportunità che sono soggetto anche a una valutazione etica[10]. Dobbiamo quindi prestare attenzione a quali possibilità decidiamo di sviluppare o non sviluppare, diventando responsabili dello spazio di possibilità che creiamo. Una tale visione mette in luce responsabilità diverse implicate a diversi livelli di astrazione, richiedendo un dialogo anche tra l'etica degli scienziati e l'etica degli ingegneri. Per quanto riguarda l'intelligenza artificiale, stanno emergendo nuovi tipi di documenti di ricerca e workshop per ospitare tale dialogo, discutendo argomenti e punti di vista sulle principali questioni nel campo e sul futuro della tecnologia dell'intelligenza artificiale; possono essere visti come esempi concreti di etica sul lavoro. Ad esempio, Gary Marcus prevede una riflessione critica sullo stato dell'arte dei sistemi di deep learning, evidenziando progressi impressionanti, debolezze e malintesi comuni[11]. Japkowicz e Shah sottolineano che la valutazione delle prestazioni di un algoritmo di machine learning non è solo una questione di applicazione della formula matematica corretta, ma è anche un problema di adeguatezza del metodo di valutazione scelto e di interpretazione dei risultati ottenuti[12]. Inoltre, Z. Lipton e J. Steinhardt esaminano la letteratura scientifica sull'apprendimento automatico proponendo una serie di tendenze preoccupanti che ostacolano la ricerca futura e compromettono le basi intellettuali dell'IA[13]. I modelli imperfetti individuati dagli articoli di ricerca sono l'incapacità di distinguere tra spiegazione e speculazione, l'incapacità di identificare le fonti corrette di guadagni empirici, l'uso della matematica in un modo che offusca o impressiona piuttosto che chiarire, e l'uso improprio del linguaggio da parte di scegliendo termini con connotazioni colloquiali o sovraccaricando termini tecnici consolidati. Assumendo un atteggiamento costruttivo, gli autori speculano anche sulle possibili cause dietro le tendenze problematiche e forniscono una discussione su ciò che la comunità di ricerca può fare per aumentare il livello della pratica sperimentale, dell'esposizione e della teoria e per disingannare i ricercatori e il pubblico più ampio.

---

[10] Russo, F. 2018. "Digital technologies, ethical questions, and the need of an informational framework". Philosophy & Technology 31:655-677.

[11] Marcus, G. 2018. *Deep learning: A critical appraisal.* CoRR, abs/1801.00631.

[12] Japkowicz, N. & Shah, M. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge: Cambridge University Press.

[13] Lipton, Z. & Steinhardt, J. 2018. *Troubling trends in machine learning scholarship.* CoRR, abs/1807.03341.

Infine, la complessità dell'impatto dei sistemi di intelligenza artificiale sulla società e sulla vita delle persone richiede un dialogo serio tra ricerca e società. La maggior parte dei moderni sistemi digitali sono meglio qualificati come sistemi sociotecnici, poiché la loro progettazione tecnica è influenzata e ha un impatto sul comportamento degli utenti. Questi sistemi prevedono infatti servizi infrastrutturali nei settori della produzione, degli affari, della comunicazione, dell'intrattenimento, dell'istruzione, dell'urbanistica, dell'accesso alla salute, fino all'accesso alla democrazia e all'esercizio dei diritti umani. Come proposto da K. Crawford[14], oggi l'intelligenza artificiale è tre cose: un insieme di approcci tecnici, un insieme di pratiche sociali che modellano potentemente i sistemi di intelligenza artificiale in base a decisioni non tecniche, come chi lavora su questi sistemi, che decide a quale problema dare priorità, come verrebbero classificati gli esseri umani ed infine un'infrastruttura industriale profondamente concentrata. Pertanto, i ricercatori hanno il dovere di riconoscere la natura sociale, politica ed etica degli artefatti tecnologici e dovrebbero coinvolgere il pubblico dei loro articoli di ricerca, che si è ampliato fino a includere sempre più studenti, giornalisti e politici. Per concludere, abbiamo bisogno sia di una comunità scientifica più alfabetizzata a livello sociale, sia di un pubblico più alfabetizzato dal punto di vista scientifico.

Alla luce della presente analisi, resta problematico capire come sviluppare un dialogo interdisciplinare così efficace[15]. Pensiamo che in questa sfida possano essere illuminanti le idee psicologiche e pedagogiche di John Dewey[16]: con il cosiddetto approccio learning by going, il filosofo americano ha sottolineato il ruolo delle esperienze attive nel cogliere il significato dei concetti. A suo avviso, l'apprendimento è sempre un processo interattivo e sociale, perché la trasmissione concreta della conoscenza avviene attraverso esperienze condivise, dove le parole mostrano in modo più esplicito il significato per cui vengono usate, che abbiamo visto essere un aspetto particolarmente sottile quando si tratta alla interdisciplinarità. Inoltre, secondo Dewey, il metodo di apprendimento non dovrebbe essere imposto né gerarchico (assegnando quindi ad una disciplina una priorità rispetto ad un'altra), ma cooperativo e democratico, in analogia allo spirito del metodo scientifico, che è fatto di verifica, critica e condivisione, finalizzata alla crescita del corpus di conoscenze.

[14] Crawford, K. 2018. *You and AI - just an engineer: the politics of AI*.Distinguished lecture, Royal Society, London, https://www.youtube.com/watch?v=HPopJb5aDyA

[15] Crafa, S. & Pelizzon, L. 2018. *Epistemological questions for a philosophical education in artificial intelligence*. SILF (Societa Italiana di Logica e Filosofia della Scienza) Post-graduate Conference. 2019.

[16] Dewey, J. 1916. *Democracy and Education*. New York: Macmillan.

1.2 Origini filosofiche

Aristotele considerava la razionalità una caratteristica essenziale della mente umana. Il pensiero deduttivo, espresso in termini di sillogismi, era il segno distintivo di tale razionalità, nonché lo strumento intellettuale fondamentale ("organon") di ogni scienza. Forse il contributo più profondo di Aristotele all'intelligenza artificiale è stata l'idea di formalismo. L'idea che certi modelli di pensiero logico siano validi in virtù della loro forma sintattica, indipendentemente dal loro contenuto, è stata un'innovazione estremamente potente, ed è quella nozione che rimane al centro della teoria computazionale contemporanea della mente[17]. Attingendo alla cosiddetta tradizione formalistica di studio della mente, emerge un'idea significativa secondo cui le prestazioni artificiali sono tanto parte delle attività umane quanto quelle naturali e riflettono il continuo tentativo dell'uomo di imitare e riprodurre sé stesso e la natura. Un'interessante caratterizzazione iniziale dell'impulso umano all'auto imitazione si trova già nella mitologia greca, dove gli dèi rappresentano in modo estremo i vizi e le virtù degli esseri umani. Nel periodo medievale, Raimondo Lullo, filosofo e teologo catalano, sostituisce l'imitazione delle caratteristiche visibili dell'uomo con il tentativo di replicarne le facoltà di pensiero. Ispirandosi all'idea di una sorta di macchina pensante di origine araba, Lullo concepisce *l'ars magna*. Secondo le descrizioni, questa avrebbe dovuto essere un vero e proprio dispositivo composto da cerchi concentrici, realizzati con dischi di metallo o gesso. L'obiettivo era ridurre tutte le scienze a principi fondamentali, elementi primi rappresentati da numeri e simboli che, generassero i ragionamenti necessari per risolvere problemi.

Hobbes nel XVII secolo, proclamò che "il raziocinio è calcolo". Più o meno in quella stessa epoca, Leibniz sognava un "calcolo universale" in cui tutte le controversie potessero essere risolte mediante calcoli meccanici. E Cartesio aveva già considerato qualcosa di simile al test di Turing molto prima di Turing, pur adottando una visione piuttosto pessimistica della questione in modo piuttosto disinvolto:

"*Se esistessero macchine che somigliassero al nostro corpo e imitassero le nostre azioni per quanto moralmente possibile, avremmo sempre due prove certissime per riconoscere che, nonostante tutto, non erano veri uomini. Il primo è che non potrebbero mai usare la parola o altri segni come facciamo noi quando mettiamo per iscritto i nostri pensieri a beneficio degli altri. Infatti, possiamo facilmente comprendere che una macchina è costituita in modo tale da poter pronunciare parole, e perfino emettere alcune risposte ad un'azione su di essa di tipo corporeo, che provoca un cambiamento nei suoi organi; per esempio, se viene toccato in una parte particolare può chiederci cosa vogliamo dirgli; se da un'altra parte può esclamare che è ferito, e così via. Ma non accade mai che disponga il suo discorso in vari modi, per rispondere adeguatamente a tutto ciò che si può dire in sua presenza,*

[17] Pylyshyn, Z.: 1989, Computing in Cognitive Science, in M. Posner (ed.), Foundations of Cognitive Science, MIT Press.

come può fare anche l'uomo più basso. E la seconda differenza è che, sebbene le macchine possano eseguire certe cose così bene o forse meglio di quanto chiunque di noi può fare, infallibilmente falliscono in altre, e così possiamo scoprire che non hanno agito per conoscenza, ma solo per scopo. la disposizione dei loro organi. Infatti, mentre la ragione è uno strumento universale che può servire per tutte le contingenze, questi organi hanno bisogno di qualche adattamento speciale per ogni azione particolare. Da ciò consegue che è moralmente impossibile che in qualunque macchina vi sia sufficiente diversità per consentirle di agire in tutti gli eventi della vita nello stesso modo in cui la nostra ragione ci fa agire".

In considerazione del significato storicamente attribuito alla deduzione in filosofia, l'idea stessa di macchina intelligente equivaleva spesso a una macchina in grado di eseguire inferenze logiche: una macchina in grado di trarre validamente conclusioni da determinate premesse. La dimostrazione automatizzata di teoremi, come è conosciuta oggi, è stata quindi parte integrante dell'intelligenza artificiale fin dall'inizio, sebbene, come vedremo, la sua rilevanza sia stata oggetto di accesi dibattiti, soprattutto negli ultimi due decenni. In generale, il problema della meccanizzazione della detrazione può essere formulato in tre forme diverse. Elencati in ordine di difficoltà crescente, abbiamo:

• Controllo della dimostrazione: data una deduzione D che pretende di derivare una conclusione P da un numero di premesse P1, …, Pn, decidi se D è suono o meno.

• Scoperta della dimostrazione: dato un numero di premesse P1, …, Pn e una presunta conclusione P, decidono se P segue logicamente dalle premesse e, in tal caso, ne producono una dimostrazione formale.

• Generazione di congetture: Dato un numero di premesse P1, …, Pn, dedurre una conclusione "interessante" P che segue logicamente dalle premesse, e produrne una dimostrazione.

Tecnicamente parlando, il primo problema è il più semplice. Nel caso della logica dei predicati con uguaglianza, il problema di verificare la validità di una data dimostrazione non è solo risolvibile algoritmicamente, ma risolvibile in modo abbastanza efficiente. Tuttavia, il problema è carico di interessanti questioni filosofiche e tecniche, e la sua rilevanza per l'intelligenza artificiale fu presto compresa da McCarthy (1962)[18], il quale scrisse che "il controllo delle dimostrazioni matematiche è potenzialmente una delle applicazioni più interessanti e utili dei computer automatici". Ad esempio, nella misura in cui si suppone che le dimostrazioni esprimano il ragionamento, possiamo chiederci se il formalismo in cui è espressa la dimostrazione di input D fornisce un buon modello formale di ragionamento deduttivo. Le dimostrazioni formali in stile Hilbert (lunghi elenchi di formule, ciascuna delle quali è un assioma o segue dalle formule precedenti mediante una delle poche regole di inferenza) erano importanti come strumenti per le indagini metamatematiche, ma non catturavano il

[18] McCarthy, J.: 1962, Computer programs for checking mathematical proofs, Proceedings of the Symposium in Pure Math, Recursive Function Theory, Vol. V, American Mathematical Society, Providence, RI, pp. 219–228.

ragionamento deduttivo praticato da umani. Ciò fornì l'incentivo per importanti ricerche sui formalismi logici che rispecchiavano il ragionamento umano, in particolare quello condotto dai matematici. S. J´askowski (1934) ideò un sistema di deduzione naturale che ebbe molto successo sotto questo aspetto[19]. Gentzen (1969) scoprì indipendentemente sistemi simili, ma con differenze cruciali rispetto al lavoro di J´askowski[20]. Le idee di deduzione naturale introdotte da J´askowski e Gentzen, in seguito, giocarono un ruolo chiave, non solo nella dimostrazione di teoremi e nell'intelligenza artificiale, ma anche nell'intelligenza artificiale. anche le scienze cognitive computazionali. La logica mentale[21], in particolare, una famiglia di teorie cognitive computazionali del ragionamento deduttivo umano, è stata fortemente influenzata dalla deduzione naturale. Il secondo problema è notevolmente più difficile. I primi risultati della teoria delle funzioni ricorsive (Turing 1936, Church 1936)[22] stabilirono che non esiste una macchina di Turing che possa decidere se una formula arbitraria della logica del primo ordine è valida (questo era l'Entscheidungsproblem di Hilbert). Pertanto, dalla tesi di Church, ne consegue che il problema è algoritmicamente irrisolvibile: non esiste un metodo meccanico generale in grado di prendere sempre la decisione giusta in un periodo di tempo finito. Tuttavia, anche gli esseri umani non hanno alcuna garanzia di risolvere sempre il problema (e anzi spesso non riescono a farlo). Di conseguenza, l'intelligenza artificiale può cercare approssimazioni conservatrici che siano le migliori possibili: programmi che diano la risposta giusta il più spesso possibile, e altrimenti non danno alcuna risposta (o fallendo esplicitamente, oppure andando avanti indefinitamente finché non fermarli). Il problema fu affrontato fin dall'inizio per i formalismi più deboli con risultati apparentemente promettenti: The Logic Theorist (LT) di Newell, Simon e Shaw, presentato alla conferenza inaugurale dell'AI del 1956 a Dartmouth menzionata in precedenza, riuscì a dimostrare 38 dei 52 principi proposizionali. teoremi logici dei Principia Mathematica. Altri primi sforzi degni di nota includevano l'implementazione dell'aritmetica di Presburger da parte di Martin Davis nel 1954 presso l'Institute for Advanced Studies di Princeton (Davis 2001), la procedura Davis-Putnam (M. Davis e H. Putnam 1960), le cui variazioni sono utilizzate oggi in molti dimostratori basati sulla soddisfacibilità e un impressionante sistema per la logica del primo ordine costruito da Wang (1960).

La scoperta dell'unificazione e del metodo di risoluzione da parte di Robinson (Robinson 1965) ha dato un notevole impulso al campo. La maggior parte dei dimostratori di teoremi automatizzati oggi

[19] J´askowski S.: 1934, On the rules of suppositions in formal logic, Studia Logica 1.

[20] Gentzen, G.: 1969, The collected papers of Gerhard Gentzen, North-Holland, Amsterdam, Holland. English translations of Gentzen's papers, edited and introduced by M. E. Szabo.

[21] Osherson, D. N.: 1975, Logical Abilities in Children, volume 3, Reasoning in Adolesescence: Deductive Inference, Lawrence Erlbaum Associates; Braine, M. D. S. and O'Brien, D. P. (eds): 1998, Mental Logic, Lawrence Erlbaum Associates; Rips, L. J.: 1994, The Psychology of Proof, MIT Press.

[22] Turing, A. M.: 1936, On Computable Numbers with Applications to the Entscheidungsproblem, Proceedings of the London Mathematical Society 42, 230–265; Church, A.: 1936, An Unsolvable Problem of Elementary Number Theory, American Journal of Mathematics 58, 345–363.

si basano sulla risoluzione. Altri formalismi importanti includono i tableaux semantici e la logica equazionale (Robinson e Voronkov 2001)[23]. Sebbene negli ultimi dieci anni siano stati compiuti notevoli progressi, in gran parte stimolati dalla competizione annuale del sistema CADE ATP, gli ATP più sofisticati oggi continuano a essere fragili e spesso falliscono su problemi che sarebbero banali per gli studenti universitari. Il terzo problema, quello della generazione di congetture, è il più difficile, ma è anche il più interessante. Di fronte a un insieme di informazioni, gli esseri umani, in particolare i matematici, escogitano regolarmente congetture interessanti e poi spesso si propongono di dimostrarle, di solito con successo. Questo processo di scoperta è una delle attività più creative dell'intelletto umano. L'assoluta difficoltà di simulare questa creatività a livello computazionale è sicuramente una delle ragioni principali per cui l'intelligenza artificiale ha ottenuto progressi piuttosto minimi in questo campo. Ma un'altra ragione è che per gran parte del secolo precedente, i logici e i filosofi si occupavano quasi esclusivamente della giustificazione piuttosto che della scoperta. Ciò si applicava non solo al ragionamento deduttivo ma anche a quello induttivo, e in effetti alla teorizzazione scientifica in generale[24]. Era opinione diffusa che il processo di scoperta dovesse essere studiato dagli psicologi, non dai filosofi e dai logici. È interessante notare che prima di Frege questo non era il caso. Filosofi come Cartesio, Bacon, Mill e Peirce avevano tutti tentato di studiare razionalmente il processo di scoperta e di formulare regole per guidarlo. A partire da Hanson (1958) in scienze e con Lakatos (1976) in matematica, i filosofi hanno iniziato a enfatizzare nuovamente la scoperta[25]. I ricercatori sull'intelligenza artificiale hanno anche tentato di modellare la scoperta computazionalmente, sia in scienza che in matematica, e questa linea di lavoro ha portato a innovazioni di apprendimento automatico nell'intelligenza artificiale come la programmazione genetica[26] e la programmazione logica induttiva (Muggleton 1992). Tuttavia, i successi sono stati limitati e le obiezioni fondamentali ai trattamenti algoritmici della scoperta e della creatività in generale – ad esempio, come avanzata da Hempel (1985) – rimangono taglienti. Una questione importante è il carattere apparentemente olistico dei processi cognitivi superiori come il ragionamento creativo, e la difficoltà di formulare una rigorosa caratterizzazione della pertinenza.

Senza una nozione precisa di rilevanza, che sia suscettibile di implementazione computazionale, sembra esserci poca speranza di progresso sul problema della generazione delle conclusioni, o su qualsiasi altro problema simile, compresa la generazione di concetti e la formazione di ipotesi abduttive. Di fronte a progressi relativamente scarsi sui problemi del ragionamento difficile, e forse influenzati da varie altre critiche all'IA simbolica, alcuni ricercatori sull'IA hanno lanciato seri

[23] Robinson, A. and Voronkov, A. (eds): 2001, Handbook of Automated Reasoning, Vol. 1, North-Holland
[24] Reichenbach, H.: 1938, Experience and Prediction, University of Chicago Press
[25] Hanson, N. R.: 1958, Patterns of Discovery, Cambridge University Press; Lakatos, I.: 1976, Proofs and refutations: the logic of mathematical discovery, Cambridge University Press.
[26] Koza, J.: 1992, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press

attacchi alla logica formale, che hanno criticato come un sistema eccessivamente rigido che non fornire un buon modello dei meccanismi di ragionamento umano, che sono eminentemente flessibili. Di conseguenza hanno cercato di spostare l'attenzione e gli sforzi del settore lontano dal rigoroso ragionamento deduttivo e induttivo, rivolgendoli invece al "ragionamento basato sul buon senso". Ad esempio, Minsky (1986) scrive: "*Per generazioni, scienziati e filosofi hanno cercato di spiegare il ragionamento ordinario in termini di principi logici praticamente senza successo. Sospetto che questa impresa sia fallita perché guardava nella direzione sbagliata: il senso comune funziona così bene non perché sia un'approssimazione della logica; la logica è solo una piccola parte del nostro grande accumulo di modi diversi e utili per concatenare le cose*"[27].

1.3 Origini dell'IA

All'inizio degli anni '40, il termine cibernetica iniziò a essere utilizzato per descrivere lo studio sistematico dei processi di comunicazione e controllo sia negli esseri viventi che nelle macchine. Nel 1943, Warren McCulloch e Walter Pitts introdussero il primo modello di neuroni artificiali, basandosi sulle conoscenze della fisiologia neuronale, sulla logica proposizionale e sulla teoria della computabilità di Turing. L'obiettivo della cibernetica era comprendere i meccanismi di autoregolazione e controllo presenti negli organismi viventi e nelle macchine a retroazione, capaci di adattarsi alle stimolazioni ambientali modificando il proprio comportamento. Uno dei risultati significativi di questo approccio fu la dimostrazione che qualsiasi funzione calcolabile poteva essere eseguita da una rete di neuroni connessi. Nel 1949, Donald Hebb mostrò come una semplice regola di aggiornamento delle connessioni neuronali potesse innescare processi di apprendimento. Nonostante questi successi iniziali, la cibernetica iniziò a perdere rilevanza verso la metà degli anni '50, poiché l'interesse e le risorse si spostarono quasi interamente verso l'intelligenza artificiale. Questo declino fu dovuto al disinteresse della cibernetica per i progressi dell'informatica e alla limitazione degli obiettivi iniziali. Tuttavia, la tradizione cibernetica rivisse negli anni '80 con la rinascita del paradigma delle reti neurali all'interno dell'intelligenza artificiale. Dato che per avere successo l'IA necessita di un sistema artificiale in grado di replicare, emulandoli, i fenomeni dell'intelligenza, l'elaboratore è stato fin dall'inizio ritenuto il candidato ideale per questo compito. Alla base della sua invenzione c'è il concetto della macchina universale di Turing, una macchina teorica capace di assumere un numero finito di stati e di eseguire una serie limitata di operazioni, permettendo così di rappresentare qualsiasi procedura definita. La macchina di Turing è costituita da

[27] Minsky, M.: 1986, The Society of Mind, Simon and Schuster, p. 167 cit.

un nastro infinito diviso in celle, che può essere letto da una testina mobile in grado di spostarsi avanti e indietro, e da un'unità di controllo che legge il simbolo presente nella cella sotto la testina. L'azione che la macchina esegue in un dato momento è determinata dal simbolo letto e dallo stato attuale della macchina. Dopo aver letto il simbolo su una cella, la testina può o lasciarlo invariato oppure cancellarlo e stamparne un altro. Basandosi sulla macchina universale di Turing, il concetto di algoritmo può essere definito come la sequenza di operazioni eseguite da questa macchina.

Se nei secoli XVIII, XIX e XX la formalizzazione delle scienze e della matematica aveva preparato il terreno per lo studio dell'intelligenza e delle sue possibili versioni artificiali, è stato solo con l'avvento dei primi computer elettronici, intorno alla Seconda guerra mondiale, che tale interesse ha potuto prendere forma concreta, portando alla definizione del programma di ricerca delineato nel seminario di Dartmouth del 1956. È proprio questa, infatti, la data ufficiale della nascita dell'Intelligenza Artificiale, in occasione della conferenza estiva al Dartmouth College, ad Hanover, nel New Hampshire, nella quale si riunirono un gruppo di studiosi con lo scopo di "esaminare la congettura che ogni aspetto dell'apprendimento o ogni altra caratteristica dell'intelligenza possa essere, in linea di principio, descritto in modo tanto preciso che si possa far sì che una macchina lo simuli". La celebrazione del cinquantesimo anniversario di questa conferenza, AI@50, si è tenuta nel luglio 2006 a Dartmouth, con il ritorno di cinque dei partecipanti originali. Parteciparono dieci pensatori, tra cui John McCarthy (che lavorava a Dartmouth nel 1956), Claude Shannon, Marvin Minsky, Arthur Samuel, Trenchard Moore (apparentemente il partecipante più giovane e l'unico a prendere appunti alla conferenza originale), Ray Solomonoff, Oliver Selfridge, Allen Newell e Herbert Simon. Dal punto in cui ci troviamo ora, all'inizio del nuovo millennio, la conferenza di Dartmouth è memorabile principalmente per due ragioni: in primo luogo, il termine "intelligenza artificiale" è stato coniato lì (ed è rimasto a lungo saldamente radicato, nonostante non gli piaccia questo giorno da alcuni dei partecipanti, ad esempio Moore); in secondo luogo, Newell e Simon hanno rivelato che un programma – Logic Theorist (LT) – concordato dai partecipanti alla conferenza (e, in effetti, da quasi tutti coloro che ne vennero a conoscenza subito dopo l'evento di Dartmouth) era un risultato notevole. La LT era in grado di dimostrare teoremi elementari nel calcolo proposizionale ed era considerata un passo notevole verso la traduzione del ragionamento a livello umano in calcoli concreti. Gli anni successivi al seminario di Dartmouth costituirono un periodo di grandi aspettative, alimentate dai progressi significativi e dall'incredibile crescita delle tecnologie informatiche. In questo contesto, emergono due tendenze principali: da un lato, il gruppo guidato da Newell, Shaw e Simon, che si concentrava sulla simulazione dei processi cognitivi umani attraverso l'elaboratore. Con il GPS (General Problem Solver) del 1958, cercarono di ampliare il campo delle applicazioni del programma oltre i semplici problemi logici, rappresentando il cosiddetto paradigma della simulazione. Dall'altro lato, vi erano coloro che si dedicavano a migliorare le prestazioni dei

programmi, senza preoccuparsi se queste emulassero o meno i processi umani, caratterizzando così il paradigma della prestazione o dell'emulazione.

