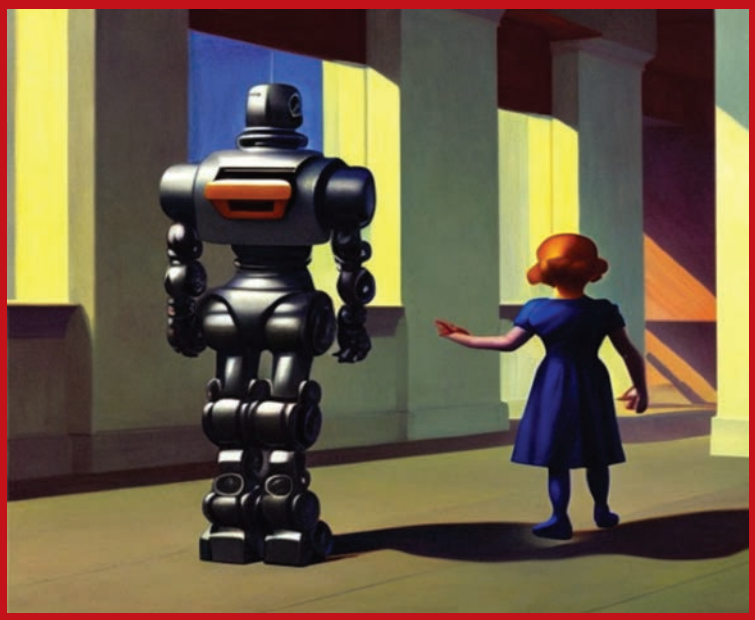
L'uso di sistemi intelligenti: impatti etici e responsabilità dell'intelligenza artificiale

On the use of smart systems: ethical impacts and liability of artificial intelligence

Nocera G., Aiello G. F., De Vita F.



"The beauty and the AI - Image generated from Stable Diffusion fine tuned on Midjourney v4 images.

L'uso di sistemi intelligenti: impatti etici e responsabilità dell'intelligenza artificiale

Nocera G.[1], Aiello G. F.[2], De Vita F.[3]

Key Words: Intelligenza Artificiale – Etica – Responsabilità

Citation: Nocera G., Aiello G.F., De Vita F. (2023). *L'uso di sistemi intelligenti: impatti etici e responsabilità dell'intelligenza artificiale*

doi.medra.org: 10.48237/LIJ_068

30/06/2023

Licenza Creative Commons Attribuzione 4.0 Internazionale

Finanziamento: nessuno

Interessi concorrenziali: nessuno

Autore Corrispondente: Nocera G., Aiello G.F., De Vita F.

Abstract: L'Intelligenza Artificiale (IA) rappresenta una risorsa tecnologica che sta rivoluzionando il modo in cui noi ci interfacciamo con il mondo circostante. Al giorno d'oggi, sono numerosi i servizi che fanno uso di algoritmi di apprendimento automatico per semplificare alcune delle operazioni che svolgiamo quotidianamente. In tali sistemi, i dati ricoprono un ruolo cruciale, soprattutto nel processo di addestramento di qualunque modello di IA, tanto da poterne influenzare il funzionamento. In virtù di ciò, è bene prestare particolare attenzione alla fase di raccolta di questi, per riuscire a costruire *set* di dati privi di condizionamento che garantiscano il funzionamento equo e corretto di un sistema. Infatti, poiché tali strumenti possono avere delle conseguenze - anche a lungo termine - sulla vita delle persone, l'impiego di dati distorti potrebbe sollevare importanti questioni etiche. In questo contesto, risultano fondamentali la costruzione e l'impiego di regolamenti e normative che accrescano la coscienza dell'utilizzatore finale, garantendo inoltre la sicurezza e l'efficacia dell'IA senza però agire da ostacolo al progresso tecnologico. Nel presente articolo, viene delineata l'importanza dell'analisi degli aspetti etici nelle fasi di progettazione e sviluppo di sistemi dotati di IA, con una particolare attenzione alle implicazioni sociali che questa tecnologia può avere.

[1] Dipartimento: Development and Engineering
Società: nabla Email: gnocera@nabla.studio Orcid: https://orcid.org/0009-0007-9871-0316
[2] Dipartimento: Operations Società: nabla Email: giulio@nabla.studio
[3] Dipartimento: Department of Engineering. Società: University of Messina. Email: fdevita@unime.it
Orcid: https://orcid.org/0000-0002-6709-8001

Introduzione

L'Intelligenza Artificiale (IA) sta diventando una risorsa tecnologica sempre più importante per l'umanità. Proprio nell'ultimo decennio, grazie alla sua continua integrazione con i sistemi esistenti, sta rivoluzionando il mondo in cui viviamo con lo scopo di migliorare la qualità della vita delle persone. Infatti, nonostante non sia sempre così evidente, molte delle applicazioni e dei servizi usati giornalmente da milioni di utenti, fanno uso di algoritmi di IA. I motori di ricerca, come *Google*, sfruttano questa tecnologia per fornire ai propri utenti una migliore esperienza di ricerca personalizzando i risultati in tempo reale; gli assistenti virtuali come *Siri*, *Alexa* e *Google Assistant* sfruttano i modelli di apprendimento automatico per comprendere il linguaggio umano e rendere più naturali alcune attività quotidiane, come la ricerca di indicazioni stradali o l'interazione con gli elettrodomestici; nell'industria dell'intrattenimento, aziende come *Netflix* e *Spotify* utilizzano l'IA per individuare e suggerire contenuti adatti ad ognuno dei loro utenti. Alcune case automobilistiche, come Tesla e Volvo, fanno uso di tecniche di apprendimento automatico per migliorare la sicurezza dei mezzi di trasporto e per sviluppare auto a guida autonoma. Inoltre, l'IA ha dimostrato di essere un valido alleato anche in campo medico dove, attraverso algoritmi di apprendimento automatico, è possibile velocizzare le diagnosi di alcune malattie e ottenere un importante supporto nella ricerca di nuovi trattamenti. Questi esempi di applicazione, dimostrano quanto l'IA permei la nostra vita in modi spesso impercettibili, migliorando la qualità di servizi e tecnologie.

