AI for Good: un approccio sostenibile all'intelligenza artificiale

Tiziana Catarci, Daniel Raffini

1. Intelligenza artificiale: storia e tipologie

L'intelligenza artificiale (IA) nasce ufficialmente come disciplina interna al campo di studi dell'Informatica nel 1956, in occasione della Conferenza di Dartmouth. In realtà già dagli anni Quaranta abbiamo alcune prime teorizzazioni riconducibili all'idea di IA. Nel 1943, McCulloch e Pitts propongono un modello teorico di rete neurale artificiale, che sarà concretizzato nel 1951 da Minsky e Edmonds, i quali creano il primo computer basato su reti neurali, chiamato SNARC. Nel 1950, Alan Turing pubblica l'articolo *Computer Machinery and Intelligence*, nel quale propone l'idea di un calcolatore in grado di simulare il ragionamento umano [1]. Nell'articolo troviamo anche la prima formulazione del cosiddetto Test di Turing, finalizzato a misurare l'efficacia di un sistema di IA. È un test sulla capacità di una macchina di mostrare un comportamento intelligente equivalente o indistinguibile da quello di un essere umano. Il test prevede che un valutatore interagisca con un essere umano e una macchina senza vederli. Le interazioni sono condotte attraverso un'interfaccia informatica, che consente al valutatore di comunicare con entrambe le parti tramite messaggi di testo. Il compito del valutatore è di determinare quale dei partecipanti sia la macchina e quale l'essere umano, basandosi esclusivamente sulle loro risposte a varie domande. Se il valutatore non riesce a distinguere la macchina

[1] A.M. Turing, *Computing Machinery and Intelligence*, in *Mind*, 49, 1950, pp. 433-460.

dall'essere umano, allora si può dire che la macchina ha superato il Test di Turing, dimostrando un livello di intelligenza che, almeno in quel contesto, è paragonabile all'intelligenza umana. Il Test di Turing è stato un concetto fondamentale nelle discussioni sull'intelligenza artificiale e ha suscitato molti dibattiti su cosa costituisca la vera intelligenza e se una macchina possa realmente possederla.

La proposta di Turing e i primi esperimenti sulle reti neurali servono da ispirazione per la Conferenza di Dartmouth, che si tenne nell'estate del 1956 al Dartmouth College nel New Hampshire e che viene considerata l'evento fondativo per la nascita dell'IA come disciplina. Più che di una vera e propria conferenza, si trattò di un workshop, ufficialmente denominato *Dartmouth Summer Research Project on Artificial Intelligence*, organizzato da John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon, che all'epoca erano figure di spicco del nascente campo dell'Informatica. Il progetto si poneva un obiettivo ambizioso: «To proceed on the basis that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it» [2]. La proposta suggeriva di usare «language, form abstractions and concept» [3] per risolvere problemi solitamente appannaggio dell'essere umano a prevedeva già anche che i sistemi di IA fossero in grado di migliorarsi autonomamente. Sebbene la conferenza non abbia portato immediatamente alla creazione di macchine intelligenti, ha posto le basi per lo sviluppo dell'IA come campo di studi, ha contribuito a cristallizzare gli obiettivi della ricerca sull'IA e ha favorito un ambiente intellettuale che ha incoraggiato i partecipanti a perseguire questi obiettivi nelle loro carriere.

L'IA è dunque una disciplina che nasce inizialmente con lo scopo di creare una macchina in grado di simulare tutti gli aspetti dell'intelligenza umana. Questo approccio, chiamato IA forte, caratterizzato i primi decenni della storia dell'IA. L'IA forte, definita anche "General AI" o "Artificial General Intelligence" (AGI), ha come obiettivo la costruzione di sistemi in grado di comprendere, apprendere e applicare la conoscenza in modo indistinguibile da un essere umano. Questo tipo di IA avrebbe la capacità di ragionare, risolvere problemi, esprimere giudizi, pianificare, imparare e comunicare in

---

[2] J. McCarthy, M. Minsky, N. Rochester, C. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, in *AI Magazine*, 27/4, 1955, p. 2.

[3] *Ibidem*.

qualsiasi ambito di competenza umana. Questo approccio non ha avuto molto successo, a causa della complessità della mappatura del cervello umano e della formalizzazione dell'intero mondo. La situazione è cambiata quando gli scienziati hanno iniziato a prendere in considerazione un altro approccio, chiamato IA ristretta o debole. Questo tipo di IA prevede sistemi progettati per svolgere compiti specifici e per operare in un contesto limitato. In questo caso, il sistema è programmato per seguire regole e schemi predefiniti e per eseguire particolari funzioni, come il riconoscimento di immagini o linguaggio. Esempi di IA debole sono i chatbot, i sistemi di riconoscimento facciale e i motori di raccomandazione e in generale tutti i sistemi di IA che utilizziamo oggi. Questi sistemi sono utili nel gestire i compiti per cui sono stati specificamente progettati, ma non hanno l'abilità di generalizzare le loro capacità. L'IA ristretta è specializzata su una singola applicazione e non vuole replicare tutti gli aspetti dell'intelligenza umana in un unico sistema.