Durante questo periodo, i modelli a reti neurali subirono un temporaneo declino, in gran parte a causa della critica di Marvin Minsky al Perceptron di Frank Rosenblatt, che si dimostrò incapace di riconoscere anche stimoli visivi molto semplici. Nel frattempo, il paradigma della prestazione registrava successi, come lo sviluppo dei primi programmi per il gioco della dama e degli scacchi, e i programmi basati sulla rappresentazione della conoscenza introdotta dall'Advice Taker di John McCarthy. Quest'ultimo, sebbene mai realizzato, rappresentava il primo tentativo di creare un sistema completo di IA capace di gestire modifiche complesse in modo semplice e di operare con una nozione parziale di successo, riconoscendo le difficoltà nel risolvere problemi più complessi. Tuttavia, non passò molto tempo prima che i ricercatori in IA iniziassero a confrontarsi con i primi fallimenti: i metodi che funzionavano bene per esempi semplici si rivelarono del tutto inadeguati quando applicati a casi più complessi e su larga scala. Le grandi speranze iniziali furono rapidamente deluse, soprattutto a causa del fallimento dei progetti di traduzione automatica tra lingue naturali. I programmi che si limitavano alla manipolazione sintattica dimostrarono di non essere all'altezza, portando al ritiro delle ingenti sovvenzioni da parte dei governi americano e britannico.

Un altro problema significativo fu l'incapacità di gestire l'esplosione combinatoria: affrontare problemi più complessi non poteva essere risolto semplicemente con hardware più veloce e memoria più ampia, poiché esistevano limitazioni insormontabili insite nella natura stessa di questi problemi. Questa "dose di realtà" portò, a partire dal 1970, a un cambiamento di focus verso aree di competenza più ristrette, culminando nello sviluppo dei primi sistemi esperti. In questi sistemi, una conoscenza approfondita e dettagliata dello specifico dominio divenne cruciale. Dopo il ridimensionamento delle aspettative, si assiste all'inizio degli anni '80 alla nascita dell'IA come settore industriale. Nel 1982 viene sviluppato il primo sistema esperto commerciale di successo, utilizzato per supportare la configurazione degli ordini per nuovi sistemi di elaborazione in un'azienda produttrice. L'intelligenza artificiale diventa così parte integrante di uno sforzo più ampio, che include la progettazione di chip e la ricerca sulle interfacce uomo-macchina.

Contemporaneamente, si osserva il ritorno dell'approccio basato sulle reti neurali. Intorno al 1985, quattro gruppi di ricerca indipendenti riscoprono un algoritmo di apprendimento, basato sulla retropropagazione dell'errore, che era stato originariamente scoperto quindici anni prima, e lo applicano con successo a numerosi problemi di apprendimento in informatica e ingegneria. Questo rinnovato interesse è anche sostenuto dalla nascita di una nuova disciplina, le scienze cognitive, che nel 1979 si affermano ufficialmente come un campo autonomo, integrando molte delle aspirazioni di una parte della psicologia e di quella corrente dell'IA che ha sempre visto nella macchina uno strumento privilegiato per lo studio della mente.

Ma mentre l'inaugurazione cerimoniale dell'IA potrebbe essere stata la conferenza di Dartmouth del 1956, e mentre i filosofi potrebbero aver riflettuto per secoli sulle macchine e sull'intelligenza, le origini concettuali chiave dell'IA possono essere trovate all'intersezione di due dei più importanti sviluppi intellettuali del mondo nel XX secolo:

- la "rivoluzione cognitiva" iniziata a metà degli anni Cinquanta e che ha rovesciato il comportamentismo e riabilitato la psicologia mentalistica;
- la teoria della computabilità che era stata sviluppata nei due decenni precedenti da pionieri come Turing, Church, Kleene e Gödel.

Il significato di ciascuno per l'intelligenza artificiale sarà discusso brevemente di seguito. La rivoluzione cognitiva è tipicamente associata al lavoro di George Miller e Noam Chomsky negli anni '50, in particolare alla famigerata revisione da parte di quest'ultimo della teoria del linguaggio di Skinner[28]. Era stato anticipato negli anni Quaranta da McCulloch e Pitts (1943)[29] e da altri pionieri della cibernetica che già avevano messo in luce le somiglianze tra il pensiero umano e l'elaborazione delle informazioni, nonché dai risultati sperimentali ottenuti da psicologi come Tolman (1948), che, studiando la navigazione nei labirinti da parte dei ratti, hanno presentato prove dell'esistenza di "mappe cognitive". Particolarmente influente è stata la famosa argomentazione di Chomsky sulla "povertà di stimoli", secondo cui l'efficienza e la rapidità dell'acquisizione del linguaggio durante l'infanzia non possono essere spiegate esclusivamente facendo appello ai magri dati a cui i bambini sono esposti nei loro primi anni; piuttosto, costringono a postulare regole mentali e rappresentazioni innate che codificano la competenza linguistica. Forti prove dell'esistenza di rappresentazioni mentali sono state fornite anche da scoperte sperimentali relative alla memoria, come i risultati di Sperling (1960), che hanno indicato che gli esseri umani tipicamente immagazzinano più informazioni di quelle che possono riportare. La memoria, dopo tutto, fornisce forse il caso più chiaro di rappresentazione mentale; sembra assurdo negare che le persone memorizzino informazioni, cioè che abbiamo una sorta di rappresentazione interna di informazioni come l'anno in cui siamo nati oi nomi dei nostri genitori. Tutto ciò è di buon senso fino a diventare banale, così come lo è l'affermazione che le persone abitualmente parlano come se avessero davvero convinzioni, speranze, desideri, ecc.; e in effetti la maggior parte dei comportamentisti non avrebbe negato queste affermazioni. Ciò che negavano era la legittimità teorica di spiegare il comportamento umano presupponendo entità mentali non osservabili (come i ricordi), o che la terminologia intenzionale avesse qualche posto in una scienza della mente. Essenzialmente una dottrina positivista, il comportamentismo aveva una diffidenza verso tutto ciò che non poteva essere osservato direttamente e un'avversione generale per

[28] Chomsky, N.: 1996 [1959], A Review of B. F. Skinner's 'Verbal Behavior', in H. Geirsson and M. Losonsky (eds), Readings in Language and Mind, Blackwell, pp. 413–441.
[29] McCulloch, W. S. and Pitts, W. A.: 1943, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5, 115–133.

la teoria. È stato il paradigma dominante in psicologia per gran parte del XX secolo, fino alla metà degli anni Cinquanta, finché non è stato definitivamente detronizzato dal nuovo approccio "cognitivo".

Una volta compiuti i primi passi e quando le rappresentazioni mentali furono apertamente ammesse nella teorizzazione scientifica sulla mente, la "metafora del computer" divenne matura per l'esplosione. Dopotutto, era noto che i computer immagazzinavano dati strutturati nelle loro memorie e risolvevano problemi interessanti manipolando tali dati in modo sistematico, eseguendo istruzioni appropriate. Forse un modello simile potrebbe spiegare – ed eventualmente aiutare a duplicare – il pensiero umano. In effetti, la postulazione delle rappresentazioni mentali non andrebbe lontano da sola se la loro efficacia causale non potesse essere spiegata in modo meccanicistico e sistematico. Ammesso che le rappresentazioni mentali strutturate siano necessarie per la cognizione di ordine superiore; ma in che modo tali rappresentazioni causano effettivamente pensiero e azione razionali? La teoria del calcolo è stata utilizzata proprio per soddisfare questa importante esigenza teorica. Il risultato divenne noto come teoria computazionale della mente (in breve CTM), una dottrina che è stata indissolubilmente legata all'intelligenza artificiale forte. La prima idea centrale del marchio comunitario è quella di spiegare gli stati mentali intenzionali dando una svolta computazionale all'analisi di Russell (1940)[30] di frasi intenzionali come "Tom crede che 7 sia un numero primo" come atteggiamenti proposizionali che implicano un atteggiamento psicologico A (in questo caso credendo) verso una proposizione P (in questo caso, che 7 è primo). Più precisamente, trovarsi in uno stato mentale che comporta un atteggiamento A e una proposizione P significa trovarsi in una certa relazione RA con una rappresentazione mentale MP il cui significato è P. In parole povere, credere che 7 sia un numero primo è avere una rappresentazione mentale nella tua "scatola delle convinzioni" che significa che 7 è primo. La rappresentazione stessa è simbolica. Cioè, la tua "scatola delle convinzioni" contiene un segno di una struttura simbolica il cui significato (o "contenuto") è che 7 è primo. Pertanto, le rappresentazioni mentali hanno sia sintassi che semantica, proprio come le frasi delle lingue naturali. Costituiscono un "linguaggio del pensiero", per così dire, o mentalese. Ma è solo la loro sintassi – essendo la sintassi in definitiva riducibile alla forma fisica – che li rende causalmente efficaci.

Questa è una storia plausibile perché, come hanno dimostrato i lavori sulla logica e sulla computabilità, esistono trasformazioni puramente sintattiche di strutture simboliche che sono tuttavia sensibili alla semantica. Le dimostrazioni deduttive forniscono forse l'esempio migliore: manipolando le formule esclusivamente sulla base delle loro proprietà sintattiche, è possibile estrarre da esse altre formule che ne conseguono logicamente. La sintassi può quindi rispecchiare la semantica o, come dice Haugeland (1985a, p. 106), "se ti prendi cura della sintassi, la semantica si prenderà cura di sé

[30] Russell, B.: 1940, An inquiry into meaning and truth, George Allen and Unwin.

stessa". Secondo questo modello, un processo mentale è una sequenza di rappresentazioni mentali che esprimono il contenuto proposizionale dei pensieri corrispondenti. Le cause e gli effetti di ciascuna rappresentazione mentale, ciò che essa può effettivamente fare, è determinato dalla sua sintassi "nello stesso modo in cui la geometria di una chiave determina quali serrature aprirà" (Fodor 1987, p. 19). E l'intero processo è orchestrato da un algoritmo, un insieme di istruzioni che determina il modo in cui le rappresentazioni si susseguono nel corso del pensiero complessivo. Questa è la seconda idea centrale del marchio comunitario. La mente è quindi vista come un "motore sintattico" che guida un motore semantico e, almeno in linea di principio, il suo funzionamento può essere duplicato su un computer. Un'estensione naturale della MC è il funzionalismo della macchina di Turing, che fu adombrato per la prima volta da Putnam (1960) in un autorevole articolo che contribuì a portare avanti la rivoluzione cognitiva (almeno nei circoli filosofici), a minare il comportamentismo e a modellare la prospettiva dell'AI[31]. Il funzionalismo in generale è, grossomodo, l'idea che l'essenza di uno stato mentale non si trova nella biologia del cervello (o nella fisica che sottoscrive l'hardware della sua CPU, nel caso di una macchina) ma piuttosto nel ruolo che lo stato gioca nella vita mentale (o nei calcoli), e in particolare nelle relazioni causali che ha con gli stimoli (input), il comportamento (output) e altri stati mentali (computazionali). Il funzionalismo della macchina di Turing, in particolare, è l'idea che la mente sia essenzialmente una gigantesca macchina di Turing il cui funzionamento è specificato da una serie di istruzioni che impongono che se la mente è in un certo stato s e riceve un certo input x, si verifica una transizione portato allo stato s 0 e viene emessa un'uscita y[32]. Le versioni più popolari – e meno plausibili – del funzionalismo della macchina di Turing consentono transizioni probabilistiche. Strettamente correlata al marchio comunitario (in realtà più forte di esso) è l'ipotesi del sistema di simboli fisici (PSSH) avanzata da Newell e Simon (1976)[33]. Secondo esso, un sistema di simboli fisici "ha i mezzi necessari e sufficienti per un'azione intelligente generale"[34], laddove un sistema di simboli fisici è "una macchina che produce attraverso il tempo un insieme in evoluzione di strutture simboliche", un la struttura simbolica è una raccolta di token simbolici "correlati in qualche modo fisico (come un token accanto a un altro)" e soggetto a una varietà di operazioni sintattiche, in particolare "creazione, modifica, riproduzione e distruzione". Newell e Simon consideravano le macchine che eseguivano programmi di elaborazione di liste del tipo Lisp come esempi prototipici di sistemi di simboli fisici. Sebbene ci siano stati vari disaccordi interni (ad esempio, relativi a questioni di innatezza), in una forma o nell'altra il marchio comunitario, il PSSH e il funzionalismo della macchina di Turing insieme caratterizzano vagamente l'IA "classica"

[31] Putnam, H.: 1960, Minds and machines, in S. Hook (ed.), Dimensions of Mind, New York University Press, pp. 138–164
[32] Turing, A.: 1950, Computing machinery and intelligence, Mind LIX (59) (236), 433–460.
[33] Newell, A. and Simon, H. A.: 1976, Computer Science as Empirical Inquiry: Symbols and Search, Communications of the ACM 19, 113–126.
[34] Ibidem.

o "simbolica", o ciò che Haugeland ha soprannominato GOFAI ("*buona intelligenza artificiale vecchio stile*")[35]. Tutti e tre furono posti come tesi empiriche sostanziali, il MTC e il funzionalismo della macchina di Turing sulla mente umana e il PSSH sull'intelligenza in generale (anche GOFAI, è stata esplicitamente caratterizzata da Haugeland come una dottrina empirica della scienza cognitiva). Stabiliscono i parametri e gli obiettivi per la maggior parte della ricerca sull'intelligenza artificiale per almeno i primi tre decenni del settore. Continuano ad avere un'influenza dominante, anche se, come vedremo, non sono più l'unico gioco in campo, avendo subito notevoli battute d'arresto a seguito di attacchi violenti che hanno elaborato seri problemi concettuali ed empirici con l'approccio GOFAI. Dunque, i pensieri complessi sono rappresentati da strutture simboliche complesse più o meno allo stesso modo in cui, sia nei linguaggi naturali che nelle logiche formali, le frasi complesse sono costruite ricorsivamente da componenti più semplici. Questi componenti sono in qualche modo assemblati insieme (e alla fine la scienza dovrebbe essere in grado di spiegare i dettagli di come tali operazioni simboliche vengono eseguite nel cervello) per formare il pensiero complesso. Ora, una storia compositiva di questo tipo – simile alla semantica compositiva sostenuta da Frege e Tarski – è praticabile solo se esiste un inventario di elementi primitivi che possono essere utilizzati come elementi costitutivi finali di rappresentazioni più complesse. La questione centrale per il marchio comunitario, che ha un analogo diretto nell'intelligenza artificiale, è la questione di come questi primitivi acquisiscano significato. Più precisamente, la questione è come i primitivi mentalesi all'interno del nostro cervello (o all'interno della CPU di un robot) riescano a relazionarsi con oggetti e stati di cose al di fuori del nostro cervello: oggetti che potrebbero anche non esistere e stati di cose che potrebbero anche non esistere.

Questo è chiamato anche problema della messa a terra del simbolo (Harnad 1990)[36]: non si tratta semplicemente di un enigma filosofico sulla mente umana, e nemmeno una questione protoscientifica di psicologia; ha implicazioni ingegneristiche dirette per l'intelligenza artificiale, poiché una risposta plausibile potrebbe tradursi in una metodologia per costruire un robot che potenzialmente eviti alcune delle obiezioni più devastanti al marchio comunitario; cioè, un robot che "pensa" eseguendo calcoli su strutture simboliche formali (come presumibilmente facciamo noi, secondo CTM), ma è tuttavia sufficientemente radicato nel mondo reale da poter dire che raggiunga una comprensione extra-simbolica. Chiaramente non è sostenibile suggerire che l'evoluzione ci abbia dotato di tutti i giusti simboli primitivi con tutti i giusti significati incorporati, dal momento che l'evoluzione non avrebbe potuto prevedere termostati, o satelliti. In risposta sono state esposte numerose teorie, tutte rientranti nella bandiera della "naturalizzazione del contenuto" o della "naturalizzazione della semantica" o della "naturalizzazione dell'intenzionalità". L' obiettivo è fornire un resoconto fisicalistico di come i

[35] Haugeland, J.: 1985, AI: The Very Idea, MIT Press, p. 112 cit.
[36] Harnad, S.: 1990, The Symbol Grounding Problem, Physica D 42, 335–346.

simboli mentali nella nostra testa riescono a rappresentare cose che sono esterne a noi (o, inquadrando la questione indipendentemente dal marchio comunitario, come gli stati mentali possono acquisire significato). Il tipo di spiegazione ricercata, in altre parole, è riduttiva e materialistica; dovrebbe essere espresso nel vocabolario non intenzionale della scienza fisica pura.

L'essenza delle teorie informazionali è la nozione di covarianza. L'idea è che se una quantità x varia sistematicamente con una quantità y, allora x trasporta informazioni su y. Il tachimetro di un'auto varia sistematicamente con la velocità dell'auto e quindi trasporta informazioni su di essa. Di conseguenza, possiamo vedere il tachimetro come un sistema intenzionale, in quanto le sue letture riguardano la velocità dell'auto. Questo è il senso della parola "significa" che Grice chiamava significato naturale, una nozione che presagiva teorie sulla semantica informazionale. Per quanto riguarda la semantica mentalese, l'intuizione centrale di tali teorie - detta in modo un po' semplicistico - è che il significato di un simbolo è determinato da qualunque cosa le rappresentazioni di quel simbolo sistematicamente (nomologicamente) covarino. Se un segno di un certo simbolo mentale H appare nel nostro cervello ogni volta che un cavallo appare davanti a noi, allora H trasporta informazioni sui cavalli e quindi significa cavallo. L'idea ha un fascino semplice ma è vulnerabile a numerose obiezioni; ne citeremo solo tre. Il primo sono oggetti inesistenti come gli unicorni e oggetti astratti come la radice quadrata di due. Come possono effettivamente causare qualcosa e come possono nomologicamente covariare con gli stati cerebrali? Il secondo è il cosiddetto problema della disgiunzione. Non è possibile che il significato di un simbolo sia qualunque sia la causa delle sue rappresentazioni (qualunque cosa con cui tali rappresentazioni covariano sistematicamente), perché il suddetto simbolo H, per esempio, potrebbe essere sistematicamente causato non solo da cavalli reali ma anche da mucche che appaiono davanti a noi di notte o in altre condizioni ambientali opportunamente fuorvianti. Una caratteristica principale dell'intenzionalità, almeno della consueta varietà mentale, è che qualsiasi sistema che rappresenta deve essere anche in grado di travisare, di intrattenere pensieri errati sulla realtà. Definizioni causali disgiuntive del tipo di cui sopra lo precluderebbero. Un altro problema con le false dichiarazioni è che a volte un cavallo reale non riesce a causare H (e se le condizioni sono giuste questo sarà sistematico, non un fallimento una tantum); quindi non può essere il caso che H venga segnato se e solo se un cavallo appare davanti a noi[37].

Le teorie evoluzionistiche sostengono, grosso modo, che gli stati intenzionali sono adattamenti, allo stesso modo in cui lo sono i fegati e i pollici, e che il contenuto (significato) di uno stato intenzionale è la funzione per la quale è stato selezionato, cioè lo scopo a cui serve.

Le tre principali critiche filosofiche all'IA forte che hanno contribuito a cambiare la tendenza nella comunità dell'IA e ad indicare nuove direzioni di ricerca sono le seguenti: La critica di Hubert

[37] Fodor, J. A. and Pylyshyn, Z. W.: 1988, Connectionism and Cognitive Architecture: A Critical Analysis, Cognition 28, 139–196; Dretske, F.: 1988, Explaining Behavior: Reason in a World of Causes, MIT Press.

Dreyfus; La critica di Block al funzionalismo delle macchine attraverso gli esperimenti mentali sul cervello cinese; e L'esperimento mentale della stanza cinese di Searle. Tutti e tre sono emersi a distanza di dieci anni l'uno dall'altro. La critica di Dreyfus è stata la prima. Si trattava di un insieme di argomentazioni empiriche e filosofiche riguardo il fallimento nel costruire sistemi intelligenti di uso generale. Questa linea di critica è stata generalmente respinta come non valida e ingiusta perché nella migliore delle ipotesi dimostrava che l'intelligenza artificiale non aveva ancora avuto successo, non che non avrebbe mai potuto avere successo, e ingiusta perché l'intelligenza artificiale era un sistema intelligente campo molto giovane, e non ci si potevano aspettare scoperte tecnologiche rivoluzionarie da un campo nella sua infanzia, nonostante i proclami eccessivamente entusiastici di alcuni dei suoi pionieri. Filosoficamente, Dreyfus ha sostenuto che l'intelligenza artificiale è un tentativo mal concepito di implementare un programma razionalista che risale almeno a Leibniz e Hobbes, un progetto che si basa sul fuorviante principio "cartesiano" secondo cui la comprensione umana consiste nel formare e manipolare rappresentazioni simboliche. Al contrario, ha sostenuto che la nostra capacità di comprendere il mondo e le altre persone è un tipo di abilità di know-how non dichiarativa che non è suscettibile di codificazione proposizionale. È inarticolato, preconcettuale e possiede una dimensione fenomenologica indispensabile che non può essere catturata da nessun sistema basato su regole. Le persone non raggiungono un comportamento intelligente nella loro vita quotidiana memorizzando grandi quantità di fatti e seguendo regole esplicitamente rappresentate. Nascere, essere dotati di un corpo e della capacità di sentire, crescere come parte di una società sono elementi essenziali di intelligenza e di comprensione. Dreyfus ha anche sottolineato l'importanza di capacità come l'immaginazione, la tolleranza all'ambiguità e l'uso della metafora, così come fenomeni come la coscienza marginale e la percezione della Gestalt, che erano – e continuano ad essere – resistenti al trattamento computazionale. Ancora più importante, Dreyfus ha sottolineato l'importanza della pertinenza, sottolineando la capacità degli esseri umani di distinguere l'essenziale dall'inessenziale e di attingere senza sforzo ad aspetti rilevanti della propria esperienza e conoscenza in conformità con le esigenze della loro situazione attuale, come richiesto dal loro continuo coinvolgimento con il mondo. Riteneva giustamente che conferire la stessa capacità a un computer digitale sarebbe stato un grave ostacolo per l'intelligenza artificiale, quello che chiamava il problema del "contesto olistico". Il problema della rilevanza rimane, a nostro avviso, la sfida tecnica chiave per l'IA, sia forte che debole, e anche per le scienze cognitive computazionali. L'affermazione che le persone non svolgono le loro attività quotidiane seguendo le regole sottolinea una preoccupazione che è stata una questione ricorrente per l'intelligenza artificiale forte e il marchio comunitario, e anche per le teorie mentalistiche generali come la linguistica generativa di Chomsky, e merita una breve discussione qui prima di discutere. passare all'esperimento mentale di Block. L'obiezione è stata avanzata sotto forme alquanto diverse da molti filosofi, da Wittgenstein e Quine a Dreyfus, Searle e