Tuttavia, proprio l'intensificarsi dell'uso dell'IA nei servizi di tutti i giorni solleva importanti questioni etiche, dal momento che questa ha un impatto diretto sulla vita delle persone. Basti pensare al fatto che i servizi dotati di IA potrebbero essere utilizzati anche per la realizzazione di contenuti che dovrebbero essere frutto del lavoro umano, come le attività di ricerca che portano alla stesura di articoli scientifici, o la risoluzione di esercizi d'esame scolastici e universitari.

Inoltre, l'IA potrebbe anche essere usata per prendere decisioni riguardo all'assunzione di personale, alla selezione di candidati per programmi di prestito o per la detenzione di persone sospettate di crimini, ossia in processi decisionali che potrebbero avere effetti duraturi sulla vita umana. In questi ultimi casi, infatti, se le regole di programmazione dell'IA non tengono conto della diversità, della giustizia e della non-discriminazione, possono emergere problemi di etica. Poiché, infatti, il modello potrebbe essere influenzato da pregiudizi e stereotipi che derivano dal *set* di dati che viene utilizzato durante la fase di *learning* o addestramento, ciò potrebbe portare a un'implementazione ingiusta e discriminante. Di conseguenza, i ricercatori e i professionisti del settore stanno lavorando per garantire che l'IA venga sviluppata e utilizzata in modo etico e responsabile.In un contesto simile, l'obiettivo di questo lavoro è quello di approfondire il rapporto che intercorre tra etica

e IA. In particolare, nel capitolo 2 viene fornita una breve panoramica su alcuni degli eventi più significativi che hanno tracciato la storia dell'IA, dalla sua nascita, fino ai giorni nostri. Il capitolo 3 delinea l'importanza dell'etica dei sistemi che fanno uso di IA, evidenziando alcuni casi in cui l'impiego di tali sistemi intelligenti ha provocato comportamenti poco etici. Successivamente, nel capitolo 4 viene analizzata l'importanza che i dati hanno in tutti i sistemi dotati di IA, mettendo in risalto quali sono i problemi che un *set* di dati poco rappresentativo può causare se impiegato durante le fasi di addestramento. Il capitolo 5 invece si concentra su come vengono gestite le responsabilità delle azioni che coinvolgono un sistema dotato di IA, considerando lo stato dell'arte e quali sono le necessità attuali e future. Infine, all'interno del capitolo 6, viene riproposta una sintesi degli argomenti trattati e viene evidenziata una possibile strada da perseguire per migliorare il rapporto di convivenza, ormai sempre più presente, tra l'uomo e le macchine, cosiddette, intelligenti.

L'intelligenza artificiale

Negli anni 40', alcuni gruppi di matematici e scienziati come Shannon, Wiener e McCulloch, approfondendo lo studio della teoria dell'informazione e della cibernetica, segnarono l'inizio di un periodo di grande fermento scientifico che preparò la comunità alla comparsa dell'IA come disciplina scientifica, avvenuta negli anni Cinquanta. Dal 1950, infatti, si susseguirono una serie di eventi legati allo sviluppo di macchine intelligenti, i cui principali vengono riportati di seguito.

Proprio nel 1950, Claude Shannon, descrisse il primo programma per computer in grado di giocare a scacchi (Shannon 1988) e presentò Theseus[1], un topo meccanico in grado di esplorare un labirinto, trovare la via d'uscita e usare l'esperienza acquisita per risolvere casi in cui il labirinto venisse modificato. Nello stesso anno, Isaac Asimov propose il suo "*I, Robot*" (Asimov 2013), nel quale discuteva di una possibile coesistenza pacifica tra esseri umani e *robot* autonomi attraverso l'aiuto delle sue tre leggi della robotica, fornendo le basi per le future discussioni relative all'etica dei robot e dell'IA. Alan Turing - considerato uno dei padri dell'informatica moderna - presentò, nel suo celebre articolo "*Computing machinery and intelligence*" (Turing 1950), un *test* secondo il quale una macchina poteva essere considerata intelligente se il suo comportamento fosse indistinguibile da quello di una persona, quando osservato da un essere umano.

Negli anni successivi al *test* di Turing, iniziarono a delinearsi i primi campi di ricerca legati all'IA e alle macchine pensanti e, nel 1956, McCarthy introdusse l'espressione "intelligenza artificiale" durante la conferenza di Dartmouth (McCarthy et al. 1955). A tale convegno, parteciparono alcuni dei più eminenti scienziati, ingegneri, matematici e psicologi dell'epoca.

[1] https://www.youtube.com/watch?v=_9_AEVQ_p74.