Un'altra distinzione che si è creata nel corso della storia della disciplina è quella tra IA simbolica e IA *data-driven*. L'approccio simbolico, cronologicamente precedente, è un approccio volto a creare una rappresentazione del mondo, completa di algoritmi per ragionare su di essa. In questo modo la macchina impara a pensare, cioè a elaborare conoscenza. Nell'IA simbolica la conoscenza è rappresentata da simboli, che possono essere parole, frasi o immagini che il computer utilizza per rappresentare oggetti, proprietà o altri concetti astratti. Il sistema manipola questi simboli utilizzando regole per ricavare nuove conoscenze o prendere decisioni. L'IA simbolica utilizza spesso sistemi basati su regole, in cui la conoscenza di un dominio specifico è codificata in un insieme di precetti, che stabiliscono come i simboli interagiscono e portano a conclusioni o azioni: ad esempio, se "A" implica "B" e "A" è vero, allora anche "B" deve essere vero. Questo approccio utilizza costrutti logici per esprimere regole e fatti. Uno dei limiti dei sistemi simbolici è la scalabilità, perché le regole e i simboli devono essere specificati e aggiornati manualmente, rendendo difficile la gestione di sistemi complessi. Inoltre, l'approccio simbolico manca di capacità di imparare da nuovi dati: una volta definito un insieme di regole, il sistema non si adatta se non viene aggiornato. Infine, delle difficoltà si incontrano anche nella gestione delle ambiguità e delle informazioni incomplete. Nonostante questi limiti, l'IA simbolica è particolarmente preziosa negli scenari in cui è richiesto un ragionamento chiaro e logico e in cui tutte le informazioni necessarie possono essere definite esplicitamente. Sebbene la sua importanza sia stata messa in ombra dall'ascesa

dell'apprendimento automatico e delle reti neurali, l'IA simbolica svolge ancora un ruolo cruciale in applicazioni specifiche di IA, soprattutto nei modelli ibridi che combinano approcci simbolici e basati sui dati.

Dall'altro lato abbiamo l'approccio *data-driven*. L'IA simbolica ha un approccio *top-down*, ovvero parte dal modello, crea algoritmi per ragionare sul modello e alla fine mette in pratica il sistema su dati reali. L'IA *data-driven*, al contrario, ha un approccio *bottom-up*, perché parte dai dati e fa emergere le informazioni da essi. L'approccio *data-driven* è reso possibile dallo sviluppo del *Machine Learning* (ML), ossia la capacità di un sistema di apprendere in maniera automatica dai dati. Per fare ciò, è necessaria una grande quantità di dati, in modo da fornire al sistema molti esempi di soluzioni. Il termine viene coniato da Arthur Lee Samuel nel 1959, ma sarà solamente a partire dagli anni Ottanta, grazie alle crescenti capacità computazionali dei computer, che il ML potrà ottenere buoni risultati. Nel 1997, Mitchell definisce il ML "the study of computer algorithms that allow computer programs to automatically improve through experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [4]. La forma più recente di ML è il *Deep Learning* (DL), così definito da Goodfellow, Bengio e Courtville:

> Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge from experience, there is no need for a human computer operator to formally specify all the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones; a graph of these hierarchies would be many layers deep [5].

Le tecniche di apprendimento automatico si basano su tre metodi di apprendimento. Nell'apprendimento supervisionato una grande quantità di dati viene fornita al sistema con informazioni complete e modelli per assicurarsi che la macchina abbia una storia di esperienze. In questo modo, quando la macchina si trova di fronte a un problema, attinge al suo database per cercare una soluzione

[4] T. Mitchell, *Machine Learning*, New York, McGraw Hill, 1997, p. 2.