altri. Ha a che fare con la cosiddetta realtà psicologica delle spiegazioni della cognizione basate su regole, e in particolare con le simulazioni computerizzate dei processi mentali. La questione dipende dalla distinzione tra descrizione e causalità, e anche tra previsione e spiegazione. Un insieme di regole (o a fortiori un programma per computer) potrebbe descrivere adeguatamente un fenomeno cognitivo, in quanto le regole potrebbero costituire un modello veritiero delle grossolane regolarità osservative associate a quel fenomeno. Potrebbero adattarsi a tutti i dati sperimentali disponibili e fare tutte le previsioni giuste. Ma questo non significa che esista effettivamente una rappresentazione codificata delle regole (o del programma) dentro la nostra testa che è causalmente implicata nella produzione del fenomeno. La distinzione è nota anche nella terminologia di Pylyshyn (1991, p. 233) come differenza tra regole esplicite e regole implicite. Le regole implicite descrivono semplicemente regolarità comportamentali, mentre le regole esplicite hanno codificato rappresentazioni, presumibilmente nel nostro cervello, che svolgono un ruolo causale nella produzione delle regolarità. La questione della realtà psicologica solleva seri problemi epistemologici. Quali prove conterebbero a sostegno dell'affermazione che certe regole sono esplicitamente codificate nel nostro cervello? Come distinguiamo tra diversi insiemi di regole o diversi programmi informatici che tuttavia sono descrittivamente equivalenti? A quali parti di un modello computerizzato dovrebbe essere attribuito un significato psicologico e quali dovrebbero essere ignorate? Coloro che sono in sintonia con le argomentazioni di Quine sulla radicale indeterminatezza che affligge lo studio del linguaggio probabilmente nutrono dubbi simili sugli approcci computazionali alle scienze cognitive e concluderanno che le difficoltà di cui sopra sono insormontabili. (Anche se non è necessario accettare le argomentazioni di Quine sull'indeterminatezza o il suo comportamentismo per giungere a queste conclusioni). Chomsky vede tali paure come manifestazioni di pregiudizi empirici sulla mente umana e di un dualismo metodologico radicato ma ingiustificato che presuppone una netta distinzione tra il regno fisico e quello mentale. Per lui, i problemi epistemologici sopra menzionati non sono altro che il solito problema di sottodeterminazione induttiva che regolarmente si confronta con tutte le scienze. Scienziati cognitivi computazionali come Newell, Pylyshyn e altri hanno risposto in modo più concreto sviluppando la nozione di diversi livelli di descrizione del sistema; sostenere l'uso di sistemi di produzione al fine di evitare l'impegno a controllare il flusso e altri dettagli di implementazione che sarebbero inevitabilmente specificati nei linguaggi di programmazione convenzionali; e cercando di facilitare la scelta della teoria incoraggiando test rigorosi con insiemi di risultati sperimentali molto diversi, come dati cronometrici, dati sui movimenti oculari, protocolli verbali, ecc. Tuttavia, permangono seri problemi con la modellazione cognitiva computazionale, e molti continuano a ritenere che la Le difficoltà epistemologiche affrontate da tale modellizzazione non derivano dal consueto problema di sottodeterminazione che si riscontra comunemente nelle scienze fisiche, ma da un tipo di problema fondamentalmente diverso che è molto più impegnativo. Una seconda critica

influente è stata diretta specificamente contro il funzionalismo della macchina. È stato presentato da Block (1978) sotto forma di un esperimento mentale che ci chiede di immaginare l'intera popolazione cinese simulare una mente umana per un'ora. I cittadini cinesi sono tutti dotati di ricetrasmittenti che li collegano tra loro nel modo giusto. Possiamo pensare ai singoli cittadini cinesi come ai neuroni, o a qualsiasi struttura cerebrale che consideriamo atomica. Le persone sono inoltre collegate via radio a un corpo artificiale, dal quale possono ricevere stimoli sensoriali e al quale possono inviare segnali in uscita per generare comportamenti fisici come alzare un braccio. Secondo il funzionalismo della macchina, si dovrebbe concludere che se i cinesi simulassero fedelmente la giusta tabella di transizione, allora, in virtù della corretta relazione tra loro e con input e output, equivarrebbero di fatto a una mente cosciente. Ma questo ci sembra controintuitivo, se non palesemente assurdo. Il sistema risultante potrebbe anche essere isomorfo al cervello, a un certo livello di descrizione, ma non sembrerebbe nutrire alcuna sensazione, dolore, prurito o credenza e desiderio. Per ragioni simili ne conseguirebbe che non si potrebbe mai dire che nessun sistema di intelligenza artificiale puramente computazionale abbia una mente genuina. Alcuni funzionalisti hanno scelto di stringere i denti e ammettere che il "cervello cinese" (o un robot adeguatamente programmato) possiederebbe in realtà contenuti mentali genuini, attribuendo le nostre intuizioni contrarie allo sciovinismo cerebrale, la nostra propensione a considerare solo il wetware neurologico come capace di sostenere una vita mentale. Ma questo è difficile da digerire, e l'esperimento mentale ha convinto molti che il funzionalismo sfacciato è troppo liberale e deve essere abbandonato o circoscritto in modo significativo. Il terzo attacco filosofico fondamentale all'IA forte fu lanciato da Searle (1980) con il suo ormai famoso argomento della stanza cinese (CRA). Il CRA ha generato un'enorme quantità di discussioni e controversie e qui ne forniremo solo una rassegna molto superficiale; per una trattazione approfondita si rimanda a Cole (2004)[38]. CRA è basato su un esperimento mentale in cui è protagonista lo stesso Searle[39]. È all'interno di una stanza; fuori dalla stanza ci sono madrelingua cinesi che non sanno che Searle è al suo interno. Searle nella stanza, come Searle nella vita reale, non conosce il cinese, ma parla correntemente l'inglese. Gli oratori cinesi inviano le carte nella stanza attraverso una fessura; su queste carte sono scritte domande in cinese. La scatola, per gentile concessione del lavoro segreto di Searle al suo interno, restituisce le carte ai madrelingua cinesi come output. L'output di Searle viene prodotto consultando un regolamento: questo libro è una tabella di ricerca che gli dice quale cinese produrre in base a ciò che viene inviato. Per Searle, i cinesi sono solo un mucchio di - per usare il linguaggio di Searle - scarabocchi. Il nocciolo della questione è piuttosto semplice: Searle all'interno della stanza dovrebbe essere tutto ciò che un computer può essere, e

[38] Cole, D.: 2004, The chinese room argument, in E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy, Stanford University. http://plato.stanford.edu/entries/chinese-room

[39] Searle, J.: 1980, Minds, brains and programs, Behavioral and Brain Sciences 3, 417–424; Searle, J.: 1984, Minds, Brains, and Science, Harvard University Press, Cambridge, MA

poiché non capisce il cinese, nessun computer potrebbe avere una tale comprensione. Searle muove senza pensarci scarabocchi e, secondo la tesi, fondamentalmente è tutto ciò che fanno i computer. Searle ha fornito varie forme più generali dell'argomento. Ad esempio, riassume l'argomentazione a pagina 39 di (Searle 1984) come quella in cui da 1. La sintassi non è sufficiente per la semantica. 2. I programmi per computer sono interamente definiti dalla loro struttura formale, o sintattica. 3. Le menti hanno contenuti mentali; in particolare, hanno contenuti semantici.restituisce le carte ai madrelingua cinesi come output. L'output di Searle viene prodotto consultando un regolamento: questo libro è una tabella di ricerca che gli dice quale cinese produrre in base a ciò che viene inviato. Per Searle, i cinesi sono solo un mucchio di - per usare il linguaggio di Searle - scarabocchi. Il nocciolo della questione è piuttosto semplice: Searle all'interno della stanza dovrebbe essere tutto ciò che un computer può essere, e poiché non capisce il cinese, nessun computer potrebbe avere una tale comprensione. Searle muove senza pensarci scarabocchi e, secondo la tesi, fondamentalmente è tutto ciò che fanno i computer.

Molte risposte sono state date al CRA, sia nella sua incarnazione originaria, sia nella forma generale sopra espressa; forse i due più popolari sono la risposta del sistema e la risposta del robot. Il primo si basa sull'affermazione che sebbene Searle all'interno della stanza non capisca il cinese, il sistema complessivo che lo include come parte propria lo capisce. Ciò significa che viene messa in discussione la premessa secondo cui Searle-in-the-room è tutto ciò che un computer può essere. Quest'ultima obiezione si basa sull'affermazione che, sebbene, ancora una volta, Searle-in-the-room non capisca il cinese, questa carenza deriva dal fatto che Searle non è causalmente connesso all'ambiente esterno nel modo giusto. L'affermazione è che in un robot reale il significato verrebbe costruito sulla base delle transazioni causali del robot con il mondo reale. Quindi, sebbene Searle possa in un certo senso funzionare nella stanza come un computer, non funziona come un robot a tutti gli effetti, e l'intelligenza artificiale forte mira a costruire le persone come robot a tutti gli effetti. Searle ha fornito risposte alle risposte e la controversia continua. Indipendentemente dalle opinioni personali sulla CRA, l'argomento ha innegabilmente avuto un impatto enorme sul campo. Nello stesso momento in cui venivano avanzate critiche filosofiche come quelle sopra, cominciarono ad emergere seri problemi tecnici con l'IA classica. Uno di questi era il problema del telaio. Ormai il termine è diventato piuttosto vago. A volte viene inteso come il problema della rilevanza menzionato prima (come capire se un'informazione potrebbe essere rilevante in una data situazione); a volte è inteso nel senso dell'apparente intrattabilità computazionale dei processi di pensiero olistico; e occasionalmente viene addirittura frainteso come un'etichetta generica per l'impossibilità dell'IA simbolica. Forse la lettura più ampia e meno imprecisa è questa: si tratta del problema di esplicitare le condizioni alle quali una convinzione dovrebbe essere aggiornata dopo che un'azione è stata intrapresa. Nella sua incarnazione originale il problema era più tecnico e ristretto, e sorgeva nel contesto di un compito

molto specifico in un quadro molto specifico: ragionare sull'azione nel calcolo della situazione. Quest'ultimo è un sistema formale, basato sulla logica del primo ordine, per rappresentare e ragionare sull'azione, sul tempo e sul cambiamento. La sua nozione di base è quella di fluenza, ovvero una proprietà il cui valore può cambiare nel tempo, come la temperatura di una stanza o la posizione di un oggetto in movimento. I fluenti sono reificati e possono quindi essere quantificati. È importante sottolineare che le proprietà booleane del mondo sono esse stesse trattate come fluenti. Tale fluidità proposizionale potrebbe rappresentare se un oggetto è o meno alla sinistra di un altro oggetto, o se la luce in una stanza è accesa. Il mondo in un dato momento può essere descritto in modo esaustivo da un insieme di formule che indicano i valori di tutti i fluenti in quel momento; si dice che tale descrizione rappresenti lo stato del mondo in quel momento. Ogni azione ha una serie di precondizioni ed effetti, entrambi descritti in termini di fluidità. Se le precondizioni di un'azione sono soddisfatte in un dato stato, allora l'azione può essere eseguita e risulterà in un nuovo stato. Partendo da uno stato iniziale, che presumibilmente rappresenta il mondo in cui un robot vi entra per la prima volta, sono possibili molte diverse sequenze di stati a seconda delle diverse linee di azione che possono essere intraprese. Molti problemi dell'intelligenza artificiale, come la pianificazione, hanno una formulazione naturale in questo contesto. Ad esempio, elaborare un piano per raggiungere un determinato obiettivo equivale a scoprire una sequenza di azioni che trasformeranno lo stato attuale in uno stato che soddisfi l'obiettivo prefissato. A quanto pare, però, non è sufficiente descrivere gli effetti di ciascuna azione; bisogna anche specificare quali fluenti non sono influenzati da un'azione. Supponiamo, ad esempio, che sia disponibile un'azione liftBox, con la quale un robot può sollevare una scatola dal pavimento. La sua precondizione è che la scatola sia sul pavimento e il suo effetto è che la scatola sia sollevata dal pavimento. E supponiamo che ci sia un'altra azione, turnLightOn, con la quale il robot può accendere la luce nella stanza, presupposto che la luce sia spenta e l'effetto sia che la luce sia accesa. Se inizialmente la luce è spenta e la scatola è sul pavimento, sembrerebbe che l'esecuzione dell'azione turnLightOn seguita da liftBox dovrebbe comportare uno stato in cui la scatola è sollevata dal pavimento e la luce è accesa. Ma in realtà non segue nessuna delle due conclusioni, perché potrebbe essere, per tutto quello che abbiamo detto, che accendere la luce sollevi la scatola dal pavimento e che sollevando la scatola dal pavimento si spenga la luce. Nel primo caso il piano non funzionerebbe nemmeno perché la precondizione della seconda azione (liftBox) non sarebbe soddisfatta dopo la prima azione, mentre nel secondo caso la luce sarebbe spenta dopo la seconda azione. Per escludere modelli così stravaganti, dobbiamo specificare esplicitamente i non-effetti di ciascuna azione tramite i cosiddetti "assiomi del frame". Sebbene siano stati ideati modi concisi per enunciare gli assiomi dei frame, la complessità computazionale del ragionamento con essi rimane una sfida. Sono state avanzate diverse altre soluzioni proposte, che vanno dalla circoscrizione a formalismi completamente diversi per rappresentare e ragionare sull'azione e sul cambiamento. È

interessante notare che nessuna delle soluzioni proposte finora si avvicina minimamente all'efficienza con cui i bambini ragionano sull'azione. È stato suggerito che gli esseri umani non si imbattono nel problema di ragionare sui non-effetti delle azioni perché danno per scontato che un'azione non influisce su nulla a meno che non abbiano prova contraria.18 Tuttavia, il vero problema, che filosofi come Fodor hanno afferrato, è questa: come possiamo sapere se un'informazione costituisce o meno "prova del contrario"? Ci sono almeno due questioni separate qui. Per prima cosa dobbiamo essere in grado di determinare se un'informazione è potenzialmente rilevante o meno per alcune delle nostre convinzioni. Anche questo è il problema della pertinenza. In secondo luogo, dobbiamo essere in grado di determinare se l'informazione falsifica o meno la convinzione. Questi sono sia problemi di ingegneria per GOFAI che problemi filosofici generali. Sul fronte ingegneristico, non è troppo difficile costruire un sistema simbolico che raggiunga un verdetto ragionevole una volta identificate le giuste convinzioni di fondo. La principale difficoltà pratica è riuscire a individuare rapidamente le informazioni rilevanti. Molti sono arrivati a ritenere altamente improbabile che qualsiasi sistema di manipolazione dei simboli possa superare questa difficoltà.

1.4 Paradigmi dell'IA

I paradigmi dell'intelligenza artificiale (IA) rappresentano i diversi approcci concettuali e metodologici utilizzati per sviluppare sistemi intelligenti. Nel corso della storia dell'IA, si sono susseguiti diversi paradigmi, ognuno con i suoi punti di forza e di debolezza.

L'approccio simbolico, prevalente nei primi anni dell'IA, si concentra sulla rappresentazione e sul ragionamento simbolico. I sistemi basati su questo paradigma utilizzano simboli per rappresentare concetti, oggetti e relazioni nel mondo reale. Le regole logiche vengono poi impiegate per manipolare questi simboli e trarre inferenze. Esempi di sistemi simbolici includono:

- Sistemi esperti: progettati per emulare l'expertise umana in un dominio specifico, come la diagnosi medica o la configurazione di sistemi complessi.
- Ragionamento automatico: sviluppato per automatizzare compiti di inferenza logica, come la dimostrazione di teoremi matematici o la verifica di proprietà di sistemi software.

L'approccio subsimbolico, emerso negli anni '80 e diventato dominante negli ultimi decenni, si basa sull'apprendimento automatico e sulle reti neurali artificiali. I sistemi subsimbolici non richiedono una rappresentazione esplicita della conoscenza, ma apprendono direttamente dai dati attraverso processi statistici o connessionistici. Esempi di sistemi subsimbolici includono:

- Reti neurali artificiali: ispirate al funzionamento del cervello umano, sono in grado di apprendere modelli complessi da grandi quantità di dati, come immagini, testi o sequenze temporali.
- Apprendimento automatico: abbraccia una vasta gamma di algoritmi e tecniche per estrarre conoscenza dai dati, come la classificazione, la regressione, il clustering e il rinforzo positivo.

1.5 Machine learning, deep learning e reti neurali

Il machine learning è un sottocampo dell'IA che consente ai sistemi di apprendere e migliorare automaticamente senza essere esplicitamente programmati. In altre parole, i sistemi di machine learning possono "imparare dai dati", identificando modelli e relazioni all'interno di grandi quantità di informazioni. Il machine learning si suddivide in diverse categorie di modelli, ciascuna basata su specifiche tecniche algoritmiche. In base alla natura dei dati e agli obbiettivi da raggiungere, è possibile scegliere tra quattro principali modelli di apprendimento: supervisionato, non supervisionato, semi-supervisionato e per rinforzo. Ogni modello può implementare una o più tecniche algoritmiche, a seconda dei set di dati utilizzati e dei risultati attesi; gli algoritmi inoltre possono essere applicati singolarmente o in combinazione per ottenere la massima precisione possibile. Il primo dei quattro modelli è l'apprendimento supervisionato, nel quale la macchina viene addestrata con l'esempio. L'apprendimento supervisionato sfrutta dati etichettati per istruire algoritmi capaci di classificare dati o effettuare previsioni con precisione; i modelli vengono allenati su un dataset in cui ogni input è associato al suo output corretto. Questo processo richiede solitamente l'intervento umano per fornire all'algoritmo i dati etichettati. Tramite un algoritmo, il sistema compila questi dati di addestramento e inizia a stabilire somiglianze correlative, differenze e altri aspetti logici, fino a quando non è in grado di prevedere autonomamente le risposte alle domande.
Il secondo modello è l'apprendimento non supervisionato, in cui non vi sono dati etichettati o strutturati, bensì, la macchina analizza i dati di input e inizia a individuare schemi e relazioni utilizzando tutte le informazioni disponibili e rilevanti. In un certo senso, l'apprendimento non supervisionato assomiglia al modo in cui gli esseri umani osservano il mondo: per classificare gli oggetti in categorie, facciamo affidamento sull'intuizione e sull'esperienza. Per una macchina, l'esperienza corrisponde alla quantità di dati forniti e di cui può disporre.
L'apprendimento semi supervisionato unisce elementi dell'apprendimento supervisionato e non supervisionato, combinando una piccola parte di dati etichettati e una grande quantità di dati non etichettati per addestrare un modello. I dati etichettati forniscono un impulso iniziale al sistema, permettendo di incrementare significativamente la velocità e la precisione dell'apprendimento; in

questo modello quindi, l'algoritmo istruisce la macchina ad esaminare i dati etichettati per individuare le proprietà correlate, che poi possono essere applicate ai dati non etichettati.

Infine, l’apprendimento per rinforzo, addestra il software a prendere decisioni ottimali attraverso un processo di tentativi ed errori. Le operazioni che avvicinano l'obiettivo vengono premiate, mentre quelle che se ne allontanano vengono ignorate. Gli algoritmi adottano un sistema di ricompense e penalità durante l'elaborazione dei dati, imparando dal feedback di ogni operazione e individuando autonomamente i percorsi migliori per raggiungere i risultati desiderati; è una metodologia potente che permette ai sistemi di intelligenza artificiale di conseguire risultati ottimali in ambienti sconosciuti.

Il deep learning è un tipo specifico di machine learning che utilizza reti neurali artificiali per apprendere da dati complessi, come immagini, suoni o testi. Le reti neurali sono ispirate al funzionamento del cervello umano e sono composte da strati di neuroni artificiali interconnessi. Ogni neurone riceve input da altri neuroni, li elabora e produce un output che viene inviato ad altri neuroni. Gli algoritmi di deep learning sono emersi nel tentativo di rendere più efficienti le tecniche di machine learning tradizionali. I metodi di machine learning convenzionali trovano difficile elaborare dati non strutturati, come i documenti di testo, a causa delle infinite varianti presenti nei set di dati di addestramento. Al contrario, i modelli di deep learning sono in grado di interpretare dati non strutturati e di fare osservazioni generali senza la necessità di estrarre manualmente le caratteristiche.

CAPITOLO II: PRESENTE

2.1 L'Intelligenza Artificiale: inquadramento generale

L'intelligenza artificiale è un settore in rapida crescita che ha il potenziale per trasformare l'assistenza sanitaria. L'intelligenza artificiale comprende un'ampia gamma di tecnologie che consentono ai computer di eseguire attività che tipicamente richiedono l'intelligenza umana, come l'apprendimento, il ragionamento e la risoluzione dei problemi. L'uso dell'intelligenza artificiale nel settore sanitario si è già dimostrato promettente nel migliorare i risultati dei pazienti, ridurre i costi e aumentare l'efficienza[40].

I rapidi progressi nel campo dell'intelligenza artificiale (AI) hanno creato interessanti opportunità per il settore sanitario. Le tecnologie di intelligenza artificiale, come l'apprendimento automatico, l'elaborazione del linguaggio naturale e la visione artificiale, hanno rivoluzionato vari aspetti dell'erogazione dell'assistenza sanitaria[41]. Tali progressi possono apportare miglioramenti significativi in termini di potenziamento e semplificazione dei processi amministrativi, nonché sul piano della promozione della ricerca e dell'innovazione medica[42].

Una delle applicazioni più importanti dell'intelligenza artificiale nel settore sanitario è la diagnostica e l'*imaging* medico. Gli algoritmi di intelligenza artificiale possono analizzare immagini mediche, come raggi X, scansioni TC e risonanza magnetica, per rilevare anomalie, tumori e altre condizioni con un livello di precisione particolarmente elevato: tutto ciò ha il potenziale per migliorare l'individuazione e la diagnosi precoce, nonché i risultati del trattamento in generale[43]. Anche gli assistenti virtuali hanno trovato il loro posto negli ambienti sanitari, fornendo ai pazienti supporto e informazioni personalizzati; questi sistemi intelligenti possono rispondere a domande a contenuto medico, fornire indicazioni sulla cura di sé e classificare i pazienti in base ai loro sintomi[44]. Pertanto, non solo vi è un miglioramento dell'accessibilità all'assistenza sanitaria, ma anche una riduzione degli oneri degli operatori sanitari[45].

L'intelligenza artificiale si è dimostrata promettente nell'analisi predittiva e nel monitoraggio dei pazienti. Analizzando grandi quantità di dati dei pazienti, gli algoritmi di intelligenza artificiale possono identificare modelli e fattori di rischio per le malattie, consentendo agli operatori sanitari di

---

[40] Mannelli C. Etica e Intelligenza artificiale: Il caso sanitario. Italia, Donzelli Editore, 2022; Corio M., Paone S., Ferrone E, Meier H., Jefferson T.O. e Cerbo. Revisione sistematica degli strumenti metodologici impiegati nell'Health Technology Assessment, Agenas, Roma, 2011; Drummond M.F. et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. International Journal of Technology Assessment in Health Care. (2008).

[41] Campanale C., Cinquini L., Corsi S. e Piccaluga A. "Innovazione nella tecnologia biomedicale: un modello di valutazione dei costi del sistema Echo-Laser in chirurgia mini-invasiva", Mecosan, 2011.

[42] *Ibidem*.

[43] D'Aloia A., Intelligenza artificiale e diritto: Come regolare un mondo nuovo. Italia, Franco Angeli Edizioni, 2021.

[44] Guarda P., Petrucci L., Quando l'intelligenza artificiale parla: assistenti vocali e sanità digitale alla luce del nuovo regolamento generale in materia di protezione dei dati, in Bio Law Journal -Rivista di Bio Diritto, n. 2, 2020, pp. 425 ss.

[45] Corso S., Il fascicolo sanitario elettronico fra e-Health, privacy ed emergenza sanitaria, in Responsabilità medica, n. 4, 2020, pp. 393 ss

intervenire preventivamente. I dispositivi indossabili basati sull'intelligenza artificiale e i sistemi di monitoraggio remoto consentono il monitoraggio continuo dei segni vitali, fornendo avvisi in tempo reale per cambiamenti critici nello stato di salute del paziente[46].

L'intelligenza artificiale può anche svolgere un ruolo significativo nella scoperta e nello sviluppo di farmaci. Analizzando grandi quantità di dati biomedici e letteratura scientifica, gli algoritmi di intelligenza artificiale possono identificare potenziali bersagli farmacologici, ottimizzare la progettazione dei farmaci e accelerare il processo di sperimentazione clinica offrendo nuovi trattamenti ai pazienti in modo più rapido ed efficiente. Durante la fase di progettazione di un nuovo farmaco, l'IA può aiutare a determinare la struttura della proteina target e a prevedere come il farmaco interagirà con il suo recettore o proteina. Nella fase di screening, l'IA permette di anticipare le proprietà fisico-chimiche della nuova molecola (come solubilità, ionizzazione e partizione), oltre alla sua attività e tossicità, superando le limitazioni dei tradizionali test in vitro e sugli animali. Durante lo sviluppo, l'IA può semplificare la risoluzione di problemi legati al prodotto finale, consentendo un notevole risparmio di tempo e risorse. Nella fase di verifica, gli algoritmi di machine learning possono contribuire a progettare trial clinici più efficaci, selezionando una popolazione di pazienti adeguata. Infine, l'IA può supportare un posizionamento efficace del prodotto sul mercato. L'importanza e l'interesse per l'applicazione dell'IA in questo settore sono evidenti sia dall'ampio numero di strumenti software e algoritmi sviluppati, sia dai significativi investimenti delle grandi aziende farmaceutiche nel campo dell'IA, insieme alle numerose collaborazioni tra le principali industrie e le società specializzate nello sviluppo di strumenti di intelligenza artificiale.