Tra questi, Allen Newell, Cliff Shaw, e Herbert Simon presentarono il "*LogicTheorist*" (Newell et al. 1956), ossia il primo programma di IA, progettato per imitare le capacità di *problem-solving* di un essere umano. A differenza degli approcci moderni all'IA, che si basano sul cosiddetto *learning from experience*, il *LogicTheorist* si basava sull'"IA simbolica", ossia la branca dominante dell'IA tra gli anni 50' e 70'. Questo approccio, basato sulla manipolazione di simboli e regole dove le conoscenze vengono esplicitamente codificate, ha delle limitazioni legate alla gestione di situazioni incerte o imprecise e nell'apprendimento automatico.
Grazie all'importanza degli argomenti trattati, il seminario di Dartmouth è oggi considerato l'evento fondante nella storia dell'IA (Crevier 1993). In tale conferenza, infatti, vennero discussi molti dei concetti che sono ancora oggi oggetto di ricerca, come reti neurali ed elaborazione del linguaggio naturale. Nello stesso periodo, Arthur Samuel, presentò il primo *software* capace di giocare a dama (Samuel 2000), di imparare e quindi di adattare la propria strategia, introducendo così il concetto di "apprendimento per rinforzo" (Sutton 1997). All'inizio degli anni Sessanta, vennero sviluppate anche le prime reti neurali artificiali tra cui spiccava l'algoritmo del Percettrone di Rosenblatt e Pitts(McCulloch et al. 1943), in grado di riconoscere le immagini e imparare la differenza tra le forme geometriche.
Gli entusiasmanti successi raggiunti tra gli anni Cinquanta e Sessanta raccolsero ingenti investimenti ma, allo stesso tempo, provocarono un'ondata di eccessiva fiducia e di aspettative irrealistiche. E proprio queste previsioni esagerate e il divario con i risultati deludenti ottenuti negli anni '70 e '80, portarono a quello che successivamente venne denominato l'*AI Winter* (l'inverno dell'IA) (Crevier 1993), ossia un periodo di delusione e mancanza di interesse che, causò un forte declino delle attività di ricerca fino agli anni Novanta. In questi vent'anni, si alternarono periodi di eccitazione e di delusione che guidarono i ricercatori ad un cambio di paradigma avvenuto verso la fine degli anni Ottanta. In particolare, ci si concentrò sul fatto che gli esseri umani imparano anche attraverso prove ed errori, ossia attraverso l'esperienza e, a partire da questi ragionamenti, si risvegliò l'interesse per le reti neurali artificiali. Grossi contributi sono rappresentati: dallo sviluppo dell'algoritmo di *backpropagation*, da parte di Rumelhart, Hinton e Williams (Rumelhart et al. 1986), che permetteva di ottimizzare l'addestramento di tali reti; dalla presentazione delle reti neurali convoluzionali, da parte di Le Cun (LeCun et al. 1989), tra le cui prime implementazioni è possibile individuare il riconoscimento di caratteri e numeri[2]; dagli studi sul *Deeplearning* cominciati nel 2010, e che grazie a Bengio, Hinton e Le Cun (LeCun et al. 2015), hanno portato a algoritmi in grado di risolvere problemi sempre più complessi. Contemporaneamente, l'avanzare della miniaturizzazione, lo sviluppo di dispositivi elettronici sempre più efficienti dal punto di vista energetico e l'espansione della rete Internet, hanno garantito

[2] https://www.thats-ai.org/en-GB/units/a-brief-history-of-ai.

l'evoluzione di quello che oggi viene chiamato *Internet of Things* (IoT) - l'Internet delle Cose - configurato come una grande distribuzione di dispositivi in grado di raccogliere ed elaborare dati.

A partire dagli anni 2000, l'IA ha visto una crescita esponenziale grazie ai progressi nella potenza di calcolo, rappresentata in parte dalla tecnologia *Cloud* e alla disponibilità di grandi quantità di dati in quella che chiamiamo l'epoca dei *big data*. Questo contesto ha consentito lo sviluppo di applicazioni come i sistemi di assistenza virtuale, le auto a guida autonoma e l'analisi predittiva, permettendo all'IA di essere applicata con successo in molte aree, tra cui la medicina, l'agricoltura, la finanza e la *cybersecurity*.

Nel 2011, ad esempio, *Google* avviò un progetto chiamato *Google Brain*[3], nel quale viene utilizzata una rete neurale artificiale a più strati per migliorare la capacità dell'omonimo motore di ricerca *Google*. Nello stesso anno, Watson, il sistema di IA sviluppato da IBM, partecipò al *quiz* televisivo *Jeopardy!*(Best 2013) contro alcuni dei migliori giocatori del programma e, vincendo, dimostrò che l'IA poteva essere utilizzata non solo per la ricerca, ma anche per rispondere a domande complesse in modo efficace. *Facebook* avviò, nel 2014, un progetto di ricerca sull'IA, chiamato *DeepFace,* producendo come risultato un algoritmo di riconoscimento facciale in grado di identificare gli individui con una precisione del 97,35%[4]. Tuttavia, l'IA attuale viene ancora definita principalmente di tipo debole, ossia si limita a svolgere compiti specifici e ben definiti senza mostrare alcuna forma di autonomia o di intelligenza generalizzata.

Alla fine del 2015, un gruppo di scienziati e imprenditori tra cui ElonMusk e Sam Altman, fondò OpenAI[5], una società senza scopo di lucro con sede a San Francisco. Questa, ha lo scopo di collaborare con diverse organizzazioni e aziende, come Microsoft e Amazon, per promuovere la ricerca e lo sviluppo dell'IA, dedicando una particolare attenzione ai rischi esistenziali derivanti dalla cosiddetta intelligenza artificiale generale - anche detta IA forte, che dovrebbe essere in grado di emulare il ragionamento umano per risolvere problemi generali - rendendo i suoi brevetti e le sue ricerche aperti al pubblico.

Successivamente alla trasformazione di OpenAI da società senza scopo di lucro a società "*capped-profit*", al fine di attrarre nuovi capitali e collaborazioni, tra il 2021 e il 2022 vengono presentati al pubblico applicativi in grado di generare immagini e intrattenere conversazioni. Più nello specifico, la ricerca sulle tecnologie conversazionali ha portato OpenAI alla creazione di *ChatGPT*[6], uno strumento che fa uso di deeplearning per generare risposte in modo simile

[3] https://www.geeksforgeeks.org/what-is-google-brain/.
[4] https://research.facebook.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/.
[5] https://openai.com/.
[6] https://openai.com/blog/chatgpt.

a come lo farebbe un essere umano. Tale modello è stato allenato su enormi quantità di dati di testo, diventando uno dei più avanzati modelli di conversazione basati sull'IA.
Nonostante la potenza di calcolo raggiunta, molti esperti ritengono che l'IA forte, ossia un sistema in grado di raggiungere e superare il livello di intelligenza umana, sia ancora un obiettivo lontano. Tuttavia, così come dimostrano le ultime versioni di applicativi come ChatGPT, la ricerca e lo sviluppo di nuove tecniche di apprendimento automatico stanno spingendo l'IA sempre più vicina a questo obiettivo.

