Stefania Costantini[*]

Intelligenza Artificiale e Agenti: sfide e opportunità etiche

Abstract
Artificial Intelligence (AI) is now a constant presence in our daily lives. From home automation to social networks, from navigation systems to healthcare, AI-powered technologies are transforming the way we live, work, and interact. But like any powerful tool, AI brings with it opportunities and risks. In this article, we will briefly review the history and main developments of AI, first of all Intelligent Agents, explore the main areas in which AI is making a difference, including healthcare, and discuss the pitfalls to avoid and the crucial role of awareness in the use of intelligent technologies.

Keywords
Artificial Intelligence (AI), Intelligent Agents, Large Language Models, Human-centered AI, Ethics

1. Introduzione

L'intelligenza artificiale (*Artificial Intelligence*, AI) è uno dei campi più affascinanti e controversi dell'informatica moderna. Dalle origini filosofiche con Aristotele all'Era Digitale con la moderna implementazione di reti neurali e agenti intelligenti, l'AI ha percorso un lungo cammino, attraversando utopie, fallimenti e rivoluzioni tecnologiche.

L'idea di replicare l'intelligenza umana ha radici antiche, che risalgono ad Aristotele e a pensatori come Raymond Lullo. Nel corso dei secoli, progetti pionieristici come le macchine di Pascal e Leibniz, e le intuizioni di Charles Babbage e Ada Lovelace, hanno gettato le basi della computazione automatica. Alan Turing, con la sua omonima macchina e il celebre test (formulato nel famosissimo articolo *Computing Machinery and Intelligence* del 1950), ha segnato una svolta fondamentale nella storia dell'AI:

[*] Università dell'Aquila

Filosofia morale/Moral Philosophy, vol. 1, 2025 • ISBN: 9791222322681 • DOI: 10.7413/fmmp0069

> Si consideri una persona che interagisce attraverso una tastiera e uno schermo con due interlocutori nascosti: uno umano e l'altro un computer. Se, dopo un certo tempo, la persona non è in grado di distinguere chiaramente quale dei due interlocutori sia umano e quale sia il computer, allora il computer ha superato il test, dimostrando così una forma di intelligenza indistinguibile da quella umana.[1]

Potremmo chiederci se esiste ad oggi un sistema intelligente che supera il test di Turing: la risposta ad oggi è negativa, anche se vi sono sistemi conversazionali che riescono in buona misura ad 'ingannare' un utente umano.

Il termine 'Intelligenza Artificiale' è stato introdotto da John McCarthy (1927-2011) nel 1956 durante un seminario tenutosi a Dartmouth (USA), riferendosi al ragionamento automatico. La capacità di ragionare logicamente è infatti un aspetto fondamentale dell'intelligenza ed è stata studiata approfonditamente sin da quel periodo. Il *Logic Theorist*, sviluppato tra il 1955 e il 1956, era in grado di dimostrare teoremi, talvolta in modo più elegante rispetto alle fonti originali. Successivamente fu introdotto il *General Problem Solver* (GPS), la cui intelligenza dipendeva però da informazioni precedentemente programmate. Questi sistemi pionieristici di intelligenza artificiale sono i precursori del moderno MATLAB, influenzandone le capacità di risoluzione logica dei problemi, automazione e calcolo simbolico.

Negli anni '70 del secolo scorso, sono stati sviluppati i cosiddetti 'sistemi esperti,' che simulano la competenza umana in domini specifici. Questi sistemi sono costituiti da una base di conoscenza e da un motore inferenziale. Esempi noti sono MYCIN per la diagnosi medica e DENDRAL per l'analisi chimica. Nonostante la loro efficienza, questi sistemi mancano di senso comune e adattabilità, poiché sono 'disincarnati,' cioè, privi di una reale interazione fisica con l'ambiente. I componenti fondamentali di un sistema esperto sono: una base di conoscenza (*Knowledge Base*, KB); un motore inferenziale (*Inference Engine*), che deriva nuove conoscenze dalla KB.

Le informazioni da inserire nella KB si ottengono intervistando esperti nell'ambito specifico di interesse. L'intervistatore, chiamato ingegnere della conoscenza (*knowledge engineer*), organizza le informazioni raccolte dagli esperti in una serie di regole, solitamente strutturate nella forma 'se-allora (*if-then*). Il motore inferenziale consente al sistema esperto di effettuare deduzioni a partire dalle regole presenti nella base di conoscenza.

[1] A. Turing, *Computing Machinery and Intelligence,* "Mind", LIX (236): 433-460, (October 1950) tr. it. in V. Somenzi, R. Cordeschi, *La filosofia degli automi. Origini dell'intelligenza artificiale*, Boringhieri, Torino 1986, pp. 157-183 (tr.it. è nostra); Cfr. A. Turing, *On computable numbers, with an application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 42, n. 1, pp. 230-265 (1936-1937).