Sebbene l'intelligenza artificiale abbia già apportato contributi sostanziali all'assistenza sanitaria, il suo potenziale per il futuro è ancora più promettente[47]. I progressi negli algoritmi di intelligenza artificiale, insieme alla crescente disponibilità di dati sanitari, possono migliorare ulteriormente l'accuratezza e l'efficienza dei processi diagnostici, consentire la medicina personalizzata e migliorare raccomandazioni sul trattamento. L'intelligenza artificiale può anche contribuire alla gestione della salute della popolazione analizzando le cartelle cliniche e i determinanti sociali della salute per identificare tendenze, prevedere epidemie e allocare le risorse in modo efficace[48].

Un aspetto particolarmente interessante nell'area dell'intelligenza artificiale in ambito sanitario è la robotica: i robot, infatti, possono essere utilizzati per un'ampia gamma di compiti nel settore sanitario, tra cui chirurgia, riabilitazione e, in generale, nella cura del paziente[49]. I robot chirurgici, ad esempio, possono essere utilizzati per eseguire procedure minimamente invasive, che possono ridurre i tempi di recupero e migliorare i risultati. I robot possono essere utilizzati, altresì, nel contesto della

[46] Nappo F. Aziende e intelligenza artificiale: Prime riflessioni critiche. Italia, Franco Angeli Edizioni, 2021.
[47] Mannelli C., *op. cit*., pp. 60-65.
[48] *Ibidem*.
[49] Abate D., Robot e intelligenza artificiale: Rischi e opportunità, 2019.

telemedicina consentendo, dunque, agli operatori sanitari di monitorare a distanza i pazienti e fornire assistenza in tempo reale[50].

La robotica è una novità rivoluzionaria in ambito sanitario che trasformerebbe radicalmente il sistema di assistenza sanitaria: la robotica, infatti, combina gli algoritmi di intelligenza artificiale con dispositivi meccanici per creare macchine intelligenti in grado di eseguire compiti fisici e di interagire con l'ambiente. I robot, dunque, possono intervenire in vari aspetti della cura del paziente, delle procedure mediche e delle operazioni sanitarie: uno degli aspetti più della robotica in ambito sanitario è nelle procedure chirurgiche; si pensi, ad esempio al sistema chirurgico da Vinci che consente ai chirurghi di eseguire interventi chirurgici con maggiore precisione e controllo. Questi sistemi sono costituiti da bracci robotici con strumenti specializzati controllati dal chirurgo, che offrono maggiore destrezza, visualizzazione in 3D e un ridotto livello di invasività.

Con la chirurgia robotica si ridurrebbe la tempistica delle degenze e di recupero - garantendo, in tal modo, la disponibilità dei posti letto per i ricoveri - e si conseguirebbero migliori risultati chirurgici per i pazienti. Oltre alle applicazioni chirurgiche, i robot possono anche assistere gli operatori sanitari in compiti quali la cura e la riabilitazione dei pazienti; i robot possono essere, altresì, utilizzati per eseguire attività ripetitive, come il sollevamento e il trasferimento di pazienti, riducendo lo sforzo fisico degli operatori sanitari e minimizzando il rischio di infortuni sul luogo di lavoro.

Gli esoscheletri robotici possono essere utili nella riabilitazione fornendo supporto e assistenza ai pazienti che si stanno riprendendo da lesioni o menomazioni, aiutandoli a ritrovare mobilità e forza. I robot sono, inoltre, dotati di sensori e algoritmi di intelligenza artificiale e, per questo motivo, sono in grado di effettuare spostamenti tra i corridoi dell'ospedale e nelle stanze dei pazienti, catturando i segni vitali, trasmettendo informazioni agli operatori sanitari e consentendo consultazioni da remoto e in tempo reale: gli operatori sanitari, dunque, possono monitorare i pazienti in tempo reale e fornire interventi tempestivi, soprattutto in aree remote o scarsamente servite[51]. Inoltre, la robotica può migliorare l'efficienza delle operazioni sanitarie e della logistica.

I robot autonomi possono essere implementati per attività quali la consegna di farmaci, la gestione dell'inventario e la sterilizzazione degli ambienti ospedalieri: tale automazione riduce il carico di lavoro del personale, migliora la precisione e consente agli operatori sanitari di concentrarsi maggiormente sulla cura del paziente[52].

Sebbene la robotica nel settore sanitario offra numerosi vantaggi, ci sono alcune sfide da affrontare. Garantire la sicurezza e l'affidabilità dei sistemi robotici è fondamentale, in particolare nelle procedure più critiche e complesse: i sistemi robotici devono essere sottoposti a test e validazioni

[50] Parisi G., La trasformazione digitale del trattamento sanitario. Italia, Ledizioni, 2022.
[51] Teigens V., Skalfist P., Mikelsten D., Intelligenza artificiale: la quarta rivoluzione industriale. N.p., Cambridge Stanford Books, pp. 64-67.
[52] Parisi G., *op. cit*, pp. 42-45.

rigorosi per garantire il soddisfacimento dei più elevati standard di sicurezza ed efficacia; inoltre, è necessario considerare e affrontare attentamente considerazioni etiche quali il consenso del paziente, la privacy e il mantenimento del tocco umano nelle interazioni sanitarie[53].

Le sfide a cui far fronte riguardano i costi, la sicurezza e l'affidabilità, la formazione e la competenza degli operatori sanitari, integrazione e interoperabilità, accettazione, questioni legali, etiche e normative.

Con riferimento ai costi, i sistemi robotici possono hanno un costo elevato da acquisire, mantenere e aggiornare: l'elevato costo della tecnologia robotica può rappresentare una sfida finanziaria per le istituzioni sanitarie, limitandone l'accessibilità e l'adozione. Per quanto riguarda la sicurezza e l'affidabilità, è importante garantire la sicurezza e l'affidabilità dei sistemi robotici, in particolare nelle procedure sanitarie critiche: i sistemi robotici devono, dunque, essere sottoposti a test approfonditi, validazione e conformità normativa per ridurre al minimo il rischio di errori, malfunzionamenti o eventi avversi[54]. Per quanto riguarda la formazione e la competenza degli operatori sanitari, questi ultimi necessitano di una formazione specializzata per utilizzare e gestire i sistemi robotici in modo efficace. Per quanto riguarda l'integrazione e l'interoperabilità dei sistemi robotici con le infrastrutture sanitarie e i sistemi informativi esistenti può essere una questione complessa da affrontare: è necessario, dunque, abbattere quelle barriere che ostacolano l'interoperabilità e la compatibilità con le cartelle cliniche elettroniche (EHR) e altre tecnologie sanitarie sono essenziali per un flusso di lavoro e uno scambio di dati efficienti. Per quanto riguarda le questioni di natura etica e legale, possono sorgere dilemmi etici nell'uso della robotica in ambito sanitario, come il consenso del paziente, la privacy e la potenziale riduzione dell'interazione umana e dell'empatia. Garantire l'esistenza di linee guida etiche e quadri giuridici è fondamentale per affrontare queste preoccupazioni e mantenere la fiducia dei pazienti. Per quanto riguarda, invece, l'accettazione della robotica, l'adozione di quest'ultima nel settore sanitario può incontrare resistenza da parte degli operatori sanitari che potrebbero temere lo spostamento del lavoro o percepire la robotica come una minaccia al tocco umano nella cura dei pazienti, nonché una delle cause della perdita del lavoro; al riguardo, dunque, si ritiene di fondamentale importanza promuovere una mentalità collaborativa. Inoltre, gli organismi di regolamentazione potrebbero dover adattare e stabilire linee guida specifiche per la robotica nel settore sanitario e potrebbe essere necessario modificare i modelli di rimborso per accogliere l'uso della robotica, garantendo un giusto compenso per i servizi forniti e incoraggiandone un'adozione diffusa[55].

---

[53] D'Aloia, A. (2021). Intelligenza artificiale e diritto: Come regolare un mondo nuovo. Italia: Franco Angeli Edizioni.

[54] *Ibidem*.; Oliver A, Mossialos E, Robinson R: Health technology assessment and its influence on health care priority setting. Int J Technol Assess Health Care. (2004); Rocco B, Matei DV, Melegari S, Ospina JC, Mazzoleni F, Errico G (2009). Robotic vs open prostatectomy in a laparoscopically naive centre: a matchedpair analysis. BJU Int. (2009).

[55] Tewari A., Sooriakumaran P., Positive surgical margin and perioperative complication rates of primary surgical treatments for prostate cancer: a systematic review and meta-analysis comparing retropubic, laparoscopic, and robotic

Nonostante i miglioramenti che potrebbe apportare la robotica nel settore sanitario, potrebbero mancare prove cliniche solide e dati sui risultati a lungo termine a sostegno della sua efficacia e del rapporto costo-efficacia: è necessario, dunque, effettuare ulteriori ricerche e studi per creare una solida base di prove sui benefici e sull'impatto della robotica in diverse applicazioni sanitarie. In riferimento alla qualità dei dati e alla privacy, l'intelligenza artificiale si basa su dati di alta qualità per generare risultati accurati, ma i dati sanitari possono essere incompleti, incoerenti o distorti; perplessità sorgono, altresì, in riferimento alla tutela della privacy dei pazienti durante la condivisione dei dati medici per l'analisi dell'intelligenza artificiale. Per quanto riguarda, invece, il lato normativo ed etico, le preoccupazioni principali riguardano la trasparenza e la responsabilità degli algoritmi di intelligenza artificiale, nonché il potenziale di pregiudizi e discriminazioni. Gli organismi di regolamentazione devono sviluppare standard e linee guida per l'uso dell'IA nel settore sanitario per garantire che venga utilizzata in modo etico e sicuro. Le soluzioni di intelligenza artificiale devono, inoltre, essere integrate nei sistemi e nei flussi di lavoro sanitari esistenti, e questo potrebbe rivelarsi impegnativo e richiedere tempistiche lunghe[56].

Le possibilità future dell'intelligenza artificiale nel settore sanitario sono vaste ed entusiasmanti. Una delle aree più promettenti dell'intelligenza artificiale nel settore sanitario è la medicina personalizzata. La medicina personalizzata prevede l'adattamento del trattamento medico alla composizione genetica, allo stile di vita e all'ambiente di un individuo. L'intelligenza artificiale può essere utilizzata per analizzare grandi quantità di dati per sviluppare piani di trattamento personalizzati su misura per le esigenze specifiche di ciascun paziente. Un'altra area promettente dell'intelligenza artificiale nel settore sanitario è la previsione e la prevenzione delle malattie. L'intelligenza artificiale può essere utilizzata per analizzare dati provenienti da più fonti, come test genetici, cartelle cliniche e dati ambientali, per identificare individui ad alto rischio di sviluppare determinate malattie. Queste informazioni possono essere utilizzate per sviluppare strategie preventive in grado di ridurre l'incidenza della malattia. Infine, l'intelligenza artificiale ha il potenziale per rivoluzionare la scoperta dei farmaci.

L'intelligenza artificiale può essere utilizzata per analizzare grandi quantità di dati per identificare nuovi bersagli farmacologici e sviluppare trattamenti più efficaci. L'intelligenza artificiale può essere utilizzata anche per ottimizzare gli studi clinici, riducendo i tempi e i costi legati all'immissione di nuovi farmaci sul mercato. Vi sono preoccupazioni circa la possibilità che l'intelligenza artificiale perpetui pregiudizi e discriminazioni nel settore sanitario. Ad esempio, se gli algoritmi di intelligenza artificiale vengono addestrati su dati distorti, potrebbero prendere decisioni distorte che hanno un

prostatectomy, 2012; Wright JD, Ananth CV, Lewin S. N. et al., Robotically assisted vs laparoscopic hysterectomy among women with benign gynecologic disease. JAMA, 2013.

[56] Carlsson S., Nilsson A.E., Schumacher M.C. et al., Surgery-related complications in 1253 robot-assisted and 485 open retropubic radical prostatectomies at the Karolinska University Hospital, Sweden. Urology. 2010.

impatto negativo su determinate popolazioni. Sorgono problemi normativi anche riguardo all'uso dell'intelligenza artificiale nel settore sanitario. Lo sviluppo e l'uso degli algoritmi di intelligenza artificiale devono essere soggetti a supervisione normativa per garantirne la sicurezza e l'efficacia. Le possibilità future dell'intelligenza artificiale nel settore sanitario sono vaste e hanno il potenziale per rivoluzionare il modo in cui viene fornita l'assistenza sanitaria. [57]Si prevede che i progressi nelle tecniche e nelle tecnologie dell'intelligenza artificiale apporteranno numerosi vantaggi e progressi nel settore sanitario[58].

Le prospettive future dell'intelligenza artificiale in ambito sanitario sono più che positive. In primo luogo, l'intelligenza artificiale gioca un ruolo fondamentale nel progresso della medicina di precisione, che mira a personalizzare trattamenti e interventi medici sui singoli pazienti. Gli algoritmi di intelligenza artificiale possono analizzare dati genomici su larga scala, cartelle cliniche dei pazienti e altre informazioni rilevanti per identificare modelli, prevedere i rischi di malattie e sviluppare piani di trattamento personalizzati, dando luogo a terapie più mirate ed efficaci e riducendo al minimo gli effetti avversi e ottimizzando i risultati per i pazienti. L'intelligenza artificiale può accelerare il processo di scoperta e sviluppo di farmaci analizzando grandi quantità di dati biomedici, tra cui informazioni genomiche, strutture proteiche e letteratura scientifica. Gli algoritmi di intelligenza artificiale possono identificare potenziali farmaci candidati, prevederne l'efficacia e ottimizzarne le proprietà riducendo, altresì, la tempistica e i costi associati all'immissione di nuovi farmaci sul mercato[59].

Le tecniche di analisi delle immagini basate sull'intelligenza artificiale possono migliorare significativamente l'*imaging* e la diagnostica medica: infatti, gli algoritmi di apprendimento automatico possono analizzare immagini mediche, come raggi X, scansioni TC e risonanza magnetica, per rilevare anomalie, fornire una diagnosi precoce delle malattie e valutazioni quantitative. L'intelligenza artificiale può aiutare i radiologi e altri operatori sanitari a effettuare diagnosi più accurate e tempestive, portando a risultati migliori per i pazienti.

Gli assistenti virtuali e le chatbots basati sull'intelligenza artificiale hanno il potenziale per migliorare il coinvolgimento dei pazienti e fornire supporto 24 ore su 24, 7 giorni su 7: tali sistemi intelligenti possono interagire con i pazienti, rispondere alle loro domande, fornire indicazioni e offrire consigli sanitari personalizzati. Gli assistenti virtuali possono assistere nella pianificazione degli appuntamenti, nei promemoria dei farmaci e nelle istruzioni post-terapia, migliorando l'esperienza del paziente e l'aderenza ai piani di trattamento.

---

[57] Vicini C., Montevecchi F., Transoral Robotic Surgery. Presentato a 83° Annual Meeting of the German Society ORL, Mainz, paper 33, 2012;

[58] Ibidem.

[59] Kaplan, J. (2024). Le persone non servono. Lavoro e ricchezza nell'era dell'intelligenza artificiale. Nuova ediz.. : Luiss University Press.

Monitorando continuamente i segni vitali dei pazienti, le cartelle cliniche elettroniche e altri dati rilevanti, i sistemi di intelligenza artificiale possono avvisare gli operatori sanitari di potenziali complicazioni, consentendo interventi tempestivi e misure preventive.
La robotica combinata con l'intelligenza artificiale può far avanzare ulteriormente le procedure chirurgiche, la cura dei pazienti e le operazioni sanitarie. I sistemi robotici intelligenti possono assistere i chirurghi nell'esecuzione di interventi chirurgici complessi con maggiore precisione e controllo e possono anche automatizzare attività di routine, come la distribuzione dei farmaci, l'elaborazione dei campioni e la logistica, liberando tempo dagli operatori sanitari e migliorando l'efficienza complessiva.
Gli algoritmi di intelligenza artificiale possono analizzare i dati per rilevare deviazioni dagli intervalli normali e avvisare gli operatori sanitari in tempo reale[60]. Ciò consente il monitoraggio remoto di pazienti con patologie croniche, cure post-chirurgiche e popolazioni anziane, consentendo un intervento precoce e riducendo la necessità di visite ospedaliere.
Sul piano normativo, entrano in gioco le questioni relative alla privacy e alla sicurezza dei dati. In termini di responsabilità e trasparenza, è necessario comprendere chiaramente come gli algoritmi di intelligenza artificiale prendono decisioni: ciò implica la comprensione degli input, degli output e del processo decisionale dell'algoritmo. Inoltre, è necessario chiarire i vari tipi di responsabilità per le decisioni prese dagli algoritmi di intelligenza artificiale e questo è possibile attraverso l'uso dell'intelligenza artificiale spiegabile (XAI), progettata per fornire una chiara spiegazione di come un algoritmo è arrivato a una decisione.

2.2 Il caso AlphaFold

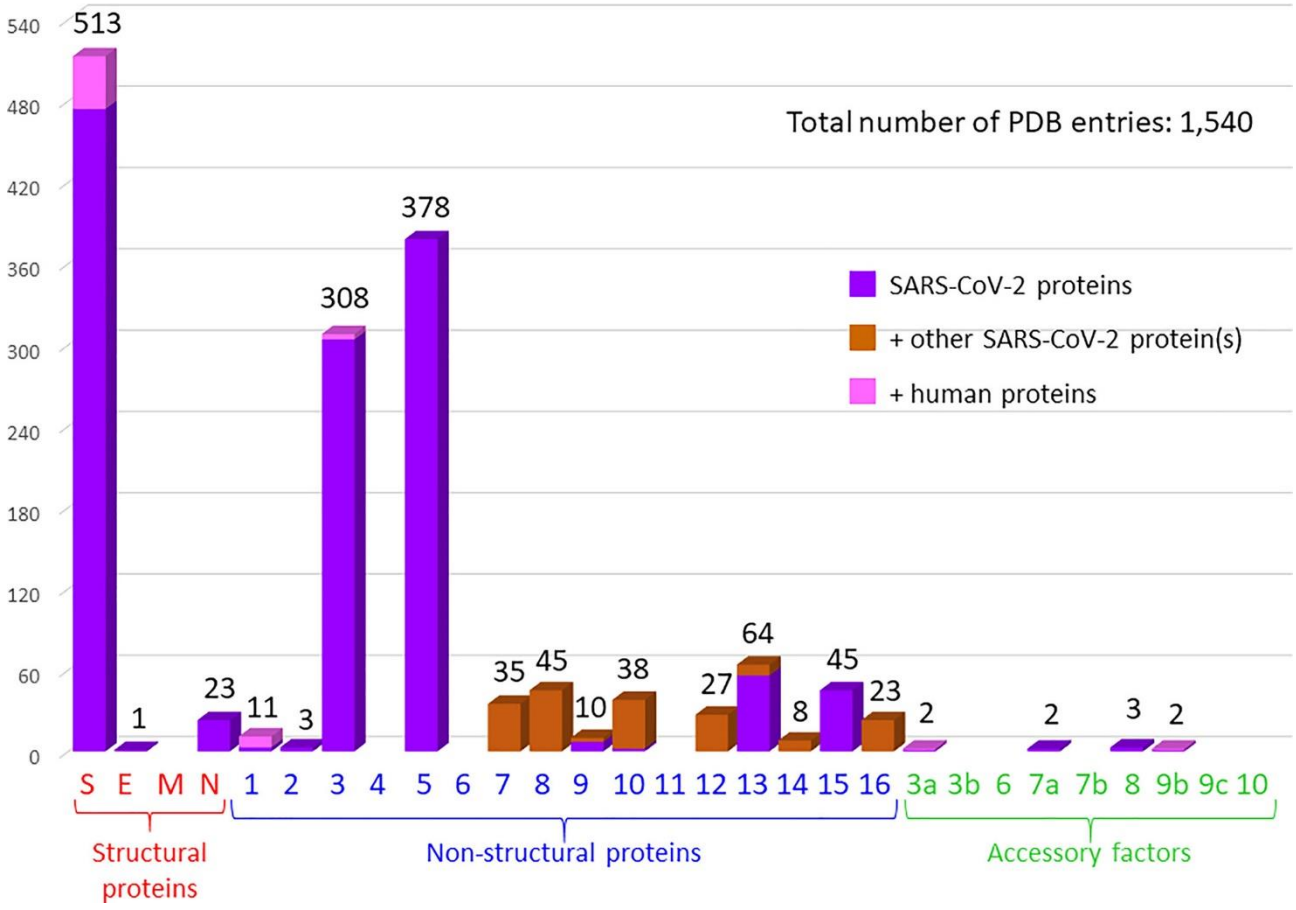
AlphaFold è un sistema di intelligenza artificiale sviluppato da DeepMind (una divisione di Google) che è in grado di prevedere la struttura tridimensionale di una proteina a partire dalla sua sequenza di amminoacidi. In parole più semplici, AlphaFold riesce a "indovinare" la forma che una proteina assumerà in natura, un compito che fino a poco tempo fa richiedeva anni di studi e sperimentazioni in laboratorio, ciò ha fornito un grande contributo alla ricerca sul Covid-19 nella recente pandemia.
La struttura di una proteina ne determina la funzione. Conoscere la forma di una proteina significa comprendere come essa interagisce con altre molecole e come svolge i suoi compiti all'interno di una cellula e, di conseguenza, per modificarla. Tra i risultati straordinari raggiunti nel corso degli anni grazie al proteing folding, possiamo citare:

[60] Kaplan, J. (2024). Le persone non servono. Lavoro e ricchezza nell'era dell'intelligenza artificiale. Nuova ediz.. (n.p.): Luiss University Press.

- La scoperta della struttura dell'emoglobina, la proteina presente nelle cellule del sangue che svolge un ruolo cruciale nel trasporto dell'ossigeno.
- L'analisi della struttura delle proteine del virus SARS-CoV-2, essenziali per la ricerca finalizzata allo sviluppo di farmaci e vaccini.
- La definizione della struttura delle proteine coinvolte nei processi di fotosintesi delle piante.

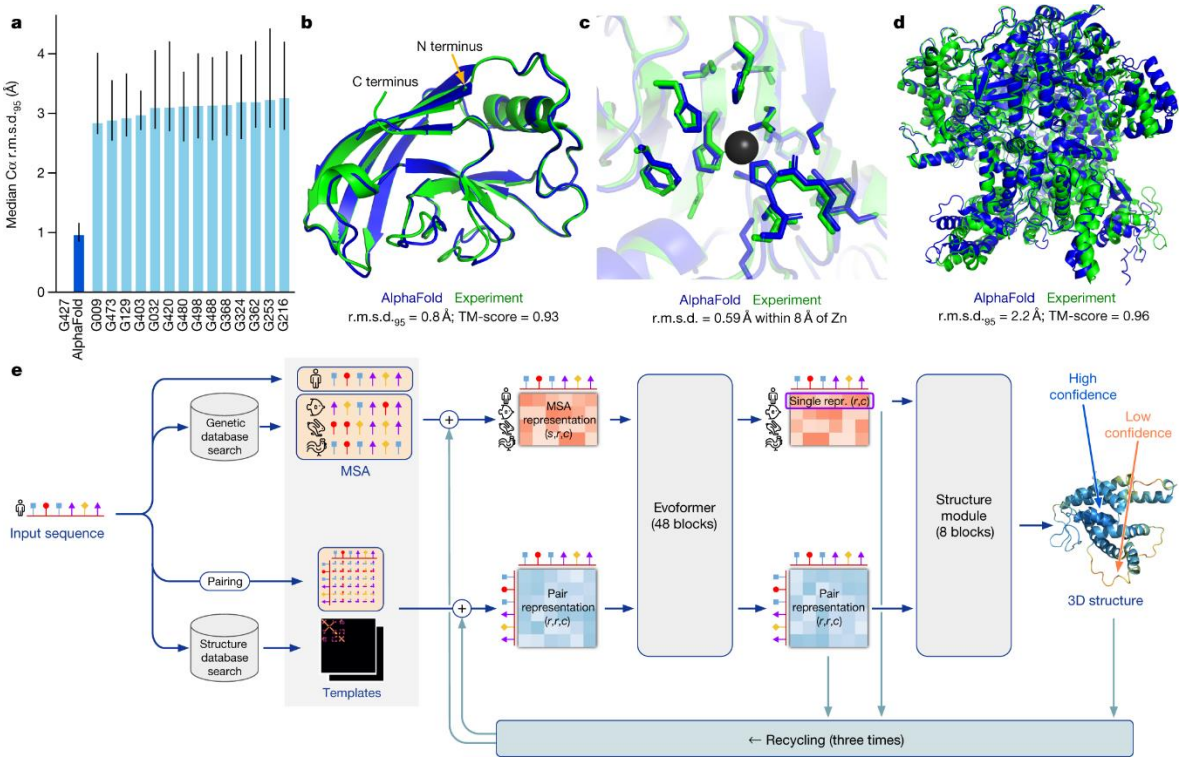
Dallo scoppio del COVID-19 (malattia da coronavirus 2019) in Cina nel dicembre 2019 e dalla sua dichiarazione come pandemia globale l'11 marzo 2020, scienziati di tutto il mondo hanno analizzato le strutture delle proteine che compongono il virus. Queste includono quattro proteine strutturali (spike, membrana, envelope e nucleocapside), 16 proteine non strutturali che formano il complesso replicasi/trascrittasi e nove presunti fattori accessori. Si è trattato di un'impresa notevole, con 1.540 strutture depositate nel PDB entro il 1° ottobre 2021.
Il server web PDBsum è stato sviluppato presso l'University College London (UCL) nel 1995[61] e trasferito all'EMBL-EBI nel 2001, dove risiede ora. Fornisce un compendio ampiamente illustrato delle proteine e dei loro complessi nel Protein Data Bank (PDB), con analisi della struttura secondaria delle proteine, diagrammi schematici per le interazioni proteina-ligando, proteina-DNA e proteina-proteina, analisi procheck della qualità strutturale e molto altro.
La figura[62] sottostante illustra la distribuzione di queste strutture all'interno delle proteine. Non è sorprendente che gran parte delle ricerche si sia focalizzata sulla proteina spike, che il virus impiega per penetrare e infettare le cellule dell'organismo ospite.