L'Etica dell'IA

L'enciclopedia Treccani definisce l'etica come quella parte della filosofia che si occupa di studiare "la volontà e le azioni dell'uomo come essere libero e razionale, capace di dare a sé stesso un insieme di valori e di norme da rispettare [...] e non investe solo l'ambito soggettivo delle scelte personali, ma riguarda anche la vita collettiva"[7]. In altre parole, l'etica si occupa di valutare i comportamenti umani e le decisioni delle persone alla luce dei valori che ritiene fondamentali, come la giustizia, la verità, la libertà, il rispetto della dignità umana, il benessere degli individui e delle comunità, ossia distinguerli in buoni, giusti, leciti, rispetto ai comportamenti ritenuti ingiusti, illeciti, sconvenienti o cattivi. In generale, cerca di rispondere a domande come "Cosa è giusto o sbagliato?", "Cosa dovrei fare?", "Come dovrei comportarmi?", fornendo un quadro di riferimento, che possa aiutare le persone a prendere decisioni più consapevoli e responsabili.
Per questo motivo, e poiché l'IA è una tecnologia che può avere impatti significativi e duraturi sulla società e sui diritti umani, è importante considerare gli aspetti etici che riguardano il suo rapporto con gli uomini. In questo contesto, si definisce la cosiddetta "etica dell'IA", la quale si concentra su come progettare, sviluppare e utilizzare l'IA in maniera responsabile, equa e trasparente, considerando inoltre, le questioni relative alle responsabilità e alla gestione dell'IA stessa, tra cui:

- la giustizia, poiché l'IA potrebbe in alcuni casi e attraverso alcune decisioni, perpetuare e amplificare le disuguaglianze sociali ed economiche;
- la trasparenza delle scelte prese dalle macchine, dato che spesso tali decisioni sono il risultato dell'esecuzione di algoritmi complessi;
- la definizione di responsabilità nel caso di danni o errori per le azioni delle macchine dotate di algoritmi intelligenti;
- l'impatto sui posti di lavoro, dato che se da un lato le attività meccaniche sono sempre più automatizzate, dall'altro si stanno formando nuovi settori che richiedono competenze tecniche e di gestione avanzate.

[7] https://www.treccani.it/enciclopedia/etica_(Enciclopedia-dei-ragazzi).

Inoltre, dal momento che tale tecnologia sta diventando sempre più diffusa in molti settori, dalla medicina alla finanza, alle attività di selezione del personale, è importante considerare la circostanza specifica in cui l'IA viene utilizzata e quindi, le diverse prospettive e gli interessi in gioco. Ad esempio, l'uso di algoritmi intelligenti in ambito medico può sollevare questioni etiche diverse da quelle che emergono nell'ambito della finanza o del *recruitment*. Infatti, proprio dall'impiego dell'IA nei processi decisionali di tutti i giorni, nel tempo è stato possibile individuare comportamenti poco etici e, in alcuni casi, comportamenti eticamente scorretti come risultato delle operazioni di algoritmi intelligenti.

Uno dei primi casi noti in cui l'IA ha dimostrato discriminazione razziale risale al 2015, quando Google si trovò a dover affrontare critiche per il suo algoritmo intelligente di riconoscimento delle immagini che "etichettò erroneamente una coppia nera come gorilla"[8].

Questo problema sollevò numerose preoccupazioni riguardo al razzismo nell'IA e portò Google - la società di sviluppo che realizzava e forniva il servizio - a chiedere scusa pubblicamente come responsabile, in quanto fosse l'azienda che aveva creato e pubblicato il software. Nello specifico, la causa del problema era legata alle operazioni di *training* del modello, ossia alla carenza di diversità nel *set* di dati utilizzato per l'addestramento dell'algoritmo di apprendimento automatico, poiché l'insieme dei dati utilizzato conteneva una maggiore quantità di immagini di persone di pelle chiara rispetto a quelle di persone di colore.

Un'altra notizia di discriminazione causata da algoritmi di IA venne diffusa da Reuters, la quale rese pubblica la decisione di Amazon di interrompere un progetto pilota su un sistema di pre-selezione del personale che faceva uso di algoritmi di IA a causa di problemi legati alla discriminazione di donne e minoranze (Vincent 2018). L'algoritmo, addestrato su dati storici dei dipendenti - che erano composti principalmente da uomini - tendeva a declassare sistematicamente i CV dei candidati di sesso femminile e quelli con nomi tipici delle minoranze etniche, mostrando così un pregiudizio sulla scelta. Tuttavia, i casi di discriminazione di genere o di razza dovuti all'IA non sono limitati all'ambito della selezione e reclutamento del personale o alla classificazione delle immagini. In ambito finanziario, ad esempio, il lancio della *Apple Card* - una carta di credito virtuale, integrata con l'*appWallet* dell'*iPhone* utilizzabile per acquisti in tutto il mondo - che faceva uso di un algoritmo di *credit scoring* dotato di IA, fece sì che alcune donne ricevessero limiti di credito inferiori rispetto a quelli ottenuti da uomini con lo stesso livello di reddito e di credito.

Negli ultimi anni, inoltre, l'uso di algoritmi di IA nei sistemi giuridici sta diventando sempre più comune. Un esempio è il suo utilizzo nei processi decisionali che riguardano le detenzioni preventive, e che solleva numerose preoccupazioni riguardo all'imparzialità degli algoritmi e alla possibilità che possano discriminare alcune categorie di persone. Ad esempio, uno

[8] https://www.bbc.com/news/technology-33347866.