[5] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, Cambridge MA, MIT Press, 2016, p. 6.

adeguata. Nel caso dell'apprendimento non supervisionato vengono forniti dati senza alcuna indicazione del risultato atteso, con l'obiettivo di trovare un modello comune che li colleghi. L'ultima tecnica è l'apprendimento per rinforzo, una metodologia in cui un software impara a prendere decisioni attraverso prove ed errori, interagendo con un ambiente e ricevendo feedback in forma di ricompense o punizioni. È importante sottolineare che, nonostante si utilizzi il termine "intelligenza" sia per l'essere umano che per il dominio dell'IA, le modalità di apprendimento dei sistemi di intelligenza artificiale sono completamente diverse da quelle degli esseri umani: mentre un bambino impara cos'è un oggetto e può dedurne l'uso attraverso pochissime interazioni, i sistemi di IA attualmente in uso non sono in grado di fare nemmeno le inferenze più semplici in così poco tempo. Se a un essere umano basta un paio di occhiali per capire che sono necessari per vedere, un sistema di intelligenza artificiale avrà bisogno di 100 milioni di esempi per poter apprendere la stessa informazione. Una volta appresa, tuttavia, sarà in grado di fornire prestazioni molto più soddisfacenti di quelle umane. Inoltre, il sistema di ML, a differenza dell'essere umano, non è in grado di concettualizzare ed astrarre e dunque non può apprendere ontologicamente le idee.

2. Il problema etico dei dati

L'approccio *data-driven* ha determinato una grande accelerazione dell'intelligenza artificiale, ma ha anche portato alla luce il problema etico dei dati. Il ML necessita di grandi quantità di dati, a cui si fa riferimento con l'espressione "*big data*". Il termine si riferisce a grandi quantità di dati eterogenei, più o meno strutturati e generati rapidamente, che richiedono procedure automatizzate per essere raccolti ed elaborati. Il concetto si riferisce non solo alla quantità di dati, ma anche alle capacità di analisi di questi insiemi di dati. La *Data Science* è la disciplina che utilizza metodi e tecnologie per elaborare grandi quantità di dati al fine di ottenere informazioni comprensibili e correlazioni non ovvie. L'attività del *data scientist* consiste, dunque, nell'estrarre valore dai dati. A questo proposito, all'inizio degli anni 2000 Dough Laney ha proposto il modello delle "3V" per descrivere i *big data* [6]. Questo modello prevedeva

[6] Cfr. F. Diebold, *On the Origin(s) and Development of the Term 'Big Data'*, PIER Working Paper No. 12-037, September 21, 2012, https://ssrn.com/abstract=2152421.

di considerare il volume, la velocità e la varietà dei dati. Con l'avanzare delle interazioni digitali, la generazione di dati è cresciuta in modo esponenziale, rendendo il volume una caratteristica significativa. Le tecnologie di elaborazione in tempo reale richiedono una lavorazione dei dati molto rapida e un'analisi quasi istantanea, rendendo necessaria una forte performatività in termini di velocità. Infine, la varietà si riferisce ai diversi tipi di dati disponibili. I tipi di dati tradizionali erano strutturati e si adattavano a un database relazionale; tuttavia, con i *big data*, abbiamo dati di diversi tipi, strutturati, semi-strutturati e non strutturati. Questa diversità rende necessario utilizzare e sviluppare metodi più complessi per aggregare, armonizzare e analizzare i dati.

In un secondo momento il modello è stato ampliato, aggiungendo altri due punti e arrivando al modello delle "5V". I due punti che sono stati integrati sono il valore, ossia il valore che i dati e le loro possibili correlazioni possono generare, e la veracità, che riguarda l'affidabilità e l'attendibilità dei dati, che diventano indispensabili affinché dati imprecisi non generino informazioni false. Quest'ultimo punto ci conduce direttamente alla questione etica dei dati. Essendo basato su dataset ampi e spesso non controllati, l'apprendimento può portare con sé contenuti errati a livello fattuale e a livello etico-sociale. Ce lo dimostra il problema dei *bias*, che è emerso di pari passo alla sempre maggiore diffusione dei sistemi di IA basati su ML. I sistemi di apprendimento automatico si basano su contenuti creati dagli esseri umani. Ciò significa che qualsiasi pregiudizio, conscio o inconscio, degli esseri umani è incorporato e a volte anche amplificato negli algoritmi. Il principio di questi algoritmi è che i processi decisionali umani siano razionali, e quindi cercano di scovare la *ratio* nei dati, ma non sempre è così. La generalizzazione e la errata interpretazione determinano che gli algoritmi in alcuni casi arrivino a riprodurre e aumentare le disuguaglianze o la discriminazione esistenti, di genere, di etnia, culturali, sociali. I sistemi dovrebbero puntare all'assenza di bias e dunque a quella che viene definita *fairness*, ossia «the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics» [7]. Tuttavia, spesso le scelte effettuate dai sistemi non sono comprensibili dall'essere umano, infatti i sistemi di IA subiscono il problema della *non explainability*, ossia

[7] N. Meherabi, F. Mornstatter, N. Saxena, K. Lerman, A. Galstayin, *A survey on bias and fairness in machine learning*, in *ACM Computing Sulveys* (CSUR), 54/6, 2021, p. 1.

l'incapacità di comprendere e interpretare chiaramente il funzionamento e le decisioni prese da modelli complessi di IA, rendendo difficile fidarsi e verificare le loro scelte.