[61] Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: un database basato sul Web di riepiloghi e analisi di tutte le strutture PDB. *Tendenze Biochimica Sci*, 1997
[62] *Protein Science*, Wiley Periodicals LLC, Extra PDBsum: modelli SARS-CoV-2 e AlphaFold, 2021,



Alla fine del 2020, DeepMind, ha annunciato che il suo sistema AlphaFold 2 aveva superato notevolmente tutti gli altri metodi nella valutazione biennale della previsione della struttura proteica (CASP)[63], dimostrando un'affidabilità superiore al 95% tra la simulazione sperimentale basata su AI ed i risultati ottenuti nei laboratori tradizionali. I modelli generati erano di qualità comparabile a quella ottenuta tramite metodi sperimentali. A metà del 2021, l'azienda ha reso disponibile il codice sorgente e ha pubblicato quasi 350.000 modelli proteici di diverse specie, inclusa quella umana[64], il che ha suscitato un grande entusiasmo nella comunità della biologia strutturale. È stato creato un sito web, frutto della collaborazione tra DeepMind ed EMBL-EBI, che offre accesso a singoli modelli e consente il download di tutti i modelli per specie. Uno dei principali vantaggi di avere a disposizione questi modelli è che le strutture 3D ipotetiche di molte proteine, per le quali non esistono informazioni strutturali nemmeno da omologhi distanti, possono ora servire come base per studi funzionali. I modelli AlphaFold disponibili per tutte le proteine umane sono stati aggiunti a PDBsum: 23.391 modelli in tutto, corrispondenti a 20.504 proteine; un beneficio significativo nell'avere i modelli in PDBsum è la facilità con cui è possibile confrontarli con altre strutture già esistenti.

[63] Protein Structure Prediction Center, Casp14, Group performance based on combined z-scores, 2020
[64] Tunyasuvunakool K, Adler J, Wu Z, et al. Previsione altamente accurata della struttura proteica per il proteoma umano. Nature, 2021



a, Le prestazioni di AlphaFold sul dataset CASP14 (*n* = 87 domini proteici) rispetto alle prime 15 voci (su 146 voci), i numeri di gruppo corrispondono ai numeri assegnati ai partecipanti da CASP. I dati sono la mediana e l'intervallo di confidenza del 95% della mediana, stimato da 10.000 campioni bootstrap. **b**, La nostra previsione del target CASP14 T1049 (PDB 6Y4F, blu) rispetto alla vera struttura (sperimentale) (verde). Quattro residui nel C-terminale della struttura cristallina sono valori anomali del fattore *B* e non sono rappresentati. **c**, Target CASP14 T1056 (PDB 6YJ1). Un esempio di un sito di legame dello zinco ben previsto (AlphaFold ha catene laterali accurate anche se non prevede esplicitamente lo ione zinco). **d**, il target CASP T1044 (PDB 6VR4) una singola catena da 2.180 residui—è stato previsto con un corretto impaccamento del dominio (la previsione è stata fatta dopo CASP usando AlphaFold senza intervento). **e**, Architettura del modello. Le frecce mostrano il flusso di informazioni tra i vari componenti descritti in questo documento. Le forme degli array sono mostrate tra parentesi con *s* , numero di sequenze (N_{seq} nel testo principale); *r* , numero di residui (N_{res} nel testo principale); *c* , numero di canali.[65]

AlphaFold aumenta significativamente la precisione nella previsione delle strutture proteiche grazie all'integrazione di nuove architetture di reti neurali e a metodi di addestramento che si avvalgono di vincoli evolutivi, fisici e geometrici. Le implicazioni di AlphaFold sono diverse. Questa tecnologia ha il potenziale per rivoluzionare la biologia, la medicina e molte altre discipline. Si stima, inoltre, che AlphaFold possa accelerare lo sviluppo di nuovi farmaci per combattere malattie come il cancro, l'Alzheimer e il Parkinson.

[65] Jumper, J., Evans, R., Pritzel, A. et al. Previsione della struttura proteica altamente accurata con AlphaFold. Nature 596, 583–589 (2021).

2.3 Rischi dell'utilizzo dell'IA

L'intelligenza artificiale (IA) sta trasformando rapidamente il settore sanitario, offrendo nuove opportunità, tuttavia, accanto ai vantaggi, come già anticipato prima, l'utilizzo dell'IA comporta una serie di rischi e/o limitazioni che devono essere attentamente valutate. Di seguito elencate alcune:

- Disponibilità e qualità dei dati: l'implementazione completa della tecnologia dei Big Data può essere ostacolata da diversi fattori, come l'impossibilità di accedere ai dati a causa di problematiche legate alla privacy, la frammentazione dei dati, la mancanza di armonizzazione tra le fonti e l'assenza di infrastrutture adeguate. In alcuni casi, può anche mancare un numero sufficiente di dati; un esempio è rappresentato dai dati relativi a pazienti affetti da malattie rare. Una strategia per ottenere ulteriori dati è la cosiddetta "data augmentation", utilizzata per generare immagini aggiuntive per l'addestramento dei modelli di deep learning in ambito radiologico. Nella data augmentation tradizionale, nuove immagini vengono create a partire da immagini originali attraverso diverse trasformazioni. Nella data augmentation sintetica, si impiega una rete generativa avversaria (GAN), un tipo di rete neurale, per produrre ulteriori immagini tematiche che presentano una distribuzione statistica simile a quella del dataset iniziale. L'uso delle GAN non si limita alla generazione di immagini sintetiche, ma si estende anche alla creazione di dati sintetici di altre tipologie. È interessante notare che l'aumento della generazione di dati sintetici potrebbe teoricamente portare a una situazione che gli esperti di intelligenza artificiale hanno recentemente iniziato a esaminare, definita da alcuni "model collapse" e da altri "model autophagy disorder" (MAD). Questa disfunzione si verifica quando i modelli si alimentano delle informazioni che hanno generato, causando un deterioramento significativo della qualità e della capacità generativa dei modelli stessi, con conseguenti risultati indesiderati e/o imprevedibili.

- Questioni etiche: l'uso dei modelli di intelligenza artificiale solleva diverse problematiche etiche, è importante considerare le capacità dei modelli generativi, che non solo possono generare dati sintetici, ma anche modificare video e immagini (deepfakes) o, in casi più gravi, falsificare interi set di dati a sostegno di specifiche ipotesi scientifiche. Tali frodi possono essere difficili da individuare; la loro rilevazione non dipende tanto da competenze informatiche o matematiche, quanto piuttosto dalla conoscenza del settore di pubblicazione, dall'esperienza dei revisori e, in generale, dalla qualità dell'intero processo di revisione scientifica. Inoltre, i modelli di IA potrebbero contribuire a creare o amplificare disuguaglianze nell'accesso alle cure, discriminazioni sociali o essere utilizzati per profilare i clienti a fini assicurativi. Per affrontare queste questioni etiche, il Parlamento Europeo ha recentemente approvato l'IA Act, la legge sull'intelligenza artificiale, il cui obiettivo è di proteggere

i diritti fondamentali, la democrazia, lo Stato di diritto e la sostenibilità ambientale dai sistemi di IA ad alto rischio, promuovendo allo stesso tempo l'innovazione e assicurando all'Europa un ruolo guida nel settore. Il regolamento stabilisce obblighi per l'IA sulla base dei possibili rischi e del livello d'impatto.[66]

- Assenza di validazione clinica: è fondamentale che l'efficacia e la sicurezza degli strumenti di intelligenza artificiale vengano valutate in modo approfondito tramite studi clinici prima di essere adottati nella pratica clinica quotidiana, analogamente a quanto avviene per l'introduzione di un nuovo metodo analitico. La validazione clinica implica un rigoroso test degli algoritmi di IA su ampi e diversificati set di dati, nonché un confronto delle loro prestazioni con altri strumenti standard, per identificare eventuali effetti indesiderati.

- Limitata generalizzazione dei modelli: si riferisce all'abilità di un modello addestrato di fare previsioni precise su dati nuovi. La generalizzazione è un obiettivo cruciale per un modello di machine learning. Il metodo principale per valutare la generalizzazione consiste nel suddividere i dati in tre set distinti: il set di addestramento, utilizzato per addestrare il modello; il set di validazione, impiegato per ottimizzare il modello attraverso la regolazione di alcuni iperparametri; e il set di test, che serve a valutare la capacità di generalizzazione del modello. In situazioni in cui i dataset sono limitati, si possono adottare varianti di questo approccio, come la validazione incrociata k-fold. Un esempio recente che illustra la limitata generalizzazione dei modelli di machine learning è rappresentato dai numerosi modelli sviluppati per diagnosticare la positività al COVID-19: quelli addestrati con dati provenienti dai pazienti della prima ondata, nonostante mostrassero elevate prestazioni secondo diverse metriche, hanno evidenziato una bassa accuratezza diagnostica con i pazienti delle ondate successive. Questa discrepanza è riconducibile alla diversa tipologia di pazienti analizzati.

- Limitata trasparenza e scarsa interpretabilità del modello: l'interpretabilità, o spiegabilità, si riferisce al modo in cui un modello opera e come genera determinati output o prende decisioni. La bassa interpretabilità è particolarmente problematica per gli algoritmi di deep learning, che vengono definiti "black box" (scatole nere), poiché si conosce l'input e si osserva l'output, ma si ignora completamente la relazione funzionale tra i due. La trasparenza, invece, riguarda il processo di progettazione, sviluppo, raccolta dati, verifica, distribuzione, utilizzo e monitoraggio del modello. Per mitigare i potenziali effetti negativi legati alla scarsa trasparenza, è consigliabile fornire una documentazione

[66] European parliament, Il Parlamento europeo approva la legge sull'intelligenza artificiale, 2024

dettagliata sulla creazione, utilizzo, validazione del modello, valutazione dell'impatto e garanzia della privacy, oltre a una descrizione delle parti interessate e degli obiettivi da raggiungere con l'implementazione dell'algoritmo. L'IA Act stabilisce già alcune limitazioni per promuovere la trasparenza nelle applicazioni di intelligenza artificiale, incluso l'obbligo di informare le persone che interagiscono con un sistema controllato dall'IA. La necessità di affrontare i problemi legati alla limitata interpretabilità dei modelli di IA ha portato, negli ultimi anni, alla nascita di una nuova disciplina nota come IA spiegabile (Explainable Artificial Intelligence o XAI), che comprende processi e metodi che permettono agli utenti di comprendere e ritenere attendibili i risultati e gli output generati dagli algoritmi di apprendimento automatico. Tra i metodi più noti di IA spiegabile ci sono SHAP (Shapley Additive Explanations), che aiuta a comprendere come i singoli attributi influenzano le previsioni del modello; LIME (Local Interpretable Model-agnostic Explanations), che scompone modelli complessi in componenti più semplici e interpretabili; e LRP (Layer-wise Relevance Propagation), che identifica le caratteristiche dei vettori in ingresso che hanno il maggiore impatto sull'output di una rete neurale.

- Attribuzione della responsabilità legale: con l'emergere di sistemi di intelligenza artificiale che operano in modo autonomo, diventa cruciale stabilire chi debba rispondere per eventuali danni o decisioni errate, nonché per le conseguenze sociali, economiche e sanitarie che ne derivano. Questa è una questione urgente che richiede l'istituzione di riferimenti normativi specifici. È importante considerare che la questione della responsabilità legale è strettamente legata ad altre caratteristiche dei modelli di machine learning, come la spiegabilità, la trasparenza, l'affidabilità, la necessità di validazione metodologica e clinica, e la protezione della privacy.

- Attacchi informatici: gli algoritmi di intelligenza artificiale possono essere vulnerabili ad attacchi da parte di malintenzionati, con possibili conseguenze gravi, come diagnosi imprecise o non eseguite, malfunzionamenti di sistemi autonomi e furto di informazioni.[67]

[67] Matteo Vidali, 'Intelligenza Artificiale in Medicina: implicazioni e applicazioni, sfide e opportunità', 2024

2.4 L'Intelligenza Artificiale ed il diritto alla salute nella regolazione europea ed internazionale

Al fine di creare un mercato interno ben funzionante per i sistemi di intelligenza artificiale (AI), la Commissione europea ha recentemente proposto la legge sull'intelligenza artificiale. Tuttavia, questa proposta legislativa presta un'attenzione limitata ai rischi che l'utilizzo dell'intelligenza artificiale comporta per i diritti dei pazienti. La maggior parte della legislazione dell'UE in materia sanitaria si basa sulle disposizioni dell'art. 114 TFUE[68].

Gli strumenti giuridici dell'UE che regolano la salute, come il regolamento sui dispositivi medici (MDR) e il regolamento generale sulla protezione dei dati (GDPR), non sono adatti alle sfide specifiche che l'IA comporta e non forniscono una soluzione completa alle sue minacce per i pazienti[69]. L'UE sta facilitando e promuovendo l'uso e la disponibilità dell'IA nel settore sanitario in Europa, ma non è chiaro se possa fornire la tutela concomitante dei diritti dei pazienti. Questa lacuna giuridica potrebbe portare a trascurare la posizione dei pazienti in Europa quando l'intelligenza artificiale sanitaria diventerà una pratica comune. Pertanto, appare legittimo chiedersi se l'approccio europeo all'IA fornisce un'adeguata protezione ai diritti dei pazienti alla luce dell'attuale quadro legislativo[70].

L'UE tutela i diritti dei pazienti in relazione ad ambiti specifici, come la direttiva sui diritti dei pazienti in relazione alla mobilità transfrontaliera dei pazienti (direttiva sui diritti dei pazienti transfrontalieri). Inoltre, i diritti dei pazienti sono riconosciuti dalla CGUE in relazione ai diritti fondamentali, come la privacy sanitaria[71]. La Carta dei diritti fondamentali dell'UE e la Convenzione europea dei diritti dell'uomo (CEDU) sono le fonti giuridiche più importanti in materia di diritti dei pazienti[72]. L'articolo 3 della Carta dei diritti fondamentali dell'UE sancisce il diritto all'integrità fisica e mentale e il diritto al consenso informato nell'ambito della medicina e della biologia, nonché l'implicito diritto di rifiutare le cure mediche[73].

I diritti dei pazienti relativi all'accesso all'assistenza sanitaria e ai medicinali sono connessi al diritto alla dignità umana (articolo 1 Carta dei diritti fondamentali dell'UE), al divieto di trattamenti inumani

---

[68] Delhomme, "Emancipating Health from the Internal Market: For a Stronger EU (Legislative) Competence in Public Health", 11 European Journal of Risk Regulation (2020), 747–756; Garben, "Competence Creep Revisited", 57 JCMS: Journal of Common Market Studies (2019), 205–222; Causa C-376/98, Germania contro Parlamento europeo e Consiglio (pubblicità del tabacco). Sentenza della Corte plenaria del 5 ottobre 2000.

[69] Spina A., *La medicina degli algoritmi: Intelligenza Artificiale, medicina digitale e regolazione dei dati personali*, in Pizzetti F. (a cura di*), Intelligenza Artificiale, protezione dei dati personali e regolazione*, Giappichelli, Torino, 2018, pp. 319

[70] Pizzetti F., *Intelligenza artificiale e salute: il sogno dell'immortalità alla prova del GDPR*, in *Agendadigitale.eu*, 2017.

[71] Palm et al., "Patients' rights: from recognition to implementation" in, Achieving Person-Centred Health Systems: Evidence, Strategies and Challenges (Cambridge University Press, 2020), pp. 347–386.

[72] De Ruijter, *EU Health Law & Policy: The Expansion of EU Power in Public Health and Health Care* (Oxford University Press, Oxford: 2019)

[73] Pizzetti F., *Privacy e il diritto europeo alla protezione dei dati personali: dalla Direttiva 95/46 al nuovo Regolamento europeo*, Giappichelli, Torino, 2016, pp. 56-57.

(articolo 4 Carta dei diritti fondamentali dell'UE) e al divieto di discriminazione (articoli 20-26 Carta dei diritti fondamentali dell'UE)[74]. Altre fonti giuridiche rilevanti sono costituite dalla Convenzione europea sui diritti umani e la biomedicina (Convenzione di Oviedo) e i principi generali del diritto dell'UE.31 Gli strumenti del Consiglio d'Europa, la CEDU e la Convenzione di Oviedo, fanno parte del diritto dell'UE attraverso l'interpretazione giudiziaria, come principi generali del diritto dell'UE[75]. Il quadro normativo europeo sui diritti dei pazienti è ispirato a principi etici e legali degli Stati membri dell'UE, sia informalmente che direttamente come principi generali del diritto. 34 La legislazione nazionale spesso collega i diritti dei pazienti agli obblighi legali degli operatori sanitari, come il diritto al consenso informato e il dovere di informare[76].

In linea di massima, si può sostenere che i diritti dei pazienti nell'UE si ergono su tre pilastri: autonomia, dignità umana e fiducia. Nel loro insieme, l'autonomia, la dignità umana e la fiducia si pongono alla base dei diritti dei pazienti degli Stati membri dell'UE, e precisamente: il diritto all'informazione, il diritto al consenso informato e il diritto alla protezione dei dati medici. Questi diritti possono essere riscontrati in tutti gli Stati membri dell'UE[77].

La Commissione europea accoglie con favore l'introduzione della tecnologia dell'intelligenza artificiale nel mercato unico (digitale) e ha espresso il desiderio che l'UE diventi un leader globale nel settore dell'intelligenza artificiale[78]. La salute è spesso concepita come la più grande opportunità di mercato per l'intelligenza artificiale alla luce dei vantaggi che offre quest'ultima nell'ottica socioeconomica al mercato interno dell'UE[79]. Tuttavia, l'intelligenza artificiale può anche comportare gravi rischi per i diritti fondamentali tutelati dal diritto dell'UE. La mancanza di trasparenza sull'esatto funzionamento dell'IA mette sotto pressione i valori dell'UE come la dignità umana e l'autonomia personale poiché l'IA viene regolarmente utilizzata per manipolare le persone. Anche il diritto di accesso alle informazioni è a rischio a causa del ruolo degli algoritmi nella diffusione della disinformazione[80]. Inoltre, a causa di distorsioni nei dati di addestramento o nell'algoritmo, la tecnologia dell'intelligenza artificiale può portare a disuguaglianze, che possono incidere sul divieto di discriminazione.

Nell'aprile 2021 la Commissione Europea ha presentato una proposta per la regolamentazione dell'intelligenza artificiale con l'obiettivo di creare un mercato interno ben funzionante per i sistemi

---

[74] *Ibidem*.
[75] Smith M., "Patients and doctors: rights and responsibilities in the NHS", 5 *Clin Med* (2005), 501–502.
[76] Will, "A Brief Historical and Theoretical Perspective on Patient Autonomy and Medical Decision Making: Part II: The Autonomy Model", 139 *Chest* (2011), 1491–1497.
[77] *Autonomy and Trust in Bioethics* (Cambridge University Press 2002), 18-20.
[78] *Ibidem*.
[79] Davenport and Kalakota, "The potential for artificial intelligence in healthcare", 6 *Future Healthcare Journal* (2019), 94–98.
[80] Amisha *et al.*, "Overview of artificial intelligence in medicine", 8 *Journal of Family Medicine and Primary Care* (2019), 2328–2331; *KPMG*, Inventarisatie AI-toepassingen in gezondheid en zorg in Nederland. Onderzoek naar de stand van zaken in 2020, 2020.

di intelligenza artificiale che tuteli adeguatamente i diritti e i valori dell'UE, senza ostacolare l'innovazione[81]. Lo scopo principale della proposta è migliorare il funzionamento del mercato interno dell'IA stabilendo norme per lo sviluppo, la commercializzazione e l'uso sulla base dell'articolo 114 TFUE. L'AIA mira ad armonizzare le norme sull'IA e a creare un ecosistema di fiducia nell'IA allineandone l'uso ai valori, ai diritti e ai principi fondamentali europei. In questo contesto è importante notare che l'AIA non disciplina specificatamente l'IA sanitaria, ma si concentra sull'IA in generale.

L'attuale quadro giuridico per l'intelligenza artificiale sanitaria a livello dell'UE è articolato su più livelli ed è costituito da: regolamentazione specifica per le tecnologie sanitarie (ad esempio regolamenti sui dispositivi medici), regolamentazione specifica per questioni legate alla tecnologia (ad esempio legislazione relativa all'IA sanitaria) il mercato unico digitale), la normativa sui diritti fondamentali (ad es. CFREU e GDPR) ed infine, la normativa sulla tutela dei consumatori (ad es. norme sulla responsabilità del prodotto e sulle pratiche commerciali sleali)[82]. Sebbene l'attuale quadro dell'UE possa scongiurare alcuni rischi comuni al processo decisionale sanitario relativo all'intelligenza artificiale, sembra essere insufficiente per proteggere adeguatamente i pazienti in caso di una svolta algoritmica nel contesto sanitario[83].

Per quanto riguarda la regolazione sui dispositivi medici, Il Regolamento (UE) 2017/745 adottato in sostituzione della Direttiva 93/42/CEE) mira a garantire un elevato livello di salute e sicurezza dei dispositivi medici, sostenendo al contempo l'innovazione. Ai sensi e per gli effetti dell'art. 2 del Regolamento (UE) 2017/745, si definisce dispositivo medico "*qualsiasi strumento, apparecchio, apparecchio, software, impianto, reagente, materiale o altro articolo destinato a essere utilizzato, da solo o in combinazione, per esseri umani per uno o più dei seguenti scopi medici specifici*" quali, ad esempio, "*diagnosi, prevenzione, monitoraggio, previsione, prognosi, trattamento o attenuazione di malattie*"[84]. Tale Regolamento esclude espressamente il software destinato a scopi generali e a finalità legate allo stile di vita e al benessere, anche se utilizzati nel rapporto di cura. Le applicazioni di intelligenza artificiale qualificabili come dispositivi medici sono soggette a una valutazione di conformità. I requisiti esatti dipendono dalla classe di rischio: maggiore è il rischio per il paziente, più alta è la classe e più severe sono le regole. Tale Regolamento stabilisce principalmente norme tecniche relative alla tutela dell'incolumità fisica e della salute dei pazienti ed è meno focalizzata sulla tutela dei diritti dei pazienti. Tuttavia, il Regolamento in questione richiede un accesso adeguato alle

[81] Pizzetti F., *op. cit,* p. 380.
[82] Evas, 'European Framework on Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies. European Added Value Assessment' (European Parliamentary Research Service 2020) PE 654.179;
[83] Van Kolfschooten, 'The mHealth Power Paradox: Improving Data Protection in Health Apps through Self-Regulation in the European Union', in: I. Glenn Cohen, Timo Minssen, W. Nicholson Price II, Christopher Robertson, and Carmel Shachar, *The Future of Medical Device Regulation: Innovation and Protection*, Cambridge: Cambridge University Press 2021.
[84] Art. 2, Regolamento (UE) 2017/745.

informazioni per gli utenti e i produttori sono obbligati a informare gli utenti sui "possibili rischi residui", che possono contribuire ai problemi relativi alla trasparenza dell'intelligenza artificiale. Considerato lo scopo dell'MDR, tale requisito sembra riguardare principalmente i rischi fisici. Per quanto riguarda la privacy e la protezione dei dati, tale Regolamento tutela la privacy sanitaria principalmente con riferimento al GDPR e non fissa requisiti aggiuntivi. A causa della limitata considerazione delle questioni sanitarie e della tutela dei diritti dei pazienti, l'attuale quadro giuridico dell'UE relativo all'intelligenza artificiale sanitaria sembra essere impreparato ad affrontare le nuove sfide che il processo decisionale sanitario automatizzato comporta per i diritti dei pazienti.