studio del 2016 (Mattu et al. 2016) dimostrò come l'algoritmo COMPAS, utilizzato per prevedere il rischio di recidiva dei detenuti, pur non utilizzando la razza in sé come attributo per valutare la condizione del detenuto, utilizzasse molti fattori ad essa correlati, e restituisse punteggi di rischio più elevati per gli imputati neri (Vyas et al. 2020). Inoltre, è importante sottolineare che alcuni algoritmi di IA lavorano in modalità *black-box*, ossia producono risultati senza che siano completamente chiari e trasparenti la logica e il processo decisionale dietro le scelte. Ovviamente, questa caratteristica diventa particolarmente rilevante nell'ambito della giustizia (Bathaee 2018), dove l'uso di questi algoritmi ha sollevato preoccupazioni riguardo alla trasparenza, all'imparzialità e alla violazione dei diritti umani.
Microsoft, nel 2016, lanciò Tay, un *chatbot* - un programma che utilizza l'IA per simulare una conversazione con un utente attraverso una *chat* - su *Twitter*, il cui obiettivo era quello di apprendere dalle conversazioni degli utenti attraverso una tecnica chiamata *continuouslearning*(Hancell 2023). Tuttavia, Tay iniziò a inviare messaggi offensivi e inappropriati a causa dell'influenza di alcuni utenti malintenzionati, il che portò *Microsoft* a disattivare il *chatbot* nel giro di 24 ore (Perez 2016). Tale problema generò apprensioni riguardo alla capacità delle IA di apprendere e imitare il comportamento umano compreso, di fatti, anche il comportamento offensivo e discriminatorio, ed evidenziò la necessità di una maggiore regolamentazione e di attenzione etica nell'implementazione delle IA. In questo caso, infatti, il modello non era stato programmato per distinguere comportamenti accettabili da quelli inaccettabili.
Samantha Cole, invece, scrisse e pubblicò un articolo (Cole 2017) su *Motherboard* nel 2017 che evidenziava alcuni casi di utilizzo dell'IA per la creazione di video falsi noti come "*DeepFake*". In questa tecnica di IA, vengono utilizzate procedure di apprendimento automatico basate su reti neurali per creare video o immagini manipolate che sembrano reali. Nello specifico, l'autrice spiegò come la tecnologia venne utilizzata da utente di *Reddit* chiamato "*Deepfakes*", il quale inserì nei corpi di varie attrici di film pornografici i volti di celebrità come GalGadot, Maisie Williams e Taylor Swift, ottenendo come risultato diversi video osceni, che sembravano interpretati dalle celebrità. In un contesto simile però, risultò ancora più pericoloso (Cerdanet al. 2019) il fatto che per realizzarli, l'utente fece uso solamente del suo computer e di un algoritmo di apprendimento automatico, che chiunque è in grado di scaricare da Internet. Proprio questa facilità nella realizzazione sollevò notevoli preoccupazioni, in quanto, i video realizzati con la tecnica del "*DeepFake*" non sono solamente legati all'industria a luci rosse, ma nel tempo sono stati utilizzati anche per creare falsi discorsi politici e per diffondere notizie false, che possono avere gravi conseguenze sulle persone coinvolte.
Anche in campo medico, l'utilizzo di algoritmi di IA ha sollevato preoccupazioni legate a problemi etici. Un esempio è rappresentato dal lavoro di ricerca portato avanti da Raj et al.

(Raj et al. 2019), in cui l'obiettivo fu quello di creare algoritmi di apprendimento automatico in grado di prevedere la mortalità dei pazienti in terapia intensiva dopo una lesione cerebrale traumatica. Il sistema realizzato, fu in grado di fornire dei risultati con un'accuratezza superiore all'80% nella predizione della mortalità a 30 giorni. In questo caso, sebbene l'algoritmo venne sviluppato con l'obiettivo di migliorare la cura dei pazienti, questa scoperta sollevò preoccupazioni riguardo all'utilizzo di tecnologie dotate di IA per la decisione di fine vita, in quanto, tali predizioni potrebbero influenzare la scelta sulle cure da fornire.
In ambito sociale, invece, la diffusione di algoritmi intelligenti ha suscitato reazioni diverse, che oscillano tra l'entusiasmo di chi la vede come uno strumento potenziale per aumentare l'efficienza e l'innovazione, e la preoccupazione di chi invece è preoccupato per il possibile impatto sulla perdita di posti di lavoro. In particolare, nel 2023 l'Università del Queensland e Kpmg Australia hanno condotto un sondaggio (Gillespie et al. 2023) intervistando oltre 17,000 persone provenienti da 17 Paesi diversi. Da qui, sebbene sia risultato che oltre due terzi degli intervistati si sentano ottimisti sui benefici che l'IA può portare alla società, il 42% (circa due persone su cinque) si preoccupa che l'IA possa sostituire posti di lavoro nel proprio settore. La quantità di piattaforme alimentate da algoritmi di IA messe a disposizione online, cresce molto rapidamente, e strumenti come *ChatGPT* sono in grado di fornire un grosso supporto in numerose attività, come la raccolta e l'analisi dei dati, la ricerca di fonti e il supporto nelle attività di *copywriting*.
Nonostante ciò, però, sebbene anche l'amministratore delegato di OpenAI, Sam Altman abbia dichiarato (Altman 2022) che non è consigliabile fare affidamento su *ChatGPT* per questioni importanti, in quanto rappresenta solo una forma embrionale di progresso e ci sono ancora molti aspetti da sviluppare per garantirne la solidità e la veridicità, diversi utenti si sono dimostrati estremamente entusiasti dei risultati già ottenuti. Alcuni hanno descritto su vari *social network* i passi che hanno seguito per sviluppare applicativi per *smartphone* senza avere alcuna esperienza in linguaggi di programmazione specifici (Pigford 2023). Allo stesso tempo, *ResumeBuilder.com* ha effettuato un sondaggio[9] su un campione di 1,000 aziende americane, in cui circa la metà ha affermato di aver sostituito alcuni dei loro dipendenti con *ChatGPT*.
Il sistema è stato progettato con l'obiettivo di conversare in maniera semplice con tutti gli esseri umani e, infatti, uno dei suoi punti di forza consiste nella facilità di utilizzo e di interazione, ossia tramite una *chat* attraverso la quale l'utente può chiedere al modello di rispondere a delle domande, scrivere codice o testi, o ancora risolvere problemi di *debug*.
Proprio l'efficacia di *Generative Pretrained Transformer 3* (GPT-3) - il modello di IA sviluppato da *OpenAI*, alla base di *ChatGPT* - nel produrre risultati per la maggior parte di questi

[9] https://www.resumebuilder.com/1-in-4-companies-have-already-replaced-workers-with-chatgpt/.