I *bias* riguardano sia i sistemi visuali che quelli testuali. Nel primo caso la questione si pone in particolar modo per i sistemi di riconoscimento e analisi facciale. Ne è un esempio il caso riportato da Bennet e Keyes, che riguarda un software per la ricerca dei sintomi dell'autismo tramite l'analisi facciale. Il sistema si basava su dati di allenamento che contenevano soprattutto immagini di bambini di etnia caucasica con tratti autistici; per questo motivo non era sempre in grado di riconoscere gli stessi tratti in bambini e bambine di altre etnie, escludendoli automaticamente dal programma di cura e supporto [8]. Per quanto riguarda i sistemi di produzione di testo, il problema diventa abbastanza rilevante dal momento che i *Large Language Models* (LLMs) vengono oggi utilizzati come fonte di informazioni da parte di molti utenti non specializzati. Un LLM è un tipo di sistema di IA progettato per comprendere, generare e interagire con il linguaggio umano. Questi modelli si basano sui principi dell'apprendimento automatico e sono addestrati su diversi set di dati che comprendono testi tratti da libri, articoli, siti web e altre forme di comunicazione scritta. Questo addestramento aiuta il modello ad apprendere modelli linguistici, grammatica, vocabolario e contesto. Un recente studio dell'*International Research Centre on Artificial Intelligence* dell'UNESCO ha sottolineato il problema dei *bias* di genere nei LLMs, analizzando gli output prodotti da alcuni dei principali modelli in circolazione [9]. I risultati mostrano che, nonostante gli sforzi per ridurre i *bias*, molti sistemi continuano a generare risultati con logiche discriminatorie. Per esempio, si nota una tendenza a collegare il genere ai ruoli tradizionali: i nomi femminili vengono spesso associati a temi come casa, famiglia e maternità, mentre i nomi maschili sono legati ai concetti di lavoro e denaro. La situazione non migliora se guardiamo al sentimento nei discorsi generati: le analisi rivelano che il 20% delle frasi generate su una donna contengono contenuti sessisti o misogini, mentre il 70% delle frasi riguardanti l'omosessualità trasmettono un messaggio negativo.

---

[8] C. Bennett, O. Keyes, *What is the Point of Fairness?*, in *Interactions*, May-June 2020.

[9] UNESCO, IRCAI, *Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models*, 2024, https://unesdoc.unesco.org/ark:/48223/pf0000038897.

3. Lavoro e controllo

Il problema etico dei dati non riguarda solamente la produzione degli output, ma anche il processo di raccolta e annotazione dei dati che vengono forniti ai sistemi nella fase di addestramento. Infatti, anche se si parla di apprendimento automatico, un grande lavoro è ancora necessario nella creazione e preparazione dei dataset. Un lavoro che viene svolto principalmente dai cosiddetti *crowdworker*, persone che svolgono mansioni a bassa specializzazione necessarie al funzionamento dei sistemi di ML. Questi lavoratori e lavoratrici provengono e operano di solito nel Sud del mondo, per sfruttare il basso costo del lavoro [10]. Uno studio del 2018 stimava in circa centomila coloro che sono impiegati in questo tipo di attività nella piattaforma Amazon Mechanical Turk [11], mentre nel 2023 è emerso all'attenzione dei media il caso di OpenAI, che impiegava in Africa personale pagato meno di 2 dollari l'ora per il funzionamento di ChatGPT [12].

La dipendenza dell'IA dal lavoro degli esseri umani sta portando dunque a una nuova forma di lavoro sottopagato e privo di norme. È importante sottolineare che ciò non dipende dalle tecnologie in sé, ma dagli esseri umani e in particolare dai vertici delle grandi aziende che decidono di creare profitto attraverso lo sfruttamento dei lavoratori, secondo un processo che è alla base della società capitalista. La questione del *tech-labor* riguarda migliaia di persone in tutto il mondo, il cui lavoro sostiene la società digitale, ma che spesso non beneficiano dei profitti. Tutti coloro che lavorano all'interno della filiera della produzione tecnologica, dai "creusers" del Congo che estraggono il cobalto, a coloro che lavorano nei call center e officine di riparazione dall'India al Kenya, agli "architetti della disinformazione" a Manila, ai moderatori dei contenuti in Arizona. Si tratta in molti casi di lavoratrici e lavoratori che vengono tracciati e lavorano da remoto, con un'organizzazione del lavoro che tende all'isolamento e non permette di creare i vincoli sociali che sarebbero la base per comunità più consapevoli. Il fatto di trovarsi

---

[10] Cfr. S. Betschon, *I precari dei robot*, in *Internazionale*, 1329, 2023, p. 66.