Marvin Minsky argomentò, nel saggio *The Society of Mind*[2], sul fatto che il ragionamento di senso comune, ossia la capacità mentale condivisa dalla maggior parte delle persone, è più complesso di molti compiti intellettuali che ricevono maggiore attenzione e considerazione. Questo perché le abilità mentali che definiamo 'esperte' coinvolgono una grande quantità di conoscenza, ma di solito impiegano solo pochi tipi di rappresentazione. Al contrario, il senso comune comprende molteplici tipi di rappresentazione e pertanto richiede una gamma più ampia di competenze diverse.

Si è in seguito compreso che il ragionamento di senso comune non è 'non logico,' ma richiede forme logiche diverse rispetto a quella deduttiva 'classica,' logiche che sono state in seguito sviluppate. Le esigenze sono: gestire le eccezioni; capacità di rivedere le proprie conclusioni; gestire scenari plausibili, e altro ancora. Sono state infatti sviluppati logiche non-monotone, ragionamento abduttivo e ragionamento di default (ragionamento '*de-feasible*'), logiche epistemiche/modali/temporali, logiche incerte e probabilistiche. Con le nuove logiche, è stato possibile costruire sistemi molto più potenti e generali dei sistemi esperti, e in particolare gli agenti intelligenti AI, che sistemi AI autonomi che combinano rappresentazione della conoscenza, ragionamento automatico e apprendimento per interagire con l'ambiente in modo proattivo e reattivo.

Sempre negli anni '70 del secolo scorso, Douglas Bruce Lenat iniziò il progetto CYC, con l'obiettivo di rappresentare in modo esplicito l'intera conoscenza umana. Da questo sforzo nacque il campo di ricerca sulle Ontologie, che ha dato origine a linguaggi specializzati di rappresentazione della conoscenza, dotati di motori inferenziali chiamati '*reasoner*.' Le ontologie sono, oggi, lo standard per la rappresentazione di dati e conoscenza. Un'ontologia è, in particolare, una rappresentazione strutturata e formale della conoscenza su un determinato dominio, che definisce chiaramente i concetti, le relazioni e le regole che lo caratterizzano. Serve a rendere comprensibile e condivisibile l'informazione tra sistemi diversi, migliorando l'interoperabilità e la comunicazione tra macchine e persone.

L'AI si è sviluppata lungo due filoni principali: simbolico e sub-simbolico. Il primo si basa su logica, regole e sistemi esperti; il secondo su reti neurali artificiali e apprendimento automatico. Le reti neurali sono sistemi informatici che si ispirano al cervello umano, capaci di imparare dai dati e riconoscere immagini, suoni e testi. Il '*Deep Learning*' è una tecnologia che utilizza reti neurali molto complesse per consentire ai computer di apprendere autonomamente, ad esempio riconoscendo

[2] Cfr. M. Minsky, *The Society of Mind,* Simon & Schuster, New York 1986, tr.it. G. Longo, *La società della mente*, Adelphi, Milano 1989.

immagini, identificando volti nelle foto o traducendo lingue diverse, grazie a tecniche come la retropropagazione e l'addestramento supervisionato. I modelli di *deep learning* (DL), basati su grandi reti neurali, come quelli alla base dei modelli linguistici (*Large Language Models*) come ChatGPT che sono in grado di decodificare e generare testo, hanno portato a risultati straordinari in molti campi. Hanno purtroppo bisogno, per la fase di 'addestramento,' di grandi quantità di dati opportunamente scelti e preparati, che non sono disponibili in tutti i campi. Se i dati a disposizione sono pochi, i sistemi DL risultano poco precisi o addirittura danno risultati errati al di fuori del 'dataset' di addestramento.

Le reti neurali sono spesso descritte come sistemi '*black-box*' (scatole nere) perché, nonostante la loro efficacia nel produrre risultati accurati, la logica interna con cui arrivano alle decisioni non è facilmente interpretabile dagli esseri umani. Ciò rappresenta un limite significativo, soprattutto in ambiti critici come la sanità o la finanza, dove la trasparenza decisionale è cruciale per assicurare fiducia, responsabilità e comprensione delle scelte effettuate dai sistemi AI. Per questo motivo, la ricerca attuale sta cercando di elaborare tecniche per migliorare l'interpretabilità e la spiegabilità delle reti neurali.

La prospettiva dell'AI neuro-simbolica (*Neuro-Symbolic AI*) mira a integrare il ragionamento simbolico con la potenza delle reti neurali, per ottenere sistemi spiegabili, affidabili e capaci di generalizzazione. Il premio Nobel per l'economia Daniel Kahneman e i vincitori del Premio Turing 2018, nonché pionieri del deep learning, Geoffrey Hinton, Yoshua Bengio e Yann LeCun, hanno indicato l'AI neuro-simbolica come la nuova prospettive per il futuro dell'intelligenza artificiale. Essi hanno sostenuto che la realizzazione di un sistema AI ricco, semanticamente solido, spiegabile e affidabile, richiederà necessariamente uno strato di ragionamento solido in combinazione con il *deep learning*.