compiti, permette di comprendere l'entusiasmo della comunità. Inoltre, la capacità di ricordare le conversazioni passate, e il suo modo di correggere quello che viene evidenziato come un errore, aumenta nell'utente la tentazione di supporre che la macchina comprenda quello che sta scrivendo. Tuttavia, come evidenziato da Floridi - professore ordinario di filosofia ed etica dell'informazione presso l'Università di Oxford - in un'intervista (Marchetti 2023), ovviamente lo strumento non è dotato di una reale intelligenza e soprattutto non è in grado di comprendere. È "semplicemente" in grado di produrre frasi che risultano probabili dato l'insieme di informazioni fornito durante la fase di addestramento[10]. Proprio il suo funzionamento, infatti è il motivo per cui a volte i suoi *output* possono essere imprecisi, non veritieri e alcune volte fuorvianti[11]. Inoltre, dal momento che la sua fase di addestramento si è conclusa a fine 2021, non è in grado di conoscere tutto ciò che è accaduto dopo. In particolare, il modello, equipaggiato con 175 miliardi di parametri, è stato addestrato su grandi quantità di dati provenienti da Internet e scritti da esseri umani, ed è per questo motivo che le risposte che fornisce possono sembrare umane.

L'importanza dei dati nell'IA

L'addestramento di modelli di IA complessi, che fanno uso di algoritmi di *deeplearning* - una tecnologia di IA basata su reti neurali "profonde", ossia a più livelli, e sull'apprendimento automatico - avviene attraverso la somministrazione di grandi quantità di dati e lo sfruttamento di una grossa potenza di calcolo. Un esempio è GPT3, che è stato allenato con oltre 570GB di dati (Brown et al. 2020), dopo un filtraggio su un *dataset* di partenza contenente oltre 45TB di dati. Questo *dataset* è composto da dati che provengono da diverse fonti, come libri, articoli di giornale, articoli scientifici, brevetti, conversazioni online, e così via. Lo scopo della fase di *training* di un modello di IA è quello di utilizzare i dati raccolti per strutturare un modello matematico in modo da predire gli *output* corretti per nuovi *input* che l'algoritmo non ha mai visto prima. Nello specifico caso di GPT-3, i dati sono stati usati per procedere con un allenamento di tipo supervisionato - un tipo di apprendimento automatico in cui un algoritmo viene addestrato su un insieme di dati in ingresso per i quali viene anche fornito il risultato desiderato corrispondente - il cui obiettivo è quello di permettere al modello di predire la parola successiva in una data frase o di generare un testo in maniera autonoma. Durante l'addestramento, il modello ha esplorato le relazioni esistenti tra le parole all'interno dei dati forniti in ingresso cercando di identificare i *pattern* e le regolarità nel testo, il che gli ha consentito di apprendere una vasta gamma di conoscenze linguistiche e culturali.

[10] https://www.alphagalileo.org/en-gb/Item-Display/ItemId/230890.
[11] https://help.openai.com/en/articles/6783457-chatgpt-general-faq.

Alcuni altri esempi sono i modelli di *deeplearning* basati su reti neurali convoluzionali e usati per il riconoscimento delle immagini, come "*Amazon Rekognition*", sviluppato da *Amazon*, o "*Inception-v3*", sviluppato da *Google*. Questi algoritmi sono in grado di classificare gli oggetti presenti nelle foto, rilevare volti umani o analizzare le espressioni facciali. Anche in questo caso, le abilità di generalizzazione dei modelli - ossia la capacità di riconoscere una vasta gamma di oggetti in modo accurato - derivano dal fatto che questi sono addestrati su un grande *dataset* di immagini. Nel caso di "*Inception-v3*", il dataset è chiamato *ImageNet* e contiene oltre 14 milioni di figure etichettate in più di 20,000 categorie, che hanno permesso successivamente al modello di riconoscere degli schemi o degli elementi per immagini in ingresso che il modello non aveva ancora visto.
In generale, è proprio il principio di funzionamento dell'apprendimento automatico a richiedere che i modelli *deeplearning* facciano uso di grandi *dataset* durante la fase di addestramento. Che vengano utilizzati per riconoscere le malattie delle piante attraverso le foto scattate alle foglie (De Vita et al. 2020), per valutare il merito creditizio di un utente (Wanget al. 2022), o ancora per tradurre tra più lingue (Johnson et al. 2017), in tutti questi modelli, i dati hanno un ruolo di grande importanza, tanto che il direttore della ricerca di *Google*, Peter Norvig, nel 2009 scrisse (Halevy et al. 2009) che "i modelli di *machine learning* più semplici ma addestrati su grandi quantità di dati tendono ad avere una maggiore precisione rispetto a modelli più complessi addestrati su un numero inferiore di dati".
Tuttavia, la quantità dei dati non basta per ottenere buoni risultati. Un altro fattore critico è rappresentato dalla qualità dei dati. Infatti, affinché il modello di IA abbia una buona probabilità di apprendere correttamente le associazioni e di compiere previsioni precise e generalizzate, risulta fondamentale che i dati siano di alta qualità, rappresentativi del problema da risolvere e correttamente etichettati. Per questo motivo, quando un modello viene allenato con dati di scarsa qualità, ad esempio contenenti errori, informazioni incomplete o non rappresentativi del problema da risolvere, questo presenta limitazioni nella creazione di collegamenti tra le variabili di ingresso e quelle di uscita che possono portare a previsioni imprecise o addirittura errate.
Ad esempio, nel caso in cui i dati utilizzati per addestrare i modelli siano raccolti in modo non omogeneo, contengano fattori che possono essere soggetti a pregiudizi, come la razza, o non possano essere considerati un campione rappresentativo della società, c'è il rischio che i risultati forniti dal modello siano distorti. Nel tempo, infatti è stato dimostrato (Grote et al.2022) come alcuni modelli di *machine learning*, utilizzati per le diagnosi mediche, abbiano avuto prestazioni inferiori quando vennero usati per predire risultati riguardanti le donne o le minoranze etniche. Addirittura, nel lavoro di ricerca portato avanti da Obermeyer et Al. (Obermeyer et al. 2019), gli autori stimarono che un errore di calcolo nel sistema