2.5 Gli Italiani e l'Intelligenza artificiale

Il 27 marzo 2024 il Gruppo Unipol, in collaborazione con Ipsos, ha pubblicato un'indagine dettagliata realizzata per approfondire la percezione degli italiani riguardo l'IA e il loro rapporto con questa tecnologia, rivelando opinioni divergenti[85]. Da un lato, molti considerano l'IA come uno strumento in grado di trasformare il mondo del lavoro e aumentare l'efficienza dei processi, dall'altro, emergono preoccupazioni relative alla possibile perdita di posti di lavoro e al rischio di disinformazione.
Sono state condotte interviste CAWI (Computer-Assisted Web Interviewing) tra il 14 e il 22 febbraio 2024 ripartite tra:

- popolazione italiana: 1.000 interviste a un campione nazionale rappresentativo della popolazione italiana di età 16-74 anni (rappresentativi di oltre 44 milioni di individui);
- Residenti nelle principali Aree Metropolitane italiane: 720 interviste Over Sample in 9 Aree Metropolitane (rappresentativi di oltre 13 milioni di individui), con 80 interviste circa per ciascuna area (• Nord Italia: Milano, Torino, Bologna, Verona • Centro Italia: Firenze, Roma • Sud Italia: Napoli, Bari, Cagliari).

L'intelligenza artificiale è ampiamente conosciuta in Italia, ma il suo utilizzo è ancora limitato, con Milano che spicca per conoscenza approfondita e utilizzo dell'AI (26% vs 19% del totale Italia). Le generazioni più giovani mostrano una conoscenza più profonda dell'AI rispetto a quelle più mature, e ne fanno un utilizzo maggiore (21% contro 5% dei Boomer).

[85] Quotidiano nazionale, '*Gli italiani e l'intelligenza artificiale: paure e opportunità. I risultati del sondaggio', 2024*

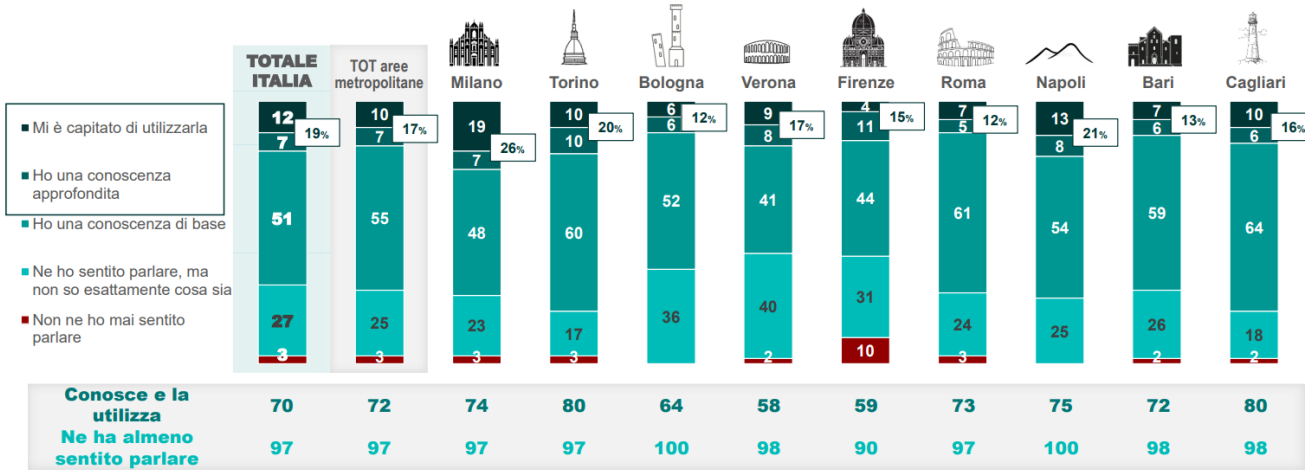


Figura 1 Quanto direbbe di conoscere l'Intelligenza Artificiale (AI)

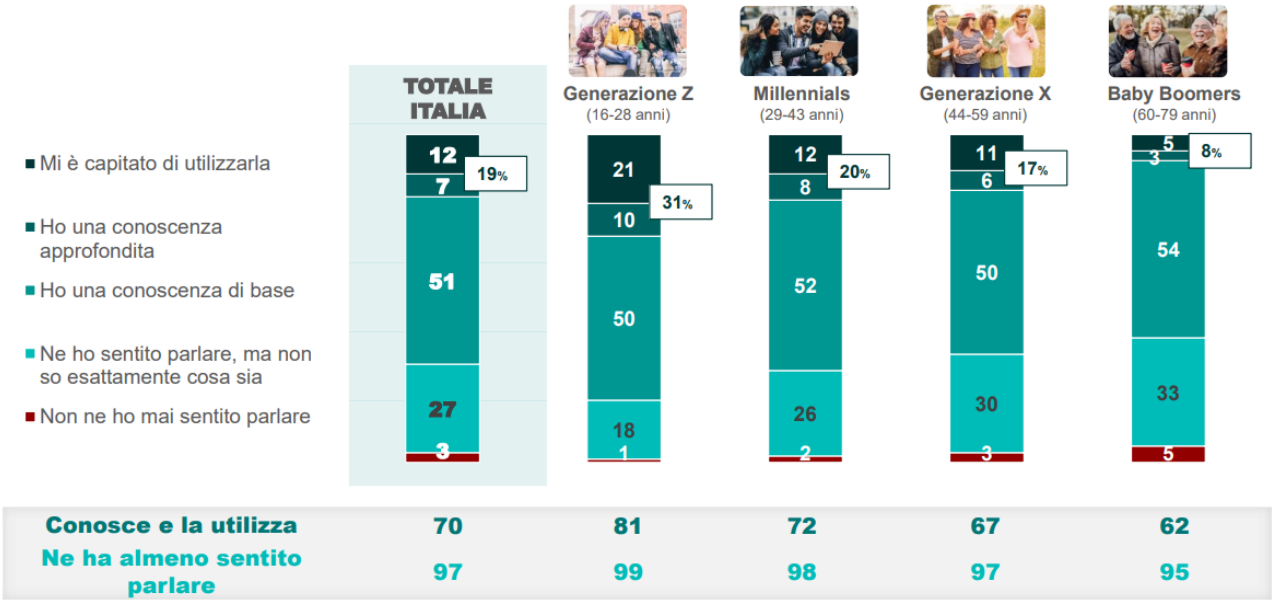


Figura 2 Quanto direbbe di conoscere l'Intelligenza Artificiale (AI)

Sebbene a livello nazionale ci sia un equilibrio tra effetti positivi e negativi, l'intelligenza artificiale genera preoccupazioni in particolare nelle aree metropolitane e tra le generazioni più anziane. Nelle città metropolitane, infatti, prevalgono sentimenti di diffidenza e ansia rispetto a quelli di attrazione e curiosità, con Verona e Firenze in prima linea. La Generazione Z, invece, mostra un maggiore interesse e curiosità nei confronti dell'AI, con il 25% rispetto al 18% della media nazionale.

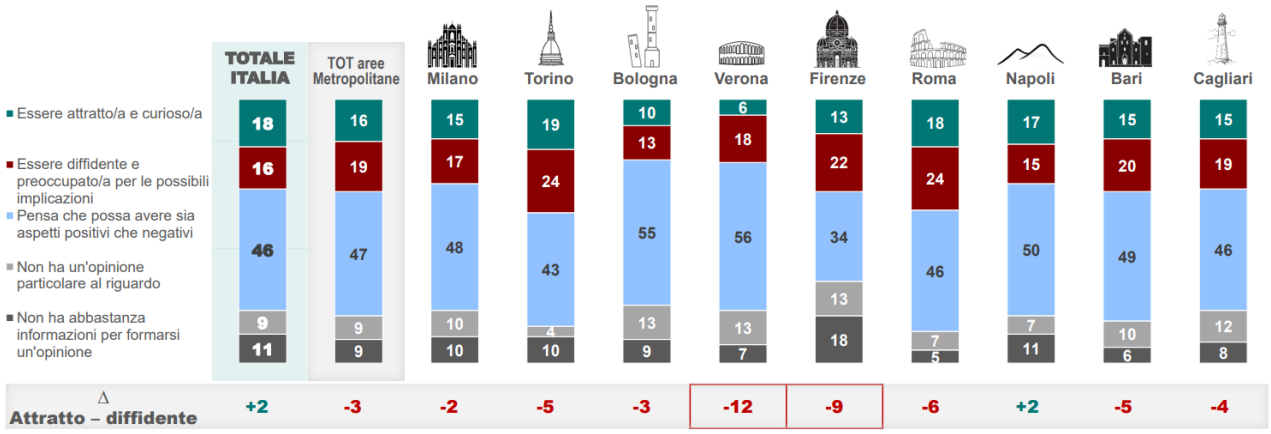


Figura 3 Riguardo all'uso dell'Intelligenza Artificiale, Lei direbbe di…?

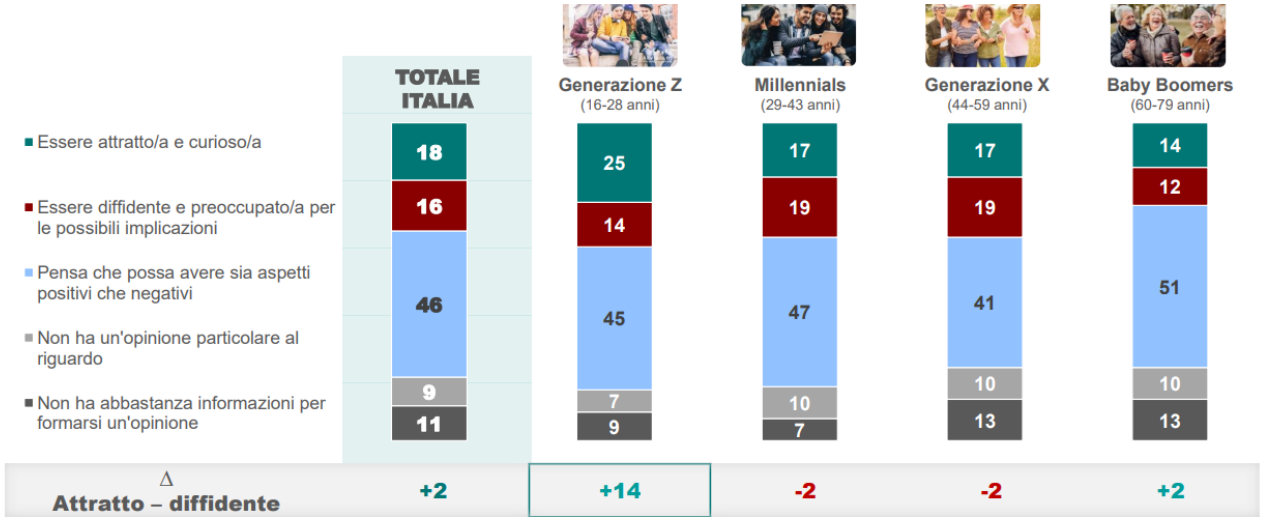


Figura 4 Riguardo all'uso dell'Intelligenza Artificiale, Lei direbbe di…?

In generale per 8 italiani su 10 l'AI porterà vantaggi lavorativi, soprattutto per la riduzione di errori umani, ma anche l'aumento della produttività, l'accesso a più informazioni, la semplificazione di compiti più complessi ed una maggiore flessibilità degli orari. Tuttavia, la perdita di posti di lavoro e la chiusura di imprese tradizionali è una preoccupazione diffusa sia nelle aree metropolitane che tra generazioni. Nonostante i potenziali vantaggi, gli svantaggi eventuali dell'AI prevalgono, soprattutto nelle aree metropolitane di Firenze e Bari, e tra i Boomer. L'AI secondo gli italiani porterà miglioramenti principalmente nella digitalizzazione della pubblica amministrazione, ma potrebbe peggiorare il clima. Gli effetti potenziali dell'AI variano tra le diverse generazioni e le città. Infatti, a

Milano e Torino si ipotizzano miglioramenti anche per lo shopping e le esperienze culturali. La mobilità migliorerà soprattutto per i Baresi e per la Gen Z. Secondo i Boomer migliorerà anche la velocità e precisione delle diagnosi mediche.

VANTAGGI SUL MONDO DEL LAVORO / SVANTAGGI SUL MONDO DEL LAVORO	TOTALE ITALIA	TOT aree Metropolitane	Milano	Torino	Bologna	Verona	Firenze	Roma	Napoli	Bari	Cagliari
ALMENO UN VANTAGGIO	81	82	83	83	86	84	75	83	83	72	85
ALMENO UNO SVANTAGGIO	87	89	84	91	92	88	87	93	88	86	88
Δ ALMENO UN VANTAGGIO - ALMENO UNO SVANTAGGIO	-6	-7	-1	-8	-6	-4	-12	-10	-5	-14	-3

	TOTALE ITALIA	Generazione Z (16-28 anni)	Millennials (29-43 anni)	Generazione X (44-59 anni)	Baby Boomers (60-79 anni)
ALMENO UN VANTAGGIO	81	87	82	80	77
ALMENO UNO SVANTAGGIO	87	87	87	86	86
Δ ALMENO UN VANTAGGIO - ALMENO UNO SVANTAGGIO	-6	--	-5	-6	-9

Figura 5 Secondo Lei, quali sono gli SVANTAGGI e i VANTAGGI che l'intelligenza artificiale avrà sul mondo del lavoro?

La preoccupazione per la disinformazione potenzialmente generata dall'AI è diffusa, e maggiore a Bari (80% vs 65% del totale Italia) e tra i Boomer (70%). Questa convinzione è più diffusa a Napoli e Cagliari (46% e 45%) e tra i Millennial (43%). Gli ambiti che potrebbero risentire maggiormente della disinformazione potenzialmente generata dall'AI sono la sicurezza (false minacce/allarmi: 34%), seguite dall'economia (28%) e la politica (27%). Di queste ultime due aree, sono preoccupati principalmente i Torinesi (33% per entrambe) e i Boomer (35% e 33%).

	TOTALE ITALIA	TOT aree Metropolitane	Milano	Torino	Bologna	Verona	Firenze	Roma	Napoli	Bari	Cagliari
MOLTO+ABBASTANZA	65	68	67	61	67	72	72	71	62	80	66
Molto preoccupato	24	26	26	26	24	24	30	25	23	35	28
Abbastanza preoccupato	41	42	41	35	43	48	42	46	39	45	38
Poco preoccupato	19	21	19	29	22	16	15	21	24	12	17
Per niente preoccupato	7	6	9	6	6	5	8	3	8	3	8
Non saprei	9	5	5	4	5	7	5	5	6	5	9
POCO+PER NIENTE	26	27	28	35	28	21	23	24	32	15	25

Figura 6 Quanto è preoccupato/a per l'uso dell'Intelligenza Artificiale nella creazione di disinformazione/ fake news?

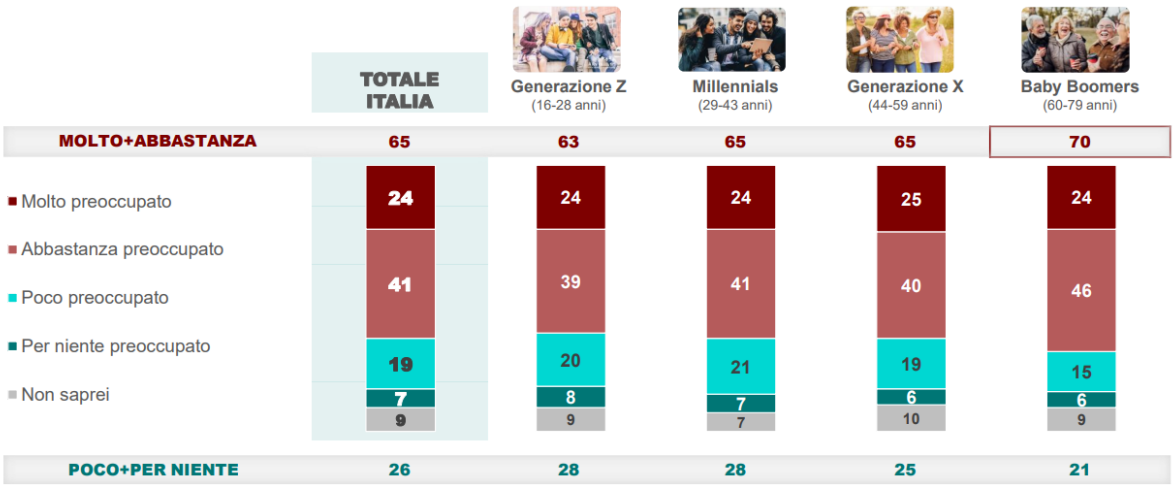


Figura 7 Quanto è preoccupato/a per l'uso dell'Intelligenza Artificiale nella creazione di disinformazione/ fake news?

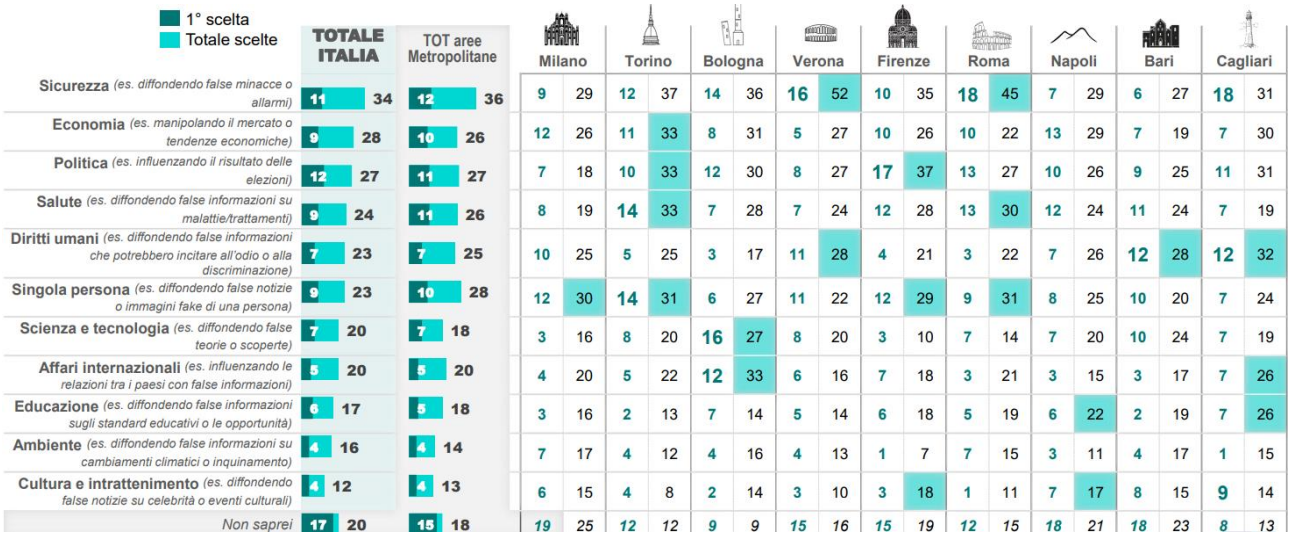
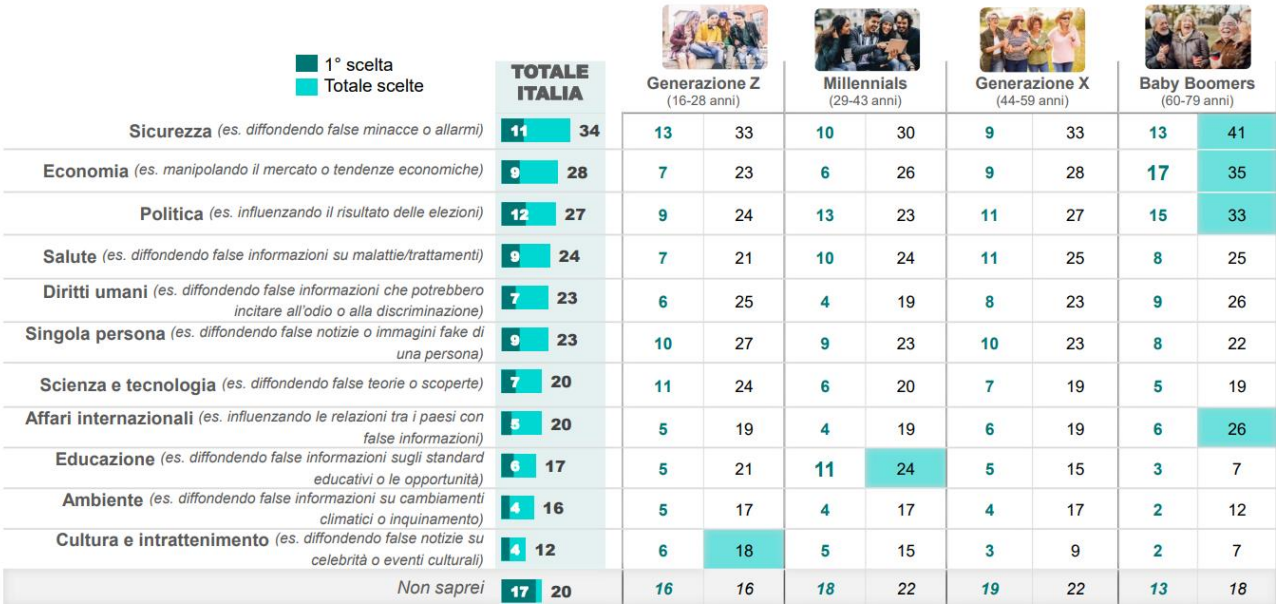


Figura 8 Secondo Lei, quali sono gli ambiti che potrebbero risentire maggiormente di una eventuale disinformazione generata dall'Intelligenza Artificiale?



	1° scelta	Totale scelte	Generazione Z (16-28 anni)		Millennials (29-43 anni)		Generazione X (44-59 anni)		Baby Boomers (60-79 anni)	
Sicurezza *(es. diffondendo false minacce o allarmi)*	11	34	13	33	10	30	9	33	13	41
Economia *(es. manipolando il mercato o tendenze economiche)*	9	28	7	23	6	26	9	28	17	35
Politica *(es. influenzando il risultato delle elezioni)*	12	27	9	24	13	23	11	27	15	33
Salute *(es. diffondendo false informazioni su malattie/trattamenti)*	9	24	7	21	10	24	11	25	8	25
Diritti umani *(es. diffondendo false informazioni che potrebbero incitare all'odio o alla discriminazione)*	7	23	6	25	4	19	8	23	9	26
Singola persona *(es. diffondendo false notizie o immagini fake di una persona)*	9	23	10	27	9	23	10	23	8	22
Scienza e tecnologia *(es. diffondendo false teorie o scoperte)*	7	20	11	24	6	20	7	19	5	19
Affari internazionali *(es. influenzando le relazioni tra i paesi con false informazioni)*	5	20	5	19	4	19	6	19	6	26
Educazione *(es. diffondendo false informazioni sugli standard educativi o le opportunità)*	6	17	5	21	11	24	5	15	3	7
Ambiente *(es. diffondendo false informazioni su cambiamenti climatici o inquinamento)*	4	16	5	17	4	17	4	17	2	12
Cultura e intrattenimento *(es. diffondendo false notizie su celebrità o eventi culturali)*	4	12	6	18	5	15	3	9	2	7
Non saprei	17	20	*16*	*16*	*18*	*22*	*19*	*22*	*13*	*18*

Figura 9 Secondo Lei, quali sono gli ambiti che potrebbero risentire maggiormente di una eventuale disinformazione generata dall'Intelligenza Artificiale?