[11] D. Difallah, E. Filatova, P. Ipeirotis, *Demographics and Dynamics of Mechanical Turk Workers*, in *WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Marina del Rey, ACM, 2018, p. 135.

[12] Cfr. F. Ngila, *OpenAI underpaid 200 Kenyans to perfect ChatGPT – then sacked them*, in *Quartz*, 2023, https://qz.com/open-ai-underpaid-200-kenyans-toperfect-chatgpt-1850005025.

in zone del mondo in cui i diritti umani e del lavoro non sempre godono di tutele rende la questione ancora più problematica. Il *tech-labor* è un problema allo stesso tempo fisico, sociale, economico e politico, è pervasivo e specifico, e per essere compreso necessita di guardare ai contesti, alle infrastrutture, alle politiche e alle regole sociali che entrano in gioco.

C'è poi un altro aspetto legato al mondo del lavoro. La vulgata vuole che l'IA arriverà a rendere obsoleti molti degli attuali lavori, "rubandoci" il lavoro. La scelta terminologica ci mostra come si tratti in realtà di una costruzione retorica: "rubare il lavoro" è un'espressione che ci suona tristemente familiare dal momento che spesso è stata utilizzata per gli immigrati. È il segnale di un'opposizione tra un noi e un loro creata su basi puramente propagandistiche, spesso per distogliere l'attenzione da altri problemi sociali, che sono nel caso degli immigrati la mancanza di politiche adeguate sull'accoglienza e sulla povertà interna e nel caso dell'IA le colpe delle grandi aziende nello sfruttamento dei lavoratori. Il confronto con l'altro da sempre genera paura, che è insita nella stessa natura umana, ma che la cultura e il confronto devono portarci a superare. Il mito della ribellione delle macchine è un topos della letteratura fantascientifica, e come tale rimanda a un immaginario letterario, non reale. Nei romanzi di fantascienza la macchina si ribellava e sostituiva l'essere umano perché prendeva coscienza e volontà. Non è questo il caso dei sistemi attuali di IA, che, per quanto possano arrivare a simulare alcuni comportamenti e ragionamenti umani, continuano a non essere senzienti e viventi. La coscienza artificiale è, e rimarrà, nel reame del fantastico. Una volta sgombrato il campo da scenari apocalittici, rimane la questione reale di come l'IA modificherà il mondo del lavoro. Si tratta infatti di una tecnologia che senz'altro introduce nuove modalità di esecuzione dei compiti e in molti casi automatizza dei processi attualmente svolti dagli esseri umani.

Un problema reale che emerge nel mondo del lavoro in relazione all'IA è quello del controllo. Le nuove tecnologie possono fornire alle aziende strumenti non normati per monitorare, gestire e motivare la propria forza lavoro. La tecnologia rende più facile per le aziende mantenere uno stretto controllo sui lavoratori al fine di massimizzare i profitti e minimizzare i costi: i software di monitoraggio per i lavoratori da remoto possono seguire ogni secondo della giornata lavorativa di una persona davanti al computer; le aziende di consegna possono utilizzare i sensori di movimento per tenere traccia degli spostamenti dei conducenti e misurare la

produttività in relazione al tempo impiegato; i lavoratori della *gig economy* si trovano in balia di applicazioni e sistemi che favoriscono la competizione tra loro per una paga talmente bassa che le mance fanno la vera differenza. Le aziende che perseguono più aggressivamente queste tattiche hanno solitamente caratteristiche simili: un grande bacino di lavoratori mal pagati, facilmente sostituibili, spesso part-time, e al vertice un piccolo gruppo di lavoratori ben pagati che progettano il software che li gestisce.

L'automazione non implica necessariamente perdita di lavoro; tuttavia, non si dovrebbe puntare alla creazione di nuovi mezzi di sfruttamento delle lavoratrici e dei lavoratori, ma ad utilizzare al meglio le potenzialità dell'IA. Si dovrebbe cercare di sfruttare l'IA per ridimensionare l'intervento umano in molti lavori ripetitivi e alienanti e favorire nuovi lavori creativi, nei quali l'IA venga integrata quale strumento per la risoluzione del problema specifico ma l'*agency* e la definizione del processo rimangano in mano all'essere umano. Questi nuovi lavori richiederanno competenze multi/interdisciplinari per affrontare problemi sempre più complessi, come il cambiamento climatico, il consumo di energia, l'invecchiamento della popolazione o le pandemie, per citarne solamente alcuni. Per questo è necessario definire un approccio etico e sostenibile, in cui l'IA sia parte della soluzione e non del problema.