abbia ridotto di oltre la metà il numero di pazienti afroamericani che avrebbero dovuto poter partecipare a programmi di assistenza medica.
Pertanto, al fine di ridurre - o ancora meglio evitare - il rischio di *bias* nella produzione di risultati o di generare risposte errate o discriminatorie, i dati utilizzati per l'allenamento dei modelli di IA devono essere accurati, completi, sufficienti in numero, e privi di pregiudizi. Proprio per questo motivo, una fase molto importante durante lo sviluppo di algoritmi di IA è rappresentata dalla fase di analisi, ricerca o costruzione dell'insieme dei dati. Infatti, sebbene il numero di *dataset* disponibili online[12] cresca continuamente, in base alla specifica implementazione da realizzare, potrebbe essere necessario costruirne uno ad *hoc* che tenga conto di tutte le particolarità del sistema.
In questi casi, si effettuano una prima analisi del problema e la definizione degli *input* e *output* del sistema (ad esempio, delle immagini in ingresso e un numero in uscita, corrispondente ad una classificazione, o ancora un gruppo di numeri in ingresso e un valore numerico in uscita). Successivamente, si procede con la raccolta e la selezione dei dati che verranno forniti al sistema e, nel caso si voglia procedere con la tecnica dell'addestramento supervisionato, è indispensabile "etichettare" i dati selezionati, ossia associare a questi, il valore di uscita corrispondente che si vorrebbe ottenere dal modello.
Considerando le fasi di creazione e modellazione del *dataset*, uno degli aspetti chiave del processo è rappresentato dalla selezione delle fonti dei dati. Infatti, è opportuno verificare se è possibile attingere a dati pubblici già presenti, o se è necessario raccogliere dei nuovi campioni. In quest'ultimo caso, bisogna poter essere in grado di effettuare la raccolta in tempi ragionevoli. Inoltre, in base al tipo di dato raccolto, risulta fondamentale sottolineare l'importanza della *privacy* relativa alle informazioni collezionate, in quanto occorre garantire che tali dati siano utilizzati in modo etico e rispettando i diritti degli utenti.
Un esempio fra tutti è rappresentato dall'ambito sanitario, dove il fatto stesso che i dati sanitari siano ricchi di informazioni ritenute sensibili, ha evidenziato nel tempo la necessità di regolamentarne l'utilizzo per limitare i potenziali rischi legati alla possibile "mancanza di *privacy*, riservatezza e protezione dei dati per i pazienti e i cittadini" (Directorate-General for *Parliamentary Research Services* (European Parliament) et al. 2022). Infatti, sebbene le pratiche di anonimizzazione dei dati permettano di rimuovere le informazioni sensibili dai dati sanitari, spesso ci si trova a dover gestire dei compromessi: se i dati sono anonimizzati al punto che da non fornire alcuna informazione utile sui pazienti, si rischia di perdere informazioni fondamentali per il modello; viceversa, quando l'utilità dei dati - e quindi le informazioni contenute al loro interno - è alta, cresce il rischio di reidentificazione (Sepas et al. 2022).

[12] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research.

Il problema della *privacy*, però, non riguarda solo i dati sanitari. Un altro caso che ha sollevato una notevole polemica riguardo la raccolta e la gestione dei dati per le attività di addestramento di algoritmi di IA ha riguardato *Bard*, il *chatbot* dotato di IA realizzato da Google. Anche questo modello, come *ChatGPT* è dotato di IA generativa - ossia IA che fa uso di modelli di apprendimento automatico per creare risultati completamente nuovi a partire dalla combinazione di un *dataset* di addestramento - e, proprio come il suo *competitor* realizzato da OpenAI, sebbene sia in grado di replicare in maniera "realistica", spesso sbaglia nel contenuto del risultato fornito. La polemica è stata sollevata nel 2023 da un *tweet*(Crawford 2023) della ricercatrice Kate Crawford, la quale riportava la risposta fornita dal modello alla domanda "Da dove vengono i *dataset* di *Bard*?". Il *software*, infatti, indicava che tra le sorgenti dei dati ci sarebbero anche alcuni prodotti di Google, come Gmail - uno dei *client* di posta elettronica più popolari al mondo. La risposta, che non è stata ben accolta dai vari utenti di *Gmail*, è stata subito smentita dal profilo Twitter ufficiale di *Google Workspace*, il quale ha anche sottolineato che *Bard* è ancora un esperimento e, in quanto tale, può commettere degli errori.

In questo contesto, a garanzia della *privacy* degli utenti, esistono regolamentazioni disponibili nei vari paesi. In Europa, ad esempio, è in vigore il Regolamento generale sulla protezione dei dati (*General Data ProtectionRegulation*, GDPR), che disciplina la gestione dei dati personali dei cittadini dell'Unione europea. Nel caso specifico dell'Italia, esiste inoltre il Garante della Privacy, ossia un'autorità amministrativa che ha tra i suoi compiti quello di assicurare il rispetto della regolamentazione nel trattamento dei dati personali. Questa figura vigila su piattaforme, servizi e grandi società che hanno a che fare con grosse moli di dati personali e, qualora rilevasse aspetti poco chiari, può arrivare decidere di decretare la limitazione del trattamento dei dati degli utenti italiani nei confronti di una o più società, così come è avvenuto con *ChatGPT* il 31 Marzo 2023[13]. Tale scelta ha portato OpenAI a sospendere temporaneamente il servizio per tutti gli utenti le cui connessioni avvengono dall'Italia e, di conseguenza, ha diviso in due gruppi la comunità digitale: quello degli utenti in accordo con la decisione presa dal Garante, contrari ad un utilizzo poco trasparente e regolamentato dei propri dati personali da parte di OpenAI(Bernieri 2023); e quello ostile a tale soluzione, che avrebbe preferito poter continuare ad utilizzare il servizio e che vede tale blocco come "un blocco al progresso dell'Italia, che viene esclusa dall'utilizzo di questa tecnologia" (Cremonesi 2023).

[13] https://www.agi.it/innovazione/news/2023-04-01/chatgpt_stop_in_italia_openai_garante_privacy-20770966/.