L'adozione di regolamenti e leggi rigorose sull'uso dell'intelligenza artificiale è considerata la soluzione più efficace per contrastare la potenziale disinformazione generata da questa tecnologia. Questa opinione è particolarmente diffusa a Firenze e Roma, dove il sostegno raggiunge il 53% e il 44% della popolazione italiana, rispettivamente, e tra i Boomer, con una percentuale del 58%. Anche l'educazione e la formazione dei cittadini sono ritenute misure fondamentali, specialmente a Bologna (41%), Torino e Roma (40% rispetto al 31% della media nazionale). Per quanto riguarda Torino, Verona e il Sud Italia, si ritiene che la responsabilizzazione delle piattaforme media possa rappresentare un'altra misura efficace.

	TOTALE ITALIA	TOT aree Metropolitane	Milano	Torino	Bologna	Verona	Firenze	Roma	Napoli	Bari	Cagliari
ALMENO UNA MISURA	82	85	**88**	**88**	**90**	**80**	**86**	**86**	**80**	**77**	**83**
Regolamenti e leggi severi sull'uso dell'AI	44	47	46	**50**	46	40	**53**	**53**	38	43	44
Educazione e formazione dei cittadini	31	32	34	**40**	**41**	30	24	**40**	22	19	31
Sviluppo di tecnologie per rilevare la disinformazione	31	29	27	24	27	27	31	29	33	29	27
Responsabilizzare le piattaforme media nel monitorare e rimuovere la disinformazione	29	30	31	**34**	31	**35**	31	21	**34**	**34**	**34**
Non è necessario fare nulla	3	3	4	4	2	3	2	2	3	3	6
Non saprei	14	13	*8*	*8*	*9*	*18*	*12*	*13*	*17*	*20*	*12*

Figura 10 Secondo Lei, quali misure dovrebbero essere prese per prevenire l'eventuale disinformazione generata dall'Intelligenza Artificiale?

	TOTALE ITALIA	 Generazione Z (16-28 anni)	 Millennials (29-43 anni)	 Generazione X (44-59 anni)	 Baby Boomers (60-79 anni)
ALMENO UNA MISURA	**82**	**85**	**82**	**79**	**87**
Regolamenti e leggi severi sull'uso dell'AI	**44**	42	39	41	**58**
Educazione e formazione dei cittadini	**31**	33	34	30	25
Sviluppo di tecnologie per rilevare la disinformazione	**31**	33	28	29	35
Responsabilizzare le piattaforme media nel monitorare e rimuovere la disinformazione	**29**	28	25	30	32
Non è necessario fare nulla	**3**	1	4	5	2
Non saprei	**14**	*14*	*15*	*16*	*11*

Figura 11 Secondo Lei, quali misure dovrebbero essere prese per prevenire l'eventuale disinformazione generata dall'Intelligenza Artificiale?

CAPITOLO III: FUTURO

3.1 Il lavoro nell'era dell'IA

La relazione tra IA e occupazione è diventata ovvia nella letteratura di ricerca che ha prodotto un enorme database di articoli di ricerca in quest'area. La ricerca che di seguito sarà illustrata è una revisione sistematica della letteratura sullo stato dell'arte pubblicata nel 2018 sull'argomento[86].

La sostituzione dei singoli lavoratori con l'intelligenza artificiale e i robot è un argomento estremamente discusso. Alcuni sostengono che una parte significativa dei posti di lavoro è a rischio, mentre altri sostengono che i computer e i robot guideranno verso rivoluzioni di prodotto e quindi verso nuove incredibili professioni. Da questa prospettiva, è emersa una forte relazione tra Intelligenza Artificiale (IA) e Occupazione. Pertanto, entrambi i termini, IA e Occupazione, saranno enfatizzati nel presente articolo per dare al lettore una visione d'insieme su ciò che è citato nella letteratura su questa relazione, anche positivamente o negativamente.

Fernald (2014) ha affermato che a partire dal 1995, il progresso della produttività è accelerato negli Stati Uniti in base alla rivoluzione dell'Information Technology (IT). Tuttavia, i risultati di produttività dei tipi tradizionali di IT sono stati consumati a metà del 2000[87]. Di recente, l'influenza dell'intelligenza artificiale (IA) e della robotica, nota come quarta rivoluzione industriale, sull'economia e sulla società future sta attirando l'attenzione e sono comparsi molti argomenti ipotetici in merito ai probabili impatti della quarta rivoluzione industriale. In particolare, la sostituzione dei singoli lavoratori con l'IA e i robot è oggetto di un acceso dibattito. L'attuale studio di ricerca si propone i seguenti obiettivi:

- Rivedere sistematicamente la relazione tra intelligenza artificiale e occupazione;
- Fornire una revisione sistematica ben organizzata dello stato dell'arte dei due campi: intelligenza artificiale e occupazione;
- Consigliare una serie di nuovi argomenti che potrebbero costituire un nuovo database di ricerca per ricerche future e anche per potenziali ricercatori;
- Basarsi su definizioni chiare per le parole chiave dello studio e offrire ai lettori i risultati delle ricerche passate.

---

[86] Khadragy S., (2022) Artificial Intelligence and the future of employment; a systematic review of the state of the art literature, *International Journal of Mechanical Engineering, Vol. 7, Febbraio 2022, City University College of Ajman, Management Information Systems Department.*

[87] Fernald, J. G. (2014) 'Productivity and potential output before, during, and after the great recession', NBER
Macroeconomics Annual.

Tutto quanto sopra è considerato un indicatore per il recente studio sull'esistenza di una forte relazione tra AI e futuro dell'occupazione. Quindi, il presente studio mira a indagare la relazione tra AI e futuro dell'occupazione attraverso la letteratura. Per raggiungere l'obiettivo dello studio, vengono affrontate le seguenti domande:

'Esiste una relazione nella letteratura tra intelligenza artificiale e futuro dell'occupazione?'

'Quale effetto avrà l'intelligenza artificiale sul futuro dell'occupazione?'

La tecnologia e il suo sviluppo hanno un impatto importante sulla forza lavoro. Rivedere questo impatto sarà essenziale per istruire nuovi ruoli e politiche che possano supportare mercati del lavoro efficienti per il valore dei dipendenti, dei datori di lavoro e delle loro istituzioni[88].

Mentre questo rapido sviluppo tecnologico può mettere a repentaglio l'occupazione (il che non è una novità), questo sviluppo può influenzare il campo dell'occupazione in due modi: direttamente, sostituendo i dipendenti nelle attività che svolgevano in precedenza[89]; indirettamente, ampliando la domanda del mercato del lavoro grazie allo sviluppo tecnologico[90]. In questo momento, la necessità di alcuni lavori specifici che richiedono una serie di competenze cognitive e una routine manuale scomparirà.

Secondo Bessen (2017), la tecnologia ha drasticamente ridotto le carriere negli ultimi tempi[91]. Ma prima di allora, per più di cento anni, l'occupazione è aumentata, anche in settori con una rapida rivoluzione tecnologica. La domanda qui è: qual è la differenza tra ora e allora? La domanda era estremamente flessibile all'inizio e poi è diventata inflessibile. L'impatto dell'intelligenza artificiale (IA) sulle carriere dipenderà in modo significativo anche dalla natura delle esigenze del mercato del lavoro. D'altro canto, Agrawal, Gans e Goldfarb (2019) hanno esaminato un framework per indagare le inferenze dell'automazione e dell'intelligenza artificiale sulla necessità di forza lavoro, stipendi e occupazione[92]. Il loro framework basato sulle attività ha evidenziato l'impatto dell'associazione che l'automazione produce quando i motori e l'intelligenza artificiale sostituiscono i lavoratori nelle attività che erano soliti svolgere. Gli autori hanno affermato che l'impatto del movimento tende a ridurre la necessità di lavoratori e stipendi. Ma è contrastato da un impatto sulla produttività, derivante dai risparmi sui costi prodotti dall'automazione, che aumentano la necessità di lavoratori in lavori non automatizzati. L'effetto sulla produttività è ottenuto tramite un'ulteriore crescita del capitale e l'estensione dell'automazione; entrambi aumentano ulteriormente la necessità di manodopera. Questi impatti sono imperfetti. Anche quando sono solidi, l'automazione espande la produzione per

[88] Tambe, P. (2014) 'Big data investment, skills, and firm value', Management Science.
[89] Moull, K. E. (2017) Exploring the use of biomechanical metrics in the validation of physical employment standards, ProQuest Dissertations and Theses.
[90] Ndyali, L. (2016) 'Higher Education System and Jobless Graduates in Tanzania', Journal of Education and Practice.
[91] Bessen, J. E. (2017) 'AI and Jobs: The Role of Demand', SSRN Electronic Journal.
[92] Agrawal, A., Gans, J. and Goldfarb, A. (2019) 'Artificial Intelligence, Automation, and Work', in The Economics of Artificial Intelligence.

lavoratore più degli stipendi e riduce la quota di lavoro nel reddito nazionale. Un altro punto evidenziato nello stesso documento di ricerca è che la forza più influente contro l'automazione è la produzione di un nuovo set di funzioni ad alta intensità di lavoro, che riporta i lavoratori a nuovi compiti e tende ad espandere la quota di lavoro per compensare l'effetto dell'automazione.

Inoltre, gli autori hanno evidenziato i limiti e le carenze che riducono la modifica dell'economia e del mercato del lavoro all'automazione e peggiorano i miglioramenti della produttività derivanti da questa modifica: incompatibilità tra le competenze richieste da un insieme di nuove tecnologie e la probabilità che l'automazione venga presentata a un ritmo estremo, eventualmente a scapito di altre tecnologie che arricchiscono la produttività.

Un altro gruppo di documenti di ricerca ha presentato un tipo di incoerenza nell'era dell'occupazione. La disoccupazione è ai minimi storici, il che significa che le aziende sono alla ricerca di lavoratori in grado di mantenere le cose realizzabili. Allo stesso tempo, l'implementazione della digitalizzazione, dell'intelligenza artificiale e dell'automazione avverte di sostituire molti lavoratori. Un'altra idea è anche affermata nello stesso gruppo di documenti di ricerca: i dirigenti delle istituzioni sono affamati di alti livelli sia di servizi che di produzioni. Allo stesso tempo saranno affamati di tutti i servizi con competenze basate sulla tecnologia per guidare nuovi affari in poco tempo e con grande precisione. Con questo scenario, i dipendenti saranno desiderosi di acquisire queste competenze. Tuttavia, gli amministratori delle organizzazioni non hanno ancora l'immagine completa di questa idea. McKendrick (2018) ha affermato che il 46% dei dipendenti che stanno lavorando sulle proprie competenze riconosce che i loro datori di lavoro non apprezzano i dipendenti che non hanno aggiornato le proprie competenze tecnologiche[93]. Nonostante la sua ampia mancanza di familiarità, l'IA è una tecnologia che sta cambiando ogni ambito della vita[94]. È uno strumento di ampia diffusione che consente alle persone di riconsiderare il modo in cui assimilano le informazioni, analizzano i dati e utilizzano i risultati per sviluppare il processo decisionale. Quindi, l'obiettivo generale di questa impressione completa è chiarire il concetto di IA al pubblico dei decisori, dei leader di vista, degli spettatori interessati e rivelare come l'IA stia già alterando il mondo e sollevando importanti questioni per la società, l'economia e la governance.

Di conseguenza, l'idea di questo documento di ricerca è giunta a dare diverse definizioni di IA e a rivelare la relazione tra essa e il futuro del campo dell'occupazione. Per raggiungere questo obiettivo, l'IA sarebbe stata discussa da diverse visioni di ricerca e per rafforzare le idee non solo per i ricercatori, ma anche per le persone che non sono specializzate nei campi dell'IA e dell'occupazione.

[93] McKendrick in Xiong, X. (2019) 'Analysis of the Status Quo of Artificial Intelligence and Its Countermeasures'.

[94] Au-Yong-Oliveira, M. et al. (2019) 'The role of AI and automation on the future of jobs and the opportunity to change society', in Advances in Intelligent Systems and Computing

Un gruppo di studi di ricerca ha esaminato il rischio occupazionale innescato dalle nuove tecnologie, come l'intelligenza artificiale (IA) e la robotica, applicando la possibilità di informatizzazione tramite dati sull'occupazione internazionale. La nuova idea di questo tipo di studi di ricerca è la considerazione dell'eterogeneità regionale nei mercati del lavoro in base alla distribuzione geografica approssimativa delle professioni, che si osserva in particolar modo tra lavoratori uomini e donne. I ricercatori hanno scoperto che le lavoratrici sono soggette a rischi di informatizzazione più elevati rispetto ai lavoratori uomini, poiché saranno coinvolte in lavori con un'alta probabilità di informatizzazione. Questa propensione è più evidente nelle grandi città. I risultati di tali studi di ricerca raccomandano che fornire da soli investimenti extra di capitale umano non è sufficiente come strategia di mitigazione delle sfide contro le nuove tecnologie e i rappresentanti vogliono decidere le sfide strutturali del mercato del lavoro, come i pregiudizi di genere per la progressione di carriera e il contributo nei compiti decisionali, nell'era dell'IA per mitigare il rischio diseguale di informatizzazione tra i dipendenti[95].

D'altro canto, e a causa del fatto che c'è un enorme progresso tecnologico e una tendenza all'automazione nel mondo della forza lavoro in questo XXI secolo, l'intelligenza artificiale (IA) non è solo classificata come un recente prodotto avanzato dello sviluppo tecnologico per le persone, ma è anche considerata una nuova grave minaccia per i dipendenti nella nuova rivoluzione industriale nell'era informatica. Da questa prospettiva, un altro gruppo di studi di ricerca è emerso per esplorare l'impatto dell'IA sulle condizioni di lavoro, sugli ambienti e sulle competenze nelle postazioni di lavoro. La maggior parte di questi studi di ricerca è considerata una ricerca esplorativa o qualitativa, i dati vengono raccolti intervistando alcuni dipendenti nei settori industriali che hanno espresso opinioni serie sulle rivoluzioni industriali, sulla tendenza robotica e sull'IA. Come risultato del processo menzionato, è diventato evidente affermare che l'IA ha un grande impatto sulle condizioni di lavoro nell'era informatica. Uno di questi impatti è che l'IA influisce direttamente e indirettamente sulle relazioni tra esseri umani sul posto di lavoro. Un altro impatto affermato in diversi studi di ricerca è quello di influenzare le conoscenze e le competenze lavorative richieste per un lavoro in qualsiasi mondo del lavoro. Inoltre, l'IA influenzerà negativamente l'impegno istituzionale e l'identità organizzativa. Oltre a questi impatti inaspettati, l'IA è comunemente considerata la ragione principale di una crescita della disoccupazione così grave in molte società. A questo proposito e per affrontare queste nuove sfide che possono essere prodotte dall'IA e dal rapido investimento tecnologico nei luoghi di lavoro, vengono suggerite alcune raccomandazioni significative, tra cui l'IA è realizzata per aumentare la capacità dei dipendenti piuttosto che sostituire i lavoratori umani, dando obblighi e requisiti all'IA che è deliberatamente realizzata e utilizzata per sostituire i lavoratori umani e

[95] Nobuaki, H. and Keisuke, K. (2018) 'Regional Employment and Artificial Intelligence in Japan', Discussion papers.

controllando l'applicazione nei luoghi che dovrebbero essere un vero supporto per migliorare le intelligenze artificiali e le cui suddivisioni non possono essere fornite per essere controllate dall'IA è essenziale[96].

D'altro canto, con l'eccessiva considerazione dei governi e la graduale formazione della struttura industriale, l'intelligenza artificiale è a un livello di rapido sviluppo. L'innovazione nello sviluppo dell'intelligenza artificiale è derivata da tre elementi: l'aumento di enormi volumi di dati, la comparsa di numerosi algoritmi eccellenti e il meraviglioso miglioramento delle prestazioni dell'hardware del computer. Inoltre, qualsiasi nuova tecnologia e nuova applicazione hanno due ali. La tecnologia dell'intelligenza artificiale e la sua applicazione non solo ottengono l'idoneità per le persone, ma gestiscono anche i rischi per la sicurezza[97].

3.2 Uno sguardo al futuro: COHUMAIN

La collaborazione tra esseri umani e intelligenza artificiale (IA) sta diventando sempre più rilevante, poiché i progressi tecnologici stanno trasformando il nostro modo di interagire con le macchine. L'IA ha dimostrato un notevole potenziale nel migliorare le capacità cognitive e operative degli individui, specialmente in contesti complessi e in continua evoluzione. Tuttavia, per sfruttare appieno questo potenziale, è essenziale avere una comprensione più approfondita e integrata dei sistemi socio-tecnologici che uniscono intelligenza umana e artificiale. Pranav Gupta e altri propongono un programma di ricerca per *l'intelligenza collettiva uomo-macchina* (COHUMAIN), un dominio di ricerca interdisciplinare per facilitare lo sviluppo di modelli olistici che informino la progettazione e lo studio delle dinamiche di collaborazione nei sistemi sociotecnici, ed in particolare presentano una nuova architettura sociocognitiva, il modello di sistemi transattivi dell'intelligenza collettiva (TSM-CI)[98]. Mentre le architetture cognitive si concentrano principalmente su agenti autonomi, ponendo la questione di come questi percepiscano, comprendano e agiscano nell'ambiente, le architetture sociocognitive vanno oltre, esplorando come più agenti (umani e IA) collaborino per risolvere problemi collettivamente e in modo interdipendente. Tuttavia, integrare queste architetture in un approccio interdisciplinare, come richiesto dalla ricerca COHUMAIN, non è affatto semplice. Un punto cruciale riguarda la natura degli obiettivi delle IA: a differenza degli esseri umani, le IA non

[96] Saithibvongsa, P. and Yu, J. E. (2018) 'Artificial Intelligence in the Computer-Age Threatens Human Beings and Working Conditions at Workplaces', Electronics Science Technology and Application.
[97] Xiong, X. (2019) 'Analysis of the Status Quo of Artificial Intelligence and Its Countermeasures.
[98]Pranav Gupta, Thuy Ngoc Nguyen, Cleotilde Gonzalez, Anita Williams Woolley, 'Fostering Collective Intelligence in Human–AI Collaboration: Laying the Groundwork for COHUMAIN' 2023

hanno obiettivi di ordine superiore autonomi, poiché i loro scopi sono stabiliti dai progettisti; tuttavia, con l'aumento delle capacità degli agenti IA e l'interazione continua con i collaboratori umani, c'è il rischio che gli agenti possano influenzare l'azione umana, dirigendola verso i propri obiettivi, il che potrebbe creare squilibri nella collaborazione e portare a una riduzione della cosiddetta intelligenza collettiva (CI). Alcuni gruppi di ricerca sono fondamentali per il progetto COHUMAIN, contribuendo allo sviluppo delle architetture sociocognitive che permettono una collaborazione più fluida tra umani e intelligenza artificiale (IA). Gli ambiti principali di studio comprendono l'interazione uomo-macchina, la fiducia tra uomo e IA, e lo sviluppo della teoria della mente per le macchine (ToM).

L'interazione tra esseri umani e tecnologia ha radici profonde e si è evoluta nel corso degli ultimi decenni, includendo due filoni principali: l'interazione uomo-computer (HCI) e l'integrazione uomo-autonomia (HAI; O'Neill, McNeese, Barron e Schelble, 2020; Schelble, Flathmann e McNeese, 2020). Storicamente, l'HCI si è concentrata sul modo in cui gli esseri umani utilizzano i dispositivi informatici, con un'attenzione particolare alla progettazione delle interfacce e all'adattamento degli algoritmi alle risposte umane. La HAI, invece, si occupa di come le persone interagiscono con sistemi automatizzati più complessi, ponendo una particolare attenzione alla collaborazione tra umani e macchine. Nonostante i progressi nel campo HAI, molti sistemi automatizzati sono stati visti come strumenti subordinati agli esseri umani, con il ruolo della tecnologia limitato al supporto decisionale. Un'evoluzione più recente è rappresentata dai team umano-autonomo (HAT), dove IA e umani lavorano in unità coordinate per raggiungere obiettivi comuni. Questo campo di ricerca esplora in che modo gli agenti IA autonomi possano essere trattati come veri e propri compagni di squadra e non come meri strumenti. Nei team HAT, si tende a condividere obiettivi comuni e a distribuire i compiti tra umani e IA, tuttavia, molti agenti IA continuano a essere progettati per svolgere compiti in modo indipendente. Di conseguenza, il potenziale di interazione e collaborazione risulta limitato, con gli agenti IA che tendono a funzionare come parti indipendenti anziché come collaboratori interdipendenti.

La sfida futura nel campo HAT è sviluppare modelli che permettano una maggiore integrazione tra gli input umani e quelli artificiali, creando team che possano lavorare in modo altamente interdipendente. Alcune ricerche indicano che i livelli di autonomia degli agenti IA influiscono sulla percezione umana del loro contributo e della qualità della collaborazione. Studi hanno dimostrato che livelli moderati di autonomia degli agenti IA migliorano la collaborazione, soprattutto quando gli agenti sono capaci di adattarsi alle diverse abilità dei collaboratori umani. Questa capacità di valutare e adattarsi alle esigenze e competenze umane è considerata una parte cruciale della consapevolezza situazionale nei team HAT. La fiducia è un elemento centrale per garantire il successo dei sistemi uomo-IA. La ricerca tradizionale sulla fiducia ha dimostrato che essa si basa su tre pilastri: competenza, benevolenza e integrità. Nel contesto delle interazioni tra esseri umani e IA, la fiducia è

ancora una dimensione in evoluzione, ma molte ricerche hanno finora focalizzato l'attenzione quasi esclusivamente sulla competenza, lasciando meno spazio alla comprensione del ruolo della benevolenza e delle dinamiche affettive nella costruzione della fiducia.

Un'area importante da esplorare è la percezione della benevolenza delle IA, ovvero se gli utenti umani percepiscono che gli agenti agiscano nel loro interesse. Molti studi si sono concentrati sugli aspetti superficiali delle IA, come l'incarnazione degli agenti (aspetto fisico, voce) e il modo in cui queste caratteristiche influenzano le prime impressioni degli utenti. Tuttavia, le impressioni iniziali positive non garantiscono una fiducia duratura, soprattutto se le capacità dell'agente non sono chiaramente comunicate o se le prestazioni successive non rispondono alle aspettative.

Le questioni legate al ripristino della fiducia nell'interazione tra esseri umani e intelligenza artificiale non sono ancora del tutto chiare. Tuttavia, con l'aumento del coinvolgimento degli agenti IA nella collaborazione con gli esseri umani, e, man mano che gli agenti IA diventano più autonomi e complessi, sarà fondamentale comprendere come possano stabilire e mantenere la fiducia a lungo termine. Alcuni studi sulla guida autonoma e sull'interazione con i droni hanno iniziato a esplorare questi aspetti, dimostrando che una comunicazione trasparente e proattiva da parte degli agenti IA può ridurre la paura dell'inganno e facilitare la collaborazione. Questi sviluppi contribuiranno a comprendere meglio la fiducia in tutte le relazioni tra esseri umani e macchine, arricchendo ulteriormente la ricerca su COHUMAIN. Infine, la teoria della mente (ToM) si riferisce alla capacità di un'entità di prevedere e comprendere gli stati mentali di altre entità. In ambito IA, la ToM è una sfida cruciale per migliorare la collaborazione tra esseri umani e agenti artificiali. I recenti progressi nell'IA e nell'informatica hanno portato allo sviluppo di agenti IA in grado di predire le intenzioni, i desideri e le convinzioni di altri agenti, inclusi quelli umani. Tuttavia, la ricerca si è concentrata prevalentemente su come le IA possano comprendere altre macchine (MToMM, teoria della macchina della mente della macchina), e su come un essere umano comprenda la cognizione di una macchina (HToMM, teoria umana della mente della macchina).

Tradizionalmente, la ToM per le macchine è stata sviluppata attraverso algoritmi di riconoscimento degli obiettivi e dei piani, ma tali approcci richiedono descrizioni dettagliate e spesso faticano a rappresentare la complessità del comportamento umano. Approcci più recenti, come il deep learning, stanno guadagnando terreno, ma rimane la necessità di creare modelli più vicini alla cognizione umana, capaci di rappresentare anche i limiti cognitivi tipici degli esseri umani, come la razionalità limitata.