2. Gli Agenti Intelligenti

2.1. Cosa sono e a cosa servono

Cosa sono esattamente gli Agenti, e in particolare gli Agenti Intelligenti, di cui si sente molto parlare? Gli agenti sono moduli software con una particolare caratteristica: sono autonomi, ossia, una volta attivati, a meno che non vengano fermati, continuano a funzionare portando avanti le loro attività. Come il nome suggerisce, sono capaci di agire sul loro ambiente, mediante opportuni 'attuatori,' e dunque devono innanzitutto essere in grado di percepire l'ambiente stesso, mediante 'sensori.' Gli

agenti sono 'situati' in un ambiente, che può essere puramente software oppure fisico. Gli agenti possono costituire la mente pensante dei robot, ossia disporre di un corpo fisico.

Le capacità principali di un agente sono: Reattività, ossia la capacità di reagire opportunamente ad eventi esterni (cioè, provenienti dall'ambiente in cui l'agente è situato) mettendo in atto una o più azioni; Proattività, ossia la capacità di intraprendere attività ed effettuare azioni per perseguire i propri obiettivi.

Infatti, gli agenti possono essere programmati per avere intenzioni, perseguire obiettivi ed eseguire compiti. Per perseguire i propri obiettivi, gli agenti devono essere capaci di costruire un piano (quindi devono avere capacità di pianificazione) ed eseguirlo, e, nel caso, riadattarlo in caso di problemi. La capacità di pianificare è indubbiamente una componente fondamentale dell'intelligenza, e presuppone la capacità di costruire un modello interno del mondo esterno (o almeno del frammento di mondo a cui si è al momento interessati) e delle azioni su di esso possibili da parte dell'agente, date le sue capacità e le risorse disponibili. Sulla base di questa descrizione, un processo di pianificazione identifica, per un dato obiettivo, una o più sequenze di azioni che possano condurre dallo stato attuale del mondo ad uno stato in cui l'obiettivo sarà stato raggiunto.

Gli agenti possono costituire 'Sistemi Multi-Agente' (*Multi-Agent Systems*, o MAS). Gli agenti componenti possono essere cooperativi, ed in tal caso eventualmente possono perseguire intenzioni e obiettivi condivisi. Agenti competitivi possono invece eventualmente negoziare per suddividere fra loro le risorse disponibili. Per far parte di MAS, gli agenti devono possedere 'abilità sociali,' ossia devono essere in grado di comunicare. A tale scopo sono stati sviluppati i cosiddetti *Agent Communication Languages* (ACL) che prevedono vari tipi di messaggi, con la propria sintassi e semantica. Ad esempio, un agente può informare un altro agente in merito a qualcosa, o può avanzare una richiesta, o può accettare una richiesta e inviare una risposta, ecc.

Una delle caratteristiche più promettenti degli agenti intelligenti AI è la capacità di affiancare l'essere umano nei compiti quotidiani, automatizzando processi ripetitivi e offrendo assistenza intelligente. L'AI sta infatti rivoluzionando il sistema produttivo, abilitando la manutenzione predittiva, il controllo qualità, l'automazione dei progetti e l'uso di modelli linguistici nei processi documentali.

L'uso di agenti intelligenti e sistemi multi-agente può offrire, già oggi o in prospettiva, numerosi vantaggi agli utenti umani in diversi contesti. Ecco alcuni dei potenziali benefici:

Assistenza personale. Un agente intelligente può assistere l'utente umano nei suoi compiti quotidiani fornendo compagnia, supporto e si-

curezza. Esempio: Un agente personale può monitorare il 'suo' utente durante lo svolgimento di compiti tediosi o rischiosi come guida dell'auto; rilevando (anche mediane opportuni dispositivi indossabili, come ad esempio uno 'smartwatch,' condizioni di ansia o stanchezza, può fornire rassicurazione e consigli. Esempio: Un agente personale può monitorare il 'suo' utente anziano con problemi di vuoti di memoria o lieve demenza, fornendo compagnia e informazioni, ricordandogli i compiti necessari (come prendere le medicine), e sollecitandolo a svolgere attività utili (come una passeggiata).

Esempio: Un agente personale può monitorare il 'suo' utente malato o disabile monitorando i parametri vitali, e restando in costante contatto con i familiari, i medici e l'ospedale, per fornire informazioni aggiornate e se necessario ottenere supporto o soccorso.

Accessibilità e assistenza inclusiva. Gli agenti AI possono migliorare l'accessibilità per persone malate o con disabilità, fornendo supporto vocale, suggerimenti, traduzioni automatiche e interfacce intuitive. Esempio: Un assistente vocale può aiutare persone con disabilità motorie a controllare dispositivi smart home.

In ambito sanitario, gli agenti (neuro-simbolici) possono supportare medici e pazienti, offrendo diagnosi precoci o assistenza alle persone con disabilità. I robot sociali (ossia agenti intelligenti dotati di un corpo fisico) vengono impiegati con successo nell'educazione e, ad esempio, nella terapia di bambini neurodivergenti, facilitando la comunicazione e l'apprendimento.