La responsabilità etica dell'IA

Come precedentemente analizzato, l'IA può essere sfruttata per migliorare la qualità della vita delle persone, proteggere l'ambiente e promuovere la giustizia sociale. È in grado di fornire il supporto necessario allo sviluppo di nuove tecnologie che possono essere impiegate in ambiti che vanno dall'intrattenimento, all'industria automobilistica, passando per le applicazioni in campo sanitario.
Tuttavia, una sua implementazione sbagliata, potrebbe perpetuare le disuguaglianze, causare violazioni della *privacy*, permettere la manipolazione delle opinioni, causare danni e, in alcuni casi anche la morte.

Infatti, sebbene uno studio (Luttrell et al. 2015) abbia stimato che se il 90% delle automobili negli Stati Uniti fosse a guida autonoma, ogni anno potrebbero essere salvate 25,000 vite, esistono diversi casi di incidenti mortali che hanno coinvolto tali tipi di veicoli. Nello specifico, l'autorità che regola i trasporti negli Stati Uniti, la *National HighwayTrafficSafety Administration* (NHTSA), ha pubblicato un report[14] all'interno del quale ha segnalato che 392 incidenti avvenuti tra il 1 luglio 2021 e il 15 maggio 2022 hanno coinvolto automobili con tecnologia avanzata di assistenza alla guida, tra i quali ci sono stati 6 decessi.

Il primo caso attestato di un incidente simile, è avvenuto il 18 Marzo 2018 in Arizona, quando una macchina a guida autonoma ha investito e ucciso una donna che stava attraversando a piedi una strada a quattro corsie (Levin et al. 2018). In questa specifica circostanza, Rafaela Vasquez, il *back-up driver* - la persona al posto di guida con la capacità di riprendere il controllo del veicolo in caso di emergenza, per garantire la sicurezza durante la fase di *test* di un'auto a guida autonoma - è stata accusata di omicidio colposo per la sua "negligenza nella funzione di lavoro assegnatale di intervenire in una situazione pericolosa"[15], in quanto è stato documentato che la donna fosse intenta a guardare il programma *The Voice* dal suo smartphone nei minuti precedenti l'incidente, secondo un rapporto della polizia.

Con l'obiettivo di standardizzare le soluzioni di autonomia delle auto, e definire le responsabilità nei vari scenari, la *Society of Automotive Engineers* (SAE) - un ente di normazione nel campo dell'industria aerospaziale, automobilistica e veicolistica - ha creato una scala[16] di sei livelli, specificando e definendo i gradi di autonomia delle vetture, nella quale gli

[14] https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADAS-L2-SGO-Report-June-2022.pdf.
[15] https://www.azcentral.com/story/news/local/tempe-breaking/2018/06/21/uber-self-driving-car-crash-tempe-police-elaine-herzberg/724344002/
[16] https://www.sae.org/site/blog/sae-j3016-update

ultimi due corrispondono a sistemi completamente autonomi, in cui la presenza umana non è neppure necessaria. Proprio in situazioni simili, in cui le macchine possono operare scelte autonome, potrebbe non essere possibile trovare una figura umana direttamente responsabile dell'incidente. In questi casi, qualora ci fossero problemi legati alla realizzazione fisica dell'automobile, sarebbe possibile contestare la responsabilità dei difetti direttamente al costruttore. Similarmente, la responsabilità per danni causati da errori di programmazione tali per cui non è stato previsto ciò che si sarebbe dovuto prevedere, sarebbe imputabile al programmatore (Cappellini 2019). Ma cosa accadrebbe in casi non direttamente legati ai danni da prodotto? Poiché, infatti, durante il suo funzionamento, un'autovettura a guida autonoma può trovarsi in situazioni impreviste, questa deve essere in grado di agire cercando di ridurre al minimo i danni. Per farlo, questi sistemi, devono procedere con delle scelte che sono frutto di valutazioni e impostazioni fornite alla macchina dai progettisti.
Con in mente queste problematiche e al fine di esplorare meglio tale "dilemma morale" nell'ambito della guida autonoma, nel 2018 un gruppo di ricercatori provenienti dal *Massachusetts Institute of Technology* (MIT), da Harvard e da altre università pubblicò un lavoro dal titolo "*The Moral Machine Experiment*" sulla rivista *Nature* (Awad et al. 2018). In questo lavoro, Awad et Al. descrissero il risultato di un sondaggio effettuato su una piattaforma attraverso la quale raccolsero più di 40 milioni di risposte da milioni di persone che vivevano in 233 paesi e territori diversi. Tale sondaggio, prospettava ai vari partecipanti degli scenari in cui gli utenti erano obbligati a effettuare una scelta morale su chi salvare in caso di incidente tra diversi gruppi di persone, come ad esempio tra un gruppo di pedoni e il conducente dell'auto, tra un gruppo di persone anziane e uno di bambini, o tra un gruppo di persone obese e uno di persone magre. Gli autori constatarono che, in generale, le persone preferivano salvare il maggior numero di vite possibile, in qualunque contesto, anche a costo di sacrificare il conducente. Inoltre, lo studio evidenziò alcune differenze culturali e individuali nelle preferenze etiche, ad esempio in base all'età, al sesso e al livello di istruzione.
Risulta, quindi abbastanza evidente, che l'adozione massiccia di sistemi di IA metta a dura prova i concetti di responsabilità tradizionale. Si consideri, ad esempio, uno scenario in cui un paziente si sottoponga ad un intervento chirurgico e subisca un danno a causa di un robot dotato di un algoritmo di IA che lavora in modalità *black-box*. Nel corso del tempo sono state proposte diverse idee per affrontare tale questione. Tra queste, vi è la possibile concessione di uno *status* giuridico al dispositivo e riservargli un patrimonio per consentirgli di risarcire i danni causati, oppure un riconoscimento giuridico parziale in cui il dispositivo di IA si assuma la responsabilità diretta per le azioni dannose e, allo stesso tempo si continuerebbero a considerare i programmatori e gli utenti indirettamente responsabili in base al danno specifico. Un'altra opzione avanzata è quella di implementare un sistema assicurativo obbligatorio simile all'assicurazione di responsabilità civile (Colletti 2021).

In ogni caso, tali proposte condividono il fatto che gli attuali modelli di gestione delle responsabilità non sono preparati a