4. *AI for Good*

L'IA pone una serie di problemi etici, che entrano nel dominio dell'etica dell'IA, una disciplina che a partire dagli strumenti della filosofia morale indaga i processi messi in atto nella svolta, da alcuni definita epocale, causata dall'avvento dell'IA [13]. Per designare un approccio etico all'IA si utilizza spesso l'espressione *AI for Good*, che definisce un utilizzo dell'IA per risolvere problemi sociali, ambientali ed economici, in opposizione a utilizzi per scopi dannosi. *AI for Good* è un approccio che propone, attraverso un posizionamento militante

[13] Cfr. su questo aspetto M. Mori, *La IAetica come etica nuovissima per un mondo ricreato*, in R. Grimaldi (a cura di), *La società dei robot*, Milano, Mondadori Università, 2022, p. 302. Per un approccio alla disciplina cfr. L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, Raffaello Cortina, 2022; F. Fossa, G. Tamburrini, V. Schiaffonati (a cura di), *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Roma, Carocci, 2021; G. Tamburrini, *Etica delle macchine. Dilemmi morali per la robotica e l'intelligenza artificiale*, Roma, Carocci, 2020.

e un'azione pratica, modalità sostenibili di utilizzo e sviluppo dell'IA. L'obiettivo è sfruttare le potenzialità dell'AI per il beneficio dell'umanità, migliorando la qualità della vita e promuovendo lo sviluppo sostenibile e il rispetto dei diritti umani.

Il termine *AI for Good* è stato reso popolare dall'Unione Internazionale delle Telecomunicazioni (ITU), un'agenzia delle Nazioni Unite specializzata nelle tecnologie dell'informazione e della comunicazione. Nel 2017 l'ITU ha lanciato l'iniziativa *AI for Good Global Summit*, con l'obiettivo di promuovere l'uso dell'IA per affrontare le sfide globali e raggiungere gli obiettivi di sviluppo sostenibile delle Nazioni Unite dell'*Agenda 2030*. L'evento si ripete annualmente e riunisce esperti, ricercatori, politici e rappresentanti del settore privato per discutere e sviluppare soluzioni IA che possano avere un impatto positivo sulla società. Insieme all'*AI for Good Global Summit*, anche altre iniziative hanno portato l'approccio etico all'attenzione della comunità scientifica e dell'opinione pubblica. Sempre nel 2017, L'Unione europea ha pubblicato il suo *Piano d'azione europeo sull'intelligenza artificiale*, che includeva un forte richiamo all'utilizzo dell'IA per il bene sociale. Nel 2018, l'Assemblea generale delle Nazioni Unite ha adottato i *Principi guida per lo sviluppo responsabile dell'intelligenza artificiale*, che sottolineano l'importanza di utilizzare l'IA in modo etico e responsabile. Nello stesso anno il *World Economic Forum* ha lanciato l'*AI for Good Initiative*, un'iniziativa globale che mira a promuovere l'utilizzo responsabile dell'IA per il bene comune. Questi eventi hanno contribuito a creare un clima di maggiore consapevolezza e attenzione all'impatto potenziale dell'IA sulla società. Il termine *AI for Good* è emerso come un modo per definire questa attenzione e sottolineare l'importanza di utilizzare l'IA per risolvere problemi reali e migliorare la vita delle persone. La diffusione di tale approccio si è imposta soprattutto nell'ambito della ricerca e delle organizzazioni no profit, come l'*AI Now Institute* e la *Partnership on Artificial Intelligence*, fino a giungere ai policymaker e alle aziende, seppur in alcuni casi in forma di *ethics washing* [14].

[14] Cfr. G. van Maanen, *AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics*, in *Digital Society*, 1, 9, 2022; B. Wagner, *Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping?,* in E. Bayamlioğlu, I. Baraliuc, L. Janssens, *Being Profiled: Cogitas Ergo Sum. 10 Years of 'Profiling the European Citizen'*, Amsterdam, Amsterdam University Press, 2018, pp. 84-88; A. Gover, *The Ethics Washing of AI*, in *Medium*, 30 November 2023; E. Bietti, *From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics*, in *Journal of Social Computing*, 2, 3, September 2021, pp. 266-283.