2.2. Agenti basati su LLM

Gli LLM (*Large Language Models*, come GPT, LLaMa, Claude, Minerva, ecc.), o 'modelli linguistici,' sono algoritmi avanzati di apprendimento automatico in grado di svolgere un'ampia gamma di compiti legati all'elaborazione del linguaggio naturale (NLP). Sono capaci di decodificare e generare il linguaggio umano, grazie all'addestramento di grandi reti neurali ottenuto con una grande quantità di dati, che consente di fare una previsione della parola o del simbolo successivo a quello già generato.

Secondo alcune recenti proposte, gli LLM rappresenterebbero una base adatta per gli agenti, in grado di supportare tre componenti principali: ragionamento, percezione e azione. Tuttavia, l'assenza di una memoria a lungo termine e la mancanza di pianificazione esplicita negli LLM costituiscono dei limiti fondamentali per il loro utilizzo come agenti autonomi, rendendo necessaria l'integrazione con altri sistemi per ottenere capacità decisionali più sofisticate. Ad esempio, per migliorare le capaci-

tà di pianificazione, gli LLM possono essere integrati con sistemi esterni, inclusi sistemi basati su regole o ragionamento simbolico.

Relativamente alla comunicazione nei MAS, gli agenti basati su LLM possono interagire tra loro in modo naturale, ma manca completamente la semantica della comunicazione.

L'integrazione tra agenti logici e agenti LLM punta ad aumentare la flessibilità e l'adattabilità: percezione dell'ambiente in tempo reale, autonomia decisionale, accesso a informazioni dinamiche e contestualizzazione dell'interazione. Gli agenti logici mantengono il controllo decisionale, mentre i LLM arricchiscono la comprensione contestuale.

3. Human-centered AI

L'interazione tra uomo e automazione è il tema principale dell'IA centrata sull'uomo (*Human-centered AI*).

Inoltre, il concetto di *Human-AI Teaming* enfatizza la collaborazione sinergica tra esseri umani e agenti artificiali. Per rendere questa collaborazione efficace, è necessario che gli agenti AI siano capaci di empatia, comprensione del contesto emotivo e delle intenzioni, come suggerito dalla *Joint Intention Theory* (JIT).

Questa tematica rientra anche nell'ambito della AI affidabile (*Trustworthy AI*), i cui requisiti includono:

– rispetto dell'autonomia umana,
– prevenzione dei danni,
– equità e spiegabilità,
– utilizzo responsabile dell'AI, il cui scopo è quello di impiegarla in modo sicuro, etico e affidabile.
– Collaborazione tra esseri umani e agenti.

Quando gli esseri umani e gli agenti AI collaborano in un team (*Human-AI Teaming*, HAIT), possono, infatti, ottenere risultati superiori a quelli che ciascuno potrebbe raggiungere singolarmente, in quanto sono in grado di controllarsi e migliorarsi a vicenda.

Ad esempio, un conducente umano potrebbe gestire meglio pericoli contingenti lavorando in sinergia con l'agente AI installato sul veicolo. Allo stesso tempo, l'agente può assistere il conducente nelle situazioni difficili e di stress, e al contempo imparare dall'utente stesso come gestire situazioni particolari.

In questa relazione sinergica, gli esseri umani possono migliorare l'efficacia e le capacità dell'automazione, mentre l'automazione può potenziare le prestazioni umane e compensare eventuali limiti, correggendo errori

di comportamento che potrebbero derivare da stanchezza, stress o altre condizioni fisiche o emotive. Inoltre, gli agenti possono fornire suggerimenti utili per ottimizzare la sicurezza e l'efficacia delle decisioni umane.

Per adottare agenti in compiti cruciali, come ad esempio: il miglioramento dell'assistenza medica, l'innovazione nell'insegnamento, la costruzione di team efficaci tra esseri umani e AI, gli agenti dovrebbero essere dotati di moduli avanzati di riconoscimento e gestione delle emozioni, capaci di empatia e di simulare alcuni aspetti della Teoria della Mente (*Theory of Mind*, ToM).

Questo vuol dire che gli agenti dovrebbero essere in grado di ricostruire ciò che una persona sta pensando o provando, in modo da adattare il proprio comportamento di conseguenza. La modellazione della Teoria della Mente si basa spesso su tecniche di '*Affective Computing*,' ossia un insieme di metodi volti a rilevare lo stato emotivo umano attraverso segnali fisici (come espressioni facciali, tono della voce o battito cardiaco) per permettere all'agente di rispondere in modo intelligente al feedback emotivo umano. Inoltre, gli agenti dovrebbero adattarsi ai loro utenti umani, anche in termini delle loro preferenze etiche.

4. AI nell'assistenza sanitaria

4.1. AI nell'interpretazione delle immagini mediche

Uno dei progressi più rilevanti nel campo dell'Intelligenza Artificiale riguarda la Computer Vision, ovvero l'analisi automatizzata di immagini e video mediante deep learning, con importanti applicazioni nell'imaging medico. In radiologia, l'AI ha dimostrato grande efficacia nell'identificare anomalie su radiografie, scansioni TC e RM, contribuendo a diagnosi più accurate e veloci. L'uso di queste tecnologie non si limita ai radiologi, ma si sta espandendo anche ad altri operatori sanitari, migliorando l'accessibilità e riducendo gli errori diagnostici in ambienti con risorse limitate o nelle emergenze.