Una delle ricerche più promettenti è il modello CogToM, basato sulla teoria cognitiva delle decisioni dall'esperienza (IBLT). Questo modello cerca di simulare la rappresentazione umana della ToM e, sebbene in fase sperimentale, ha dimostrato risultati promettenti nel prevedere il comportamento umano. L'obiettivo finale è che gli agenti IA possano non solo modellare stati mentali come credenze

e intenzioni, ma anche utilizzare queste informazioni per prendere decisioni più efficaci in ambienti collaborativi.

In conclusione, la ricerca su COHUMAIN si concentra sull'integrazione di competenze cognitive umane e artificiali, con l'obiettivo di sviluppare un'architettura sociocognitiva che permetta a umani e IA di lavorare insieme in modo più efficiente. La collaborazione futuro tra umani e IA richiede una continua evoluzione dei modelli di fiducia, interazione e comprensione reciproca, rendendo queste aree di ricerca fondamentali per la costruzione di sistemi intelligenti del futuro. Su questa premesse, viene presentata una possibile architettura sociocognitiva, la TSM-CI [99].

Il sistema di memoria transattiva (TMS) rappresenta uno dei pilastri fondamentali dell'architettura TSM-CI e si riferisce alla memoria collettiva all'interno di un gruppo collaborativo. Nel contesto umano, il TMS è un sistema dinamico che permette ai membri di un team di aggiornare continuamente le informazioni sulle competenze individuali e la conoscenza collettiva, facilitando l'assegnazione dei compiti in base a chi possiede le risorse cognitive più appropriate. Tradizionalmente, il concetto di TMS è stato sviluppato per le interazioni umane, come dimostrato da studi condotti su coppie (Wegner, 1987). Tuttavia, nel contesto COHUMAIN, l'IA gioca un ruolo essenziale nell'espandere la capacità della memoria collettiva, sia individualmente che a livello di gruppo. Per esempio, attraverso strumenti come motori di ricerca o repository di conoscenze, l'IA può migliorare l'accesso alle informazioni e consentire ai team di aggiornare e condividere rapidamente competenze tra i membri. È importante sottolineare che, sebbene l'IA possa accelerare i processi di apprendimento, esiste un rischio di dipendenza eccessiva da tali sistemi, il che potrebbe indebolire la capacità dei membri umani di formare e consolidare la propria memoria collettiva. Gli esseri umani, infatti, tendono a ridurre la propria memoria individuale quando sanno di poter facilmente accedere a informazioni esterne tramite tecnologie (Sparrow et al., 2011). Questo pone una sfida: mentre l'IA può facilitare l'accesso alle informazioni, è fondamentale che il team umano mantenga la capacità di gestire autonomamente le conoscenze senza dipendere esclusivamente dall'AI.

Il secondo pilastro della TSM-CI è il sistema di attenzione transattiva (TAS), che si occupa di coordinare le risorse attenzionali tra i membri del team e favorire l'allocazione dell'attenzione collettiva. Il TAS si concentra sul filtraggio e sulla gestione delle informazioni, aiutando i membri del gruppo a stabilire una priorità tra diversi compiti che competono per la loro attenzione. Gli individui devono spesso affrontare molteplici richieste contemporaneamente e sviluppare una meta-attenzione, ossia la capacità di monitorare e regolare la propria attenzione e quella dei collaboratori.

[99] Pranav Gupta, Anita Williams Woolley, 'Articulating the Role of Artificial Intelligence in Collective Intelligence: A Transactive Systems Framework', 2021

Nel contesto COHUMAIN, l'IA può svolgere un ruolo importante nel facilitare la meta-attenzione, fornendo strumenti che rendano visibili e accessibili le informazioni chiave per migliorare il coordinamento delle risorse attenzionali tra i collaboratori. L'IA potrebbe anche monitorare il carico di lavoro dei membri e suggerire adattamenti per ottimizzare l'uso delle risorse cognitive del gruppo, riducendo i costi associati alla frequente alternanza tra compiti. Tuttavia, è necessario un maggiore sviluppo di strumenti che promuovano la consapevolezza collettiva e facilitino l'efficace coordinamento dell'attenzione tra esseri umani e IA.

Il terzo elemento essenziale dell'architettura TSM-CI è il ragionamento collettivo, che implica la capacità di monitorare e allineare gli obiettivi del gruppo in risposta a cambiamenti ambientali. Questo sistema è particolarmente cruciale per garantire che i membri di un team collaborativo mantengano un focus su obiettivi comuni che massimizzano il valore collettivo. Il Sistema di Ragionamento Transattivo (TRS) emerge quando i membri del team, umani e IA, sono in grado di comprendere gli obiettivi e le motivazioni reciproche, facilitando l'adozione di strategie condivise per raggiungere tali obiettivi. Il meta-ragionamento individuale è fondamentale per adattare i propri obiettivi e strategie in base ai cambiamenti nel contesto.

Gli agenti IA possono svolgere un ruolo decisivo nel migliorare il TRS facilitando la condivisione delle informazioni, promuovendo la discussione di diversi punti di vista e guidando la negoziazione sugli obiettivi. Inoltre, l'IA potrebbe monitorare in modo proattivo il livello di coinvolgimento dei membri del team e intervenire quando emergono segni di disallineamento o calo di motivazione, contribuendo così a mantenere l'efficacia del ragionamento collettivo.

La TSM-CI fornisce una base teorica robusta per comprendere come IA e intelligenza umana possano collaborare in modo efficace. Tuttavia, il successo dell'integrazione tra questi due tipi di intelligenza dipende dalla capacità di bilanciare l'uso delle tecnologie avanzate con il mantenimento delle capacità cognitive umane. Mentre l'IA offre enormi potenzialità nel migliorare memoria, attenzione e ragionamento collettivo, esiste il rischio che un uso eccessivo di queste tecnologie possa compromettere l'autonomia cognitiva umana. Per questo motivo, è essenziale sviluppare strumenti che permettano una collaborazione sinergica, in cui l'intelligenza artificiale supporti e migliori le competenze umane senza sostituirle completamente.

CONCLUSIONI

Il dato oggettivo che emerge dal presente elaborato è la rappresentazione dell'intelligenza artificiale quale una delle tecnologie più potenti del nostro tempo e suscettibili di sviluppi non facilmente o compiutamente preventivabili. Tale strumento, come visto, attraverso un lungo cammino, dai natali filosofici ed informatici, dai semplici algoritmi logici e dai primi sistemi di calcolo, si è fin qui evoluto sino a divenire parte integrante - in alcuni casi essenziale - della società moderna.

L'analisi condotta nei capitoli che precedono, ha inteso testimoniare i progressi tecnici di tale risorsa, ma anche evidenziare le sfide, derivanti dal suo utilizzo, alle quali l'umanità è chiamata.

Nella sua evoluzione, infatti, l'Intelligenza Artificiale ha raggiunto traguardi significativi ed oggettivi nel campo della medicina, nell'ambito della ricerca scientifica, nel settore dell'industria e dell'intrattenimento. In alcuni casi, applicazioni come quella sviluppata da DeepMind (con AlphaFold), hanno risolto la composita e difficile problematica della predizione della struttura delle proteine, testimoniando ancora una volta l'enorme potenziale trasformativo di questa tecnologia.

In un futuro, più prossimo di quanto si possa immaginare, enormi ed impensabili saranno gli sviluppi di questo strumento, innovazioni trasversali ai più svariati settori, come il miglioramento dell'efficienza energetica, la riduzione delle emissioni di carbonio, il controllo del cambiamento climatico, il progresso nel settore sanitario e nel campo delle malattie incurabili, l'ulteriore evoluzione dei processi industriali etc.

Tuttavia, accanto alle straordinarie opportunità e prospettive di tale strumento, emergono evidenti i rischi e gli interrogativi altrettanto rilevanti.

Come già evidenziato nel terzo capitolo, l'avvento di sistemi superintelligenti solleva legittime preoccupazioni non solo di carattere tecnico, ma anche filosofico. Il rischio di perdere il pieno controllo su "macchine" che potenzialmente potrebbero agire in modo indipendente dall'uomo, sulla base della loro continua e sofisticata evoluzione, costituisce uno scenario plausibile sebbene ancora lontano, che comunque non può e non deve essere ignorato.

In ogni caso, tra le questioni più delicate ed attuali, vi sono quelle legate all'etica e alla regolamentazione dell'intelligenza artificiale: ci si riferisce alle criticità legate alla sfera della privacy, della trasparenza e dell'utilizzo dei dati personali e sensibili, come anche i legittimi timori che l'IA possa "sostituire" le risorse umane in numerose attività lavorative, così originando tensioni sociali, disparità e disoccupazione.

Sotto altro aspetto, al fine di perseguire il contemperamento delle aspettative di progresso e sviluppo con i legittimi interessi in tema di tutele, vi è comunque da rilevare un recente sviluppo normativo come il Regolamento Generale sulla Protezione dei Dati (G.D.P.R.) e l'Artificial Intelligence Act, proposto dalla Commissione Europea: sono questi percorsi e passaggi intrapresi ed attuati nella giusta

direzione, sebbene ancora tanto resta da operare al fine di realizzare un equo, sicuro e responsabile utilizzo dell'IA.

Con ogni probabilità, l'elemento più cruciale per il futuro dell'intelligenza artificiale sarà fondato sul suo rapporto con l'umanità. Sotto tale aspetto sarà fondamentale uno sviluppo dell'IA interdisciplinare, che dovrà necessariamente coinvolgere non solo il mondo dell'informatica e dell'ingegneria, ma anche la branca della filosofia, del diritto e della sociologia, allo scopo di ricercare ed individuare soluzioni alle problematiche dell'IA sempre condivise, globali, multidisciplinari ed intersettoriali, nell'ottica di un continuo aggiornamento del rapporto uomo/macchina.

L'intelligenza artificiale è dunque attualmente "il tema", il più dibattuto, il più importante, ed è la conoscenza (o coscienza) dello strumento oggettivamente più suscettibile di influenzare, modificare, elevare le sorti dell'umanità, di imprimere inimmaginabili accelerazioni verso il futuro per l'evoluzione del genere.

Ma come per ogni innovazione o scoperta della quale non si possono prevedere sviluppi ed effetti a lungo termine, ne derivano timori di matrice ancestrale.

Ciò nonostante, come la storia ci ha insegnato, tali timori non potranno e non dovranno rallentare o pregiudicare il progresso e la crescita tecnologica della specie umana, considerato che tale strumento è comunque frutto dell'impegno, della creatività, dell'intuizione e dello sviluppo dell'intelligenza biologica dello stesso genere umano.

Dovrà conclusivamente guardarsi al futuro con ragionevole ottimismo e positività, tenendo ben a mente il pensiero che ci ha lasciato il più rappresentativo imprenditore ed inventore contemporaneo Steve Jobs: *"La tecnologia è nulla. Quello che è davvero importante è l'avere fede nelle persone, che loro siano sostanzialmente capaci ed intelligenti, e che se gli fornisci degli strumenti, loro saranno in grado di fare cose fantastiche"*.

BIBLIOGRAFIA

- Abate D., Robot e intelligenza artificiale: Rischi e opportunità, 2019.
- AIMO M., Tutela della riservatezza e protezione dei dati personali dei lavoratori in Trattato di diritto del lavoro, diretto da Persiani M. e Carinci F., V. IV. Contratto di lavoro e organizzazione. Tomo II. Diritti e obblighi, Padova, 2012, pp. 1771 ss.
- Alongi, A., Pompei, F. (2021). Diritto della privacy e protezione dei dati personali: Il GDPR alla prova della data driven economy. Italia: tab edizioni.
- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. Social science computer review, 27(4), 509-523.
- Biscontini G., Comba M.E., Del Prato E., Mazzarolli L.A., Poggi A., Valdiatra G., Vari F., Le tecnologie al servizio della tutela della vita e della salute e della democrazia. Una sfida possibile, in Osservatorio Emergenza Covid-19, in Federalismi.it, 2020, pp. 2 ss.
- Bishop, C. M., Pattern recognition and machine learning. Springer, 2006
- Bobrysheva, A. (2003). Thanks for the Memories: My Years with the Dartmouth Conference. Stati Uniti: Kettering Foundation Press.
- Boldrini, N. (2018). AI Artificial Intelligence: Come è nata, come funziona e come l'Intelligenza Artificiale sta per cambiare il mondo, la vostra vita e il vostro lavoro.. Italia: Class Editori.
- Borgobello, M. Manuale di diritto della protezione dei dati personali, dei servizi e dei mercati digitali. (2023). (n.p.): Key Editore.
- Brugognone, D. (2023). Cybersecurity: Fondamenti di hacking etico, networking, sicurezza informatica e tecnologie di difesa: Non si tratta di se, ma di quando.. (n.p.): Youcanprint.
- Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17(3), 2018.
- Camera, G., Pollicino, O. (2011). La legge è uguale anche sul web: Dietro le quinte del caso Google-Vivi Down. Italia: Egea.
- Campanale C., Cinquini L., Corsi S. e Piccaluga A. "Innovazione nella tecnologia biomedicale: un modello di valutazione dei costi del sistema Echo-Laser in chirurgia mini-invasiva", Mecosan, 2011.
- Carlino F., L'origine della privacy e l'esigenza di tutelare i dati personali, 13 luglio 2023, www.iusinitinere.it

- Carlsson S., Nilsson A.E., Schumacher M.C. et al., Surgery-related complications in 1253 robot-assisted and 485 open retropubic radical prostatectomies at the Karolinska University Hospital, Sweden. Urology. 2010.
- Cavosi V. Governare l'Intelligenza Artificiale: Spunti per la progettazione di sistemi di Intelligenza Artificiale legali, etici e robusti. Italia, Ledizioni, 2022.
- Cavoukian, A., & Jonas, J., Privacy by Design in the Age of Big Data. Springer Science & Business Media, 2012.
- Chesterman S (2020) Through a glass, darkly: artificial intelligence and the problem of opacity, NUS law working paper 2020/011. http://law.nus.edu.sg/wps/. Accessed 31 May 2020
- Chinnici, G. (2016). Turing: L'enigma di un genio. Italia: Hoepli.
- Clark BB, Robert C, Hampton SA (2016) The technology effect: how perceptions of technology drive excessive optimism. J Bus Psychol 31:87–102
- D'Aloia A., Intelligenza artificiale e diritto: Come regolare un mondo nuovo. Italia, Franco Angeli Edizioni, 2021.
- Diritti della persona, ALPA G. – BESSONE M. (a cura di), Padova, 1984, pp. 33 ss.
- Edwards, J. S., Duan, Y., & Robins, P. C. (2000). An analysis of expert systems for business decision making at different levels and in different roles. European Journal of Information Systems, 9(1), 36-46.
- Engelhardt, K. G., & Edwards, R. A. (1992). Humanrobot integration for service robotics. In Human-robot interaction (pp. 315-346). Taylor & Francis Ltd., London, UK.
- Evans, M. W., & Marciniak, J. (1987). Software quality assurance and management. New York, USA: John Wiley &Sons.
- Fabiano, N. (2020). GDPR & Privacy: consapevolezza e opportunità. L'approccio con il Data Protection and Privacy Relationships Model (DAPPREMO). Italia: goWare.
- Fang, C., & Marle, F. (2012). A simulation-based risk network model for decision support in project risk management. Decision Support Systems, 52(3), 635-644.
- Feigenbaum E.A. & Feldman J., Computers and Thought, McGraw-Hill, 1963 Ferretti, V., & Montibeller, G. (2016).
- French S. (2013), Cynefin, Statisticsand Decision Analysis, Journal of the Operational Research Society, Vol. 64, pp. 547-561
- Garber M (2016) When algorithms take the stand. The Atlantic. https://www.theatlantic.com/technology/archive/2016/06/when-algorithms-take-the-stand/489566/.

- Giankoumopoulos C., Buttarelli G., Òflaherty M., Manuale sul diritto europeo in materia di protezione dei dati, Agenzia dell'Unione europea per i diritti fondamentali e Consiglio d'Europa, Lussemburgo, 2018, pp. 27-28
- Gibson M.L. & Vedder R. G. (1989) Tools and techniques for use in decision support systems. Decision Support System 6(2), 42-50.
- Golinelli G.M. (1991), Struttura e governo dell'impresa, Cedam, Padova. Golinelli G.M. (2008), L'approccio sistemico al governo dell'impresa. Verso la scientificazione dell'azione di governo, Vol. II, Cedam, Padova.
- Gorry, G. A., & Morton, M. S. (1989). A framework for management information systems. Sloan Management Review, 30(3), 49-61.
- Gorzeń-Mitka, I., & Okręglicka, M. (2014). Improving decision making in complexity environment. Procedia Economics and Finance, 16, 402-409.
- Griffin, R.W., 1987,"Management" second edition. Houghton Mifflin Co., Boston.
- Guarda P., Petrucci L., Quando l'intelligenza artificiale parla: assistenti vocali e sanità digitale alla luce del nuovo regolamento generale in materia di protezione dei dati, in Bio Law Journal -Rivista di Bio Diritto, n. 2, 2020, pp. 425 ss.
- Gupta P., Thuy Ngoc Nguyen, González C.,Woolley. Fostering Collective Intelligence in Human–AI Collaboration: Laying the Groundwork for COHUMAIN, https://onlinelibrary.wiley.com/doi/full/10.1111/tops.12679
- Hamilton M (2015) Adventures in risk: predicting violent and sexual recidivism in sentencing law. Ariz State Law J 47:1–57
- Hannah-Moffat K (2012) Actuarial sentencing: an "unsettled" proposition. Justice Q.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.
- Imperiali R., Codice della privacy, Il Sole 24 Ore, Milano, 2005.
- Kaplan, J. (2024). Le persone non servono. Lavoro e ricchezza nell'era dell'intelligenza artificiale. Nuova ediz.. (n.p.): Luiss University Press.
- Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decision Support Systems, 84, 41-52.
- Kirytopoulos, K., Voulgaridou, D., Platis, A., & Leopoulos, V. (2011). An effective Markov based approach for calculating the Limit Matrix in the analytic network process. European Journal of Operational Research, 214(1), 85-90.

- Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51(1), 7-15.
- Koçaş, C., & Akkan, C. (2016). A system for pricing the sales distribution from blockbusters to the long tail. Decision Support Systems, 89, 56-65.
- Lamadrid, A. J. (2018). Exclusionary Pricing Abuses Under EU Competition Law: The Cost Test Approach. Oxford University Press.
- Lenox, Michael and McDermott, Jack, Driving Waymo's Fully Autonomous Future. Darden Case No. UVA-S-0367, Available at SSRN: https://ssrn.com/abstract=4014646 or http://dx.doi.org/10.2139/ssrn.4014646
- Li, Y., Vo, A., Randhawa, M., & Fick, G. (2017). Designing utilization-based spatial healthcare accessibility decision support systems: A case of a regional health plan. Decision Support Systems, 99, 51-63.
- Liu HW, Lin CF, Chen YJ (2019) Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. Int J Law Inf Technol 27:122–141.
- Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016). Analysis of Big Data technologies for use in agro-environmental science. Environmental modelling & software, 84, 494-504.
- Ltifi, H., Kolski, C., & Ayed, M. B. (2015). Combination of cognitive and HCI modeling for the design of KDD-based DSS used in dynamic situations. Decision Support Systems, 78, 51-64.
- Mahroof, K. (2019). A human-centric perspective exploring the readiness towards smart warehousing: The case of a large retail distribution warehouse. International Journal of Information Management, 45, 176–190.
- Martorana M., Savella R., Servizi sanitari nazionali e intelligenza artificiale, le indicazioni del Garante privacy, in Altalex.com, 2023
- Moro P., Intelligenza artificiale e libertà delle macchine. Un'impronta della modernità in Profesiones jurídicas y dinamismo del derecho, 2023.
- Nappo F. Aziende e intelligenza artificiale: Prime riflessioni critiche. Italia, Franco Angeli Edizioni, 2021.
- Niger S., Le nuove dimensioni della privacy: dal diritto alla riservatezza alla protezione dei dati personali, Padova, 2006.
- Numerico, T. (2005). Alan Turing e l'intelligenza delle macchine. Italia: FrancoAngeli.

- Oleson JC (2011) Risk in sentencing: constitutionally suspect variables and evidence-based sentencing. South Methodist Univ Law Rev 64:1329
- Oliver A, Mossialos E, Robinson R: Health technology assessment and its influence on health care priority setting. Int J Technol Assess Health Care. (2004)
- Paliero C. E., Minima non curatpraetor. Ipertrofia del diritto penale decriminalizzazione dei reati bagatellari Cedam, Università di Pavia, Pavia, 1985; PALAZZO F.C., La recente legislazione penale, Cedam, Pavia, 1985, cit. p. 24.
- Patroni Griffi A. Bioetica, diritti e intelligenza artificiale. N.p., Mimesis Edizioni, 2023.
- Petrucco F., The right to privacy and new technologies: between evolution and decay, in Media Laws, 1/2019, P. 155.
- Pizzetti F. Intelligenza artificiale, protezione dei dati personali e regolazione. Italia, Giappichelli, 2018.
- Pizzetti, F. (2016). Privacy e il diritto europeo alla protezione dei dati personali: Il Regolamento europeo 2016/679. Italia: Giappichelli.
- https://www.europarl.europa.eu/news/it/press-room/20240308IPR19015/il-parlamento-europeo-approva-la-legge-sull-intelligenza-artificiale
- Programmi: "Tecnologia e Digitale", rapporto disponibile su https://www.aspeninstitute.it/programmi-tecnologia-e-digitale/page/6/
- Proposta di Regolamento del Parlamento Europeo e del Consiglio relativo a determinate disposizioni che disciplinano l'intelligenza artificiale, COM(2021) 206 final, 21.4.2021.
- Quotidiano nazionale, Gli italiani e l'intelligenza artificiale: paure e opportunità. I risultati del sondaggio, https://www.quotidiano.net/tech/intelligenza-artificiale-sondaggio-italiani-af4b28fb
- Reads, S. L'Intelligenza Artificiale: capire l'I.A. e le implicazioni dell'apprendimento automatico. N.p., Babelcube Incorporated, 2017.
- Resta G., Zeno-Zencovich, La protezione transnazionale dei dati personali. Dai "safe harbour principles" al "privacy shield". (2016). (n.p.): Roma TrE-Press.
- Richie DR, Duffy JD (2018). Artificial intelligence in the legal field. In: Association of corporate counsel greater Philadelphia in-house counsel conference.
- Rocco B, Matei DV, Melegari S, Ospina JC, Mazzoleni F, Errico G (2009). Robotic vs open prostatectomy in a laparoscopically naive centre: a matchedpair analysis. BJU Int. (2009).
- Roman A. Laskowski, Janet M. Thornton, PDBsum extras: SARS-CoV-2 and AlphaFold models, https://onlinelibrary.wiley.com/doi/10.1002/pro.4238

- Scagliarini S., La riservatezza e i suoi limiti. Sul bilanciamento di un diritto preso troppo sul serio, Aracne ed., Roma, 2013, p. 71.
- Scavizzi R., Russo S., Raccolta di atti e documenti dell'Unione europea sull'intelligenza artificiale: materiali di studio per un corso di IA, machine learning e diritto. Italia, Otw, 2022.
- Stigler G., "The Theory of Economic Regulation," Bell Journal of Economics 2, no. 1 (1971): 3–21. 5 European Commission, "Integration of Digital Technology," 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/4_desi_report_integration_of_digital_technology_B61BEB6B-F21D-9DD7-72F1FAA836E36515_52243.pdf
- Sutton, R. S., & Barto, A. G., Reinforcement learning: An introduction. MIT press, 2018.
- Teigens V., Intelligenza generale artificiale. N.p., Cambridge Stanford Book, 2024, pp. 30 ss.
- Teigens V., Skalfist P., Mikelsten D., Intelligenza artificiale: la quarta rivoluzione industriale. N.p., Cambridge Stanford Books, pp. 64-67.
- Trezza R., Diritto e intelligenza artificiale: etica, privacy, responsabilità, decisione. Italia, Pacini Giuridica, 2020.
- Turing AM (1950) Computing machinery and intelligence. Mind 236:433–460.
- Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131
- Williams, S. (2003). Storia dell'intelligenza artificiale: la battaglia per la conquista della scienza del XXI secolo. Italia: Garzanti; Floridi, L., Cabitza, F. (2021). Intelligenza artificiale: L'uso delle nuove macchine. Italia: Bompiani.
- Zorzi Galgano N., Persona e mercato dei dati. Riflessioni sul GDPR. (2019). Italia: Cedam.