Un esempio promette di applicazione di *AI for Good* è l'ambito sanitario [15]. In questo settore l'IA viene utilizzata per migliorare le diagnosi mediche, personalizzare i trattamenti o prevedere la nascita e l'andamento di epidemie e contagi. In ottica predittiva, l'IA mostra la sua utilità anche nell'anticipare l'esito di possibili interventi terapeutici, permette di sviluppare campagne di prevenzione, cura e supporto, definendo sia le modalità di intervento che i beneficiari, può calcolare le aspettative di vita e i rischi di malattia dei singoli individui, aprendo nuove frontiere nella medicina personalizzata. Anche in ambito medico, l'utilizzo dell'IA non è privo di rischi, ad esempio si creano in alcuni casi discriminazioni nei pazienti sulla base di fattori come sesso, età, etnia o abilismo [16]. È il caso di software diagnostici che presentano problemi sistematici nelle diagnosi dei pazienti di etnia non caucasica [17]. È importante, come in altri casi, mantenere la supervisione umana e che l'IA sia utilizzata in maniera complementare e non sostitutiva al lavoro del medico.

Per quanto riguarda la tutela dell'ambiente, l'IA viene impiegata per sviluppare sistemi di monitoraggio e gestione delle risorse naturali, combattere il cambiamento climatico e proteggere la biodiversità. L'IA può essere utilizzata per ottimizzare la produzione di energia rinnovabile, sviluppare reti elettriche più intelligenti, monitorare l'impatto umano sull'ambiente, tenere sotto controllo la fauna selvatica in via di estinzione, combattere il bracconaggio e proteggere gli habitat naturali. Collegato al problema ambientale è anche quello della sicurezza alimentare, che riguarda l'ottimizzazione della produzione agricola, la riduzione degli sprechi alimentari e il miglioramento della distribuzione degli alimenti. Le applicazioni di *AI for Good* alla produzione agricola contemplano attualmente modalità molto diverse tra loro, come l'uso delle risorse in

---

[15] Interessante in questo caso l'intersezione che si crea tra Etica dell'IA e Bioetica: cfr. M. Tai, *The impact of artificial intelligence on human society and bioethics*, in *Tzu Chi Med Journal*, 32 (4), 2020, pp. 339-343.; J. Nabi, *How Bioethics Can Shape Artificial Intelligence and Machine Learning*, in *The Hastings Center Report*, 48, 5, 2018, pp. 10-13.

[16] Cfr. C. Mannelli, *Intelligenza artificiale in ambito sanitario. Riflessioni etiche su explainability e bias*, in *Bioetica*, XXIX, 1/2021, pp. 51-60; C. Luverà, *Il machine learning come strumento di supporto nelle decisioni mediche: questioni etiche e prospettive*, in *Bioetica*, XXIX, 3/2021, pp. 411-428.

[17] Cfr. L. Seyyed-Kalantari, H. Zhang, M. McDermott, *Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations*, in *Nat Med,* 27, 2021, pp. 2176-2182.

funzione del clima e della variabilità dei fattori di contesto, la diminuzione degli sprechi di risorse, la guida assistita dei mezzi meccanici, l'analisi dei dati dei campi per determinare la corretta concimazione o i droni per il monitoraggio. Non solo l'agricoltura, ma anche la produzione industriale trae vantaggio dall'integrazione dei sistemi di IA. Le *smart factories* prevedono l'integrazione tra operatori, macchine e strumenti, nella quale le capacità umane risultano aumentate e si mira a una produzione "zero difetti". In questo ambito risultano utili anche le pratiche di simulazione e *digital twin*, la manutenzione predittiva e l'*additive manufacturing*, che aumenta l'efficienza dei materiali.

In ambito educativo, l'IA permette la creazione di strumenti didattici personalizzati per supportare la formazione inclusiva e ridurre le barriere di accesso all'istruzione. Un sistema di apprendimento personalizzato si adatta alle esigenze individuali di ogni discente, fornendo riscontri in tempo reale e identificando gli studenti e le studentesse che hanno bisogno di un ulteriore supporto. Ma anche al di là dell'ambito educativo, la questione dell'inclusione sociale è centrale per l'approccio *AI for Good*, che mira in ogni campo di applicazione dell'IA a promuovere l'inclusione di gruppi emarginati, migliorare l'accessibilità per le persone con disabilità e ridurre le disuguaglianze. A livello sociale e politico si deve poi contemplare l'utilizzo dell'IA per la promozione della pace e della sicurezza, attraverso il monitoraggio delle minacce, con lo scopo di prevenire i conflitti e promuovere la diplomazia. Tale attitudine si oppone agli utilizzi sempre più diffusi dell'IA in ambito bellico, che mostrano in maniera chiara la contrapposizione etica che si può creare nell'utilizzo della tecnologia, dividendo nettamente un utilizzo *for good* da uno *for bad*. In questo senso pare assolutamente necessario che l'*AI for Good* sia sempre di più un approccio ideologico legato all'azione concreta e militante e che si realizzi per mezzo di una pressione continua e determinata verso i governi e verso gli enti che si occupano di regolamentare l'utilizzo dell'IA.