Tuttavia, i modelli di DL soffrono spesso di una scarsa generalizzabilità, mostrando prestazioni inferiori quando applicati a pazienti differenti rispetto a quelli utilizzati in fase di addestramento dei modelli stessi. Per un'applicazione efficace su larga scala, è essenziale sviluppare metodi di validazione trasparenti che ne assicurino affidabilità e generalizzazione. La migliore performance si ottiene anche in questo caso tramite *Human-AI Teaming*, ossia tramite una stretta collaborazione tra professionisti sanitari e AI. Comunque, l'interpretabilità dei modelli rimane un problema centrale, poiché i clinici sono riluttanti a fidarsi di sistemi percepiti come "*black-box*" ("scatola nera"), di cui non possono analizzare i ragionamenti dietro gli output.

Innovazioni recenti permettono ai sistemi DL di riconoscere patologie mai viste durante l'addestramento, migliorando significativamente la flessibilità dei modelli. Altre promettenti frontiere includono i modelli multimodali, che combinano immagini radiologiche con dati clinici testuali, e modelli auto-supervisionati, che apprendono senza bisogno di indicazioni esplicite.

I futuri modelli di AI potranno generare spiegazioni dettagliate in linguaggio naturale, annotazioni vocali e immagini contestualizzate, rispecchiando un ragionamento medico sofisticato. Questa evoluzione garantirà ai medici un supporto di alta qualità e personalizzato durante la pratica clinica.

Dunque, l'AI ha il potenziale per rivoluzionare l'assistenza sanitaria, migliorando la tempestività delle valutazioni, la gestione delle patologie e la pianificazione terapeutica. Nonostante ciò, rimane imprescindibile il ruolo del medico umano, poiché l'AI deve essere considerata uno strumento complementare e collaborativo, soprattutto per motivi di responsabilità e affidabilità.

4.2. Agenti e robot assistivi nell'assistenza ai pazienti

La fornitura di servizi sanitari di alta qualità in maniera economicamente sostenibile è una questione cruciale in tutti i paesi, soprattutto a causa dell'invecchiamento della popolazione, della ricomparsa di malattie considerate estinte e della comparsa di nuove problematiche impegnative, come le pandemie di Ebola e Covid-19. I sistemi sanitari intelligenti rappresentano una valida soluzione per affrontare efficacemente le sfide della sanità moderna, rispondendo alle esigenze di pazienti, medici, personale sanitario e famiglie. Questi sistemi possono garantire assistenza continua, monitoraggio costante, compagnia e supporto anche a coloro che non avrebbero la possibilità di accedere a caregiver umani per periodi prolungati – una risorsa, peraltro, sempre meno disponibile a causa del calo demografico e dell'invecchiamento della popolazione. L'intelligenza artificiale può dunque contribuire concretamente a rendere l'assistenza sanitaria di alta qualità accessibile a tutti, promuovendo così equità e giustizia sociale, e non soltanto ai pochi che possono permettersela.

La tutela dell'autonomia personale e il miglioramento della qualità della vita sono obiettivi centrali per tutte quelle persone – non solo anziane – che affrontano quotidianamente limitazioni fisiche o cognitive. Condizioni croniche di salute, disabilità o problemi di memoria possono rendere estremamente complesso lo svolgimento delle normali attività giornaliere. In questo scenario, i robot assistivi sociali (*Socially Assistive Robots*, SAR), basati su agenti intelligenti, rappresentano una soluzione tecnologica promettente. Essi possono supportare concretamente gli

utenti nella gestione delle attività quotidiane e della propria salute, migliorando il benessere complessivo. L'impiego di agenti intelligenti AI all'interno dei SAR può ridurre la dipendenza degli utenti dai caregiver umani, stimolare le funzioni cognitive e favorire una maggiore partecipazione sociale.

È stato sperimentalmente dimostrato che i robot sociali assistivi possono favorire l'interazione degli individui con l'ambiente circostante e possono fungere da stimolo per la conversazione, incoraggiando scambi significativi tra gli utenti e i loro pari o caregiver. I SAR hanno dimostrato il potenziale di facilitare interazioni positive e ridurre lo stress in vari contesti sanitari. Queste interazioni, percepite come empatiche, favoriscono un senso di connessione e migliorano il benessere dell'utente. Fra i primi esempi, Il robot assistivo Alice è stato sviluppato per contrastare l'"epidemia di solitudine" tra gli anziani. Il film *Alice Cares* (diretto da Sander Burger nel 2015) documenta un progetto pilota in cui il robot umanoide da compagnia viene affidato a tre donne anziane sole, residenti in strutture assistenziali.

L'empatia nei sistemi robotici deve andare oltre la semplice assistenza funzionale: implica la capacità di rispondere ai bisogni emotivi e psicologici degli utenti. Affinché un robot venga percepito come empatico, deve esibire comportamenti personalizzati che riflettano la comprensione delle circostanze, delle preferenze e dei problemi specifici dell'utente. La personalizzazione risulta quindi cruciale, poiché permette al robot di adattare le interazioni, modificando stile comunicativo, tono e risposte sulla base del profilo individuale del paziente. Dunque, i robot possono promuovere interazioni più profonde e significative, migliorando così l'esperienza complessiva e i risultati terapeutici, se integrano informazioni dettagliate sui pazienti, come storia medica, contesto sociale e preferenze personali.

Un passo per raggiungere questi obiettivi e sviluppare agenti o robot personalizzati è lo sviluppo delle cosiddette *Blueprint Personas*, uno strumento mirato a promuovere un'assistenza centrata sulla persona. Tale approccio, inizialmente introdotto nel documento della Commissione Europea *Blueprint on Digital Transformation of Health and Care for the Aging Society*" e poi ulteriormente sviluppato dai ricercatori, identifica profili di pazienti basati su diverse necessità, ambienti (in particolare domestici), e su una varietà di caratteristiche sanitarie e socioeconomiche. Inoltre, tiene conto dei potenziali benefici che le risorse digitali possono offrire ai pazienti e ad altri soggetti interessati (*stakeholders*), come ricercatori, operatori sanitari e "caregiver" formali e informali. L'adozione delle *Personas* nel settore della sanità digitale consente una migliore comprensione delle esigenze dei pazienti, migliorando la personalizzazione dei servizi e l'accesso alle tecnologie digitali per la salute.

In concomitanza con lo sviluppo delle *Personas*, è opportuno realizzare delle ontologie per codificare la conoscenza strutturata necessaria al sistema. Grazie a queste ontologie, sarà possibile creare soluzioni di assistenza personalizzate e adattive, capaci di rappresentare la complessità delle interazioni tra pazienti, caregiver e sistemi di intelligenza artificiale. In questo modo, il sistema potrà comprendere al meglio le esigenze degli utenti, garantire una comunicazione affidabile e rafforzare la fiducia nelle soluzioni sanitarie basate sull'AI.

L'obiettivo di progetti attualmente in corso[3] è implementare un agente intelligente basato anche sulle *Blueprint Personas* e sulle ontologie che, analizzando i dati raccolti, sia in grado di riconoscere lo stato di benessere o malessere del paziente sia a breve che a lungo termine. L'agente funge da assistente personale del paziente, che lo aiuta nella supervisione del proprio stato di salute. Tuttavia, questo agente potrà assistere il paziente anche nella vita quotidiana, ricordandogli, ad esempio, non soltanto di assumere i farmaci prescritti, ma anche altre attività da svolgere, supportandolo nello svolgimento delle attività giornaliere[4].

L'agente sarà comunque capace di rilevare sintomi e valutare la gravità della situazione e, se necessario, avvisare un medico o chiamare direttamente i servizi di emergenza, basandosi su dispositivi hardware che stanno diventando gradualmente meno costosi, inclusi quelli indossabili per il rilevamento dei dati vitali (come pressione sanguigna, battito cardiaco, febbre, ecc.). Infatti, i prototipi di sistema finora implementati utilizzano proprio alcuni dispositivi indossabili economicamente accessibili. L'agente può raccogliere in un database i parametri vitali, in parte raccolti automaticamente dai dispositivi medici e in parte inseriti autonomamente dal paziente[5].

Questi sviluppi rappresentano il primo passo verso sistemi capaci di garantire servizi sanitari di alta qualità e sostenibili, affrontando contestualmente anche le sfide etiche, dato che il comportamento etico deve essere una caratteristica intrinseca di tali sistemi. Questa tipologia di sistema potrebbe costituire una soluzione a lungo termine per la gestione quotidiana di una popolazione che invecchia, anche in circostanze drammatiche e urgenti come quelle di una pandemia.

[3] Cfr. A. Monaldini, A. Vozna, S.Costantini, *Blueprint Personas in Digital Health Transformation. Workshop HC@AIxIA 2024*, pagine 40-49.
[4] Cfr. L. De Lauretis, F. Persia, S. Costantini, D. D'Auria, *How to leverage intelligent agents and complex event processing to improve patient monitoring*. "Journal of Logic and Computation", volume 33, n. 4, 2023, pp. 900-935.
[5] Cfr. L. De Lauretis, F. Persia, S. Costantini, *An Intelligent Ecosystem to improve Patient Monitoring using Wearables and Artificial Intelligence*, Conferenza CILC, 2022, pp. 141153.

4.3. Pianificazione dell'uso delle risorse

L'utilizzo di ASP (*Answer Set Programming*), piattaforma altamente efficiente e flessibile per la pianificazione automatica, consente di sviluppare modelli di pianificazione sofisticati in ambito sanitario, bilanciando molteplici obiettivi, vincoli e preferenze provenienti da diverse prospettive:

– dal punto di vista dei pazienti: tempo disponibile, risorse economiche, disponibilità di caregiver, mezzi di trasporto accessibili e adeguati allo stato fisico e cognitivo del paziente, preferenze personali relative agli appuntamenti o ai trattamenti sanitari;
– dal punto di vista del sistema sanitario: budget a disposizione, specializzazioni mediche disponibili, programmazione delle tempistiche (giornaliere, settimanali e mensili) per visite ed esami, e ottimizzazione della composizione dei gruppi di pazienti specifici per determinati studi clinici o trattamenti.