Una modalità per prevenire l'utilizzo malevolo dell'IA è l'idea di costruire dei sistemi *ethics by design*, nei quali le questioni etiche vengano premesse allo sviluppo stesso del sistema e gestite nella fase di sviluppo, piuttosto che essere affrontate nella fase di utilizzo. Questo si sposa a un approccio *human-centered* e *value-centered*, in cui l'essere umano è posto al centro del design attraverso il rispetto di alcuni valori basilari, come la tutela del benessere e dei diritti delle persone. Tale risultato può essere raggiunto attraverso delle *social minded measures*, misure che valutano gli effetti degli algoritmi sulla

società sulla base del grado in cui rispecchiano fedelmente le nozioni di etica. Ciò comprende l'*algoritmic fairness*, ossia l'attenzione affinché i risultati di un algoritmo non siano influenzati da attributi, spesso protetti o sensibili, che non sono rilevanti per l'attività da svolgere, come il genere, la religione, l'età, l'orientamento sessuale e l'etnia. In questo caso bisogna far attenzione sia all'*individual-based fairness*, secondo cui individui simili devono essere trattati in modo simile (dove la nozione precisa di "simile" è definita caso per caso), che alla *group-based fairness*, per la quale gli individui sono collocati in gruppi in base ai valori di uno o più dei loro attributi protetti e tutti i gruppi devono essere trattati nello stesso modo. Per quanto riguarda i *bias*, è importante per un approccio *ethics by design* considerare il problema della correttezza dei dati e attivare processi di *data debiasing*. Un altro elemento da considerare in fase di design è la diversità, assicurando che tutti i tipi di entità siano considerati e rappresentati nell'output di un processo algoritmico. Attraverso questi processi è possibile ridurre l'ambiguità e la ridondanza dei risultati, arricchire i contenuti per aumentare il coinvolgimento degli utenti ed evitare processi di polarizzazione come *filter bubbles* ed *echo chambers* [18].

La percezione e l'opinione pubblica giocano nella società contemporanea un ruolo determinante nella buona riuscita di qualsiasi operazione. Per questo motivo il presupposto per un'IA sostenibile deve essere la divulgazione della consapevolezza. Un elemento centrale è la formazione multidisciplinare, che permette a chiunque in qualunque ambito svolga la propria attività di comprendere la complessità del mondo in cui viviamo e di individuare le modalità di interazione delle diverse discipline e dei diversi ambiti del sapere, nell'ottica di un superamento della tradizionale divisione tra le due culture. Per fidarsi della tecnologia si deve avere familiarità con la tecnologia. Per questo la diffusione dei sistemi di IA deve essere accompagnata da un maggior impegno a livello sociale verso l'alfabetizzazione digitale, con lo sviluppo di competenze digitali utili a fornire strumenti per capire i funzionamenti delle tecnologie,

[18] Il fenomeno è abbastanza diffuso nei casi di utilizzo di sistemi di IA nei social media. Cfr. E. Rodilosso, *Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization*, in *Philos. Technol*, 37, 71, 2024; H.W. Park, S. Park, *The filter bubble generated by artificial intelligence algorithms and the network dynamics of collective polarization on YouTube: the case of South Korea*, in *Asian Journal of Communication*, 34(2), 2024, 195-212; C. Krönke, *Artificial Intelligence and Social Media*, in T. Wischmeyer, T. Rademacher (a cura di), *Regulating Artificial Intelligence*, Cham, Springer, 2020, pp. 145-173.

discernere l'informazione attendibile e analizzarla in modo critico e responsabile. Per fare ciò bisogna sostenere nelle varie fasi dell'educazione e della formazione l'importanza del pensiero critico, che va applicato anche nel caso dell'utilizzo delle tecnologie. Se da una parte, come si è visto, la tecnologia genera paure infondate, dall'altra siamo sempre più portati a una visione oracolare della tecnologia, in cui i dati, soprattutto se processati digitalmente, sembrano in grado di fornirci verità assolute, che non necessitano di ulteriori verifiche e approfondimenti. Come si è mostrato, i dati sono tutt'altro che neutri e affidabili senza riserve e metterli in discussione è il primo passo verso la definizione di un approccio etico dell'IA. In questo modo è possibile avere un'IA sostenibile, che non contribuisca ad aumentare le distorsioni sociali, economiche e ambientali, ma che invece sia uno strumento al servizio di una svolta positiva nella risoluzione dei problemi sempre più pressanti che il presente ci pone davanti.