Il modulo di pianificazione basato su ASP è progettato per affrontare scenari sanitari complessi, integrando fattori cruciali come l'urgenza clinica, la posizione geografica del paziente, l'accessibilità delle strutture sanitarie e ulteriori vincoli pratici. In particolare, un modello ASP può includere:

– Fatti: informazioni specifiche sui pazienti, cliniche disponibili, orari e modalità operative;
– Vincoli: limiti dettati da budget, priorità cliniche, accessibilità fisica e logistica;
– Obiettivi di ottimizzazione: riduzione dei tempi di attesa, minimizzazione delle distanze fisiche da percorrere e del disagio sensoriale e cognitivo per i pazienti.

Un ulteriore livello di personalizzazione è introdotto dall'integrazione delle cosiddette *Blueprint Personas*, profili dettagliati relativi sia ai pazienti che ai medici. Questa innovativa metodologia permette di assegnare ogni paziente al medico più appropriato in base alle sue specifiche necessità, competenze mediche richieste e caratteristiche individuali, assicurando contemporaneamente che i medici ricevano pazienti per i quali hanno una preparazione specifica e appropriata.

L'approccio basato su ASP, attualmente in fase di sviluppo, ha il potenziale per generare numerosi benefici nel campo della pianificazione sanitaria, quali:

– Pianificazione ottimizzata dei servizi sanitari;
– Gestione più efficiente e razionale delle risorse;

– Adattabilità in tempo reale ai cambiamenti e alle esigenze emergenti;
– Maggiore soddisfazione e migliore esperienza del paziente;
– Rispetto rigoroso dei diritti umani, delle preferenze e della dignità dei pazienti.

5. Problemi Etici

Sebbene l'intelligenza artificiale (AI) presenti numerosi vantaggi e offra opportunità innovative in molteplici set-tori, essa comporta anche rischi significativi che non possono essere trascurati. Un uso improprio o malevolo dei sistemi di AI potrebbe infatti portare a gravi conseguenze, tra cui la manipolazione dell'opinione pubblica, la diffusione massiva di disinformazione e la sorveglianza invasiva e pervasiva della popolazione. Un esempio rilevante riguarda gli algoritmi impiegati dai social network, che possono inconsapevolmente influenzare e alterare le nostre opinioni, comportamenti e decisioni personali.

L'AI può rappresentare una seria minaccia per la libertà individuale, soprattutto se impiegata per profilare, controllare o isolare le persone sulla base di criteri non trasparenti e non democraticamente condivisi. Pertanto, è fondamentale impegnarsi nella progettazione e nello sviluppo di tecnologie etiche, trasparenti e responsabili, che rispettino pienamente i diritti fondamentali e i valori umani.

Con la crescita delle capacità e della diffusione dell'AI, emergono nuove criticità e aumenta la complessità dei rischi connessi: dalla potenziale perdita di controllo sui sistemi autonomi all'uso improprio o fraudolento dei dati personali, fino alla sofisticata manipolazione delle informazioni e della percezione pubblica.

L'etica dell'AI richiede dunque un impegno costante verso la trasparenza, la spiegabilità dei modelli decisionali, il rispetto rigoroso dei diritti umani e una chiara attribuzione della responsabilità ("*accountability*") per le azioni e le decisioni prese dai sistemi intelligenti. A tale scopo, l'Unione Europea ha recentemente introdotto regolamenti specifici, culminati nell'AI Act, con l'obiettivo di garantire affidabilità, sicurezza e trasparenza dei sistemi di intelligenza artificiale. Anche l'UNESCO ha definito standard globali che mirano a una governance etica e sostenibile dell'AI.

Come sottolineato chiaramente da Costantini nel suo lavoro scientifico[6] e approfondito da Russell[7], è indispensabile sviluppare, perfezionare,

[6] Cfr. S. Costantini, *Ensuring trustworthy and ethical behaviour in intelligent logical agents*, "Journal of Logic and Computation", vol. 32, Issue 2, March 2022, pp. 443-478.
[7] S. Russel, *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin Putnam Inc., New York 2020.

implementare e integrare metodi formali e rigorosi per progettare e certificare agenti intelligenti. Tali metodi sono fondamentali per assicurare comportamenti trasparenti, comprensibili, affidabili ed eticamente corretti, poiché i sistemi basati su agenti AI sono e saranno sempre più frequentemente impiegati in applicazioni critiche, che influenzano direttamente la vita, il benessere delle persone e funzioni sociali essenziali.

Bilanciare l'automazione con l'intervento umano diviene cruciale per garantire che le decisioni e i processi più delicati rimangano sempre sotto supervisione umana, minimizzando il rischio di errori o azioni dannose. L'integrazione di controlli umani nei punti chiave del processo decisionale è dunque imprescindibile. Ad esempio, nel settore sanitario, è fondamentale che un agente AI sottoponga le proprie diagnosi e raccomandazioni cliniche alla revisione di un medico umano prima di prendere decisioni autonome, al fine di evitare errori con potenziali conseguenze irreversibili.

Un ulteriore punto critico riguarda il possibile impatto negativo derivante da un uso eccessivo e incontrollato dell'AI, specialmente tra i più giovani. Un eccessivo affidamento all'intelligenza artificiale può ostacolare lo sviluppo di competenze cognitive fondamentali quali il pensiero critico, la capacità di risolvere problemi autonomamente e la creatività. Delegare troppe attività all'AI potrebbe infatti indebolire la nostra autonomia e capacità di pensiero indipendente, rendendoci progressivamente più dipendenti dalla tecnologia.

Diventa pertanto essenziale trovare il giusto equilibrio tra l'utilizzo di strumenti basati su AI e il rafforzamento delle abilità cognitive e decisionali degli esseri umani. L'intelligenza artificiale dovrebbe rappresentare uno strumento che amplifica e arricchisce le capacità umane, piuttosto che sostituirle o indebolirle. Solo adottando questa prospettiva potremo assicurare che l'AI contribuisca davvero al progresso umano e sociale.

6. Discussione e conclusione

L'Intelligenza Artificiale, da sogno filosofico, è diventata una realtà concreta e complessa, in grado di influenzare significativamente la società. Tuttavia, ogni progresso tecnologico implica responsabilità: trasformare l'AI da potenziale minaccia a reale opportunità richiede un impegno coordinato tra ricerca, normative, principi etici e sensibilizzazione pubblica.

L'introduzione di agenti intelligenti e sistemi multiagente può aumentare notevolmente efficienza, sicurezza, personalizzazione e innovazione in numerosi settori. Nel campo sanitario, ad esempio, permette di offrire trattamenti personalizzati e un monitoraggio costante anche a pazienti che tradizionalmente non ne avrebbero accesso.

Ciononostante, il rapido incremento dell'autonomia degli agenti AI introduce inevitabilmente rischi che vanno attentamente gestiti. È pertanto essenziale bilanciare l'automazione con una supervisione umana attenta e costante, assicurando così un utilizzo responsabile e sicuro delle nuove tecnologie.

L'approccio *Human-Centered Artificial Intelligence* (HCAI) offre un'importante strategia progettuale, ponendo l'essere umano al centro dello sviluppo tecnologico. Preservare e valorizzare le competenze umane, riconoscere i limiti delle tecnologie e creare sistemi AI che amplifichino, anziché sostituire, le capacità umane costituisce una via essenziale per un futuro sostenibile.

In definitiva, l'attuale impegno nella ricerca su supervisione umana, controlli tecnici e protocolli rigorosi di verifica e certificazione rappresenta la chiave per ridurre i rischi connessi all'impiego degli agenti autonomi. Solo attraverso un equilibrio attentamente mantenuto tra automazione e sicurezza, sarà possibile assicurare l'integrazione efficace e sicura dei sistemi intelligenti nelle applicazioni pratiche e nella società del futuro.

Bibliografia

C. Accoto, *Il mondo dato. Cinque brevi lezioni di filosofia digitale*, Egea, Milano 2017

A. Fabris, Etica delle nuove tecnologie, La Scuola, Brescia 2012.

L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Raffaello Cortina Editore, Milano 2022.

L. De Lauretis, F. Persia, S. Costantini, D. D'Auria, *How to leverage intelligent agents and complex event processing to improve patient monitoring*. "Journal of Logic and Compution", volume 33, n. 4, 2023, pp. 900-935.

L. De Lauretis, F. Persia, S. Costantini, *An Intelligent Ecosystem to improve Patient Monitoring using Wearables and Artificial Intelligence*, Conferenza CILC, 2022, pp. 141153.

S. Costantini, *Ensuring trustworthy and ethical behaviour in intelligent logical agents*, "Journal of Logic and Computation", vol. 32, Issue 2, March 2022, pp. 443-478.

J. McCarthy, *The Inversion of Fuctions Defined by Turing Machines*, in C.E. Shannon, J. McCarthy (eds.), *Automata Studies*, Princeton University Press, Princeton 1956

M. Minsky, *The Society of Mind,* Simon & Schuster, New York 1986, tr.it. G. Longo, *La società della mente*, Adelphi, Milano 1989.

A. Monaldini, A. Vozna, S.Costantini, *Blueprint Personas in Digital Health Transformation. Workshop HC@AIxIA 2024*, pagine 40-49.

A. Turing, *Computing Machinery and Intelligence,* "Mind", LIX (236): 433-460, (October 1950) tr. it. in V. Somenzi, R. Cordeschi, *La filosofia degli automi. Origini dell'intelligenza artificiale*, Boringhieri, Torino 1986, pp. 157-183

Turing, A., *On computable numbers, with an application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 42 (1937),
A. Turing, *On computable numbers, with an application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, 42, n. 1, pp. 230-265 (1936-1937).
M.B. Saponaro, *La traduzione algoritmica del pensiero relazionale*, "TEORIA" 2020/2, pp. 187-205.
S. Russel, *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin Putnam Inc., New York 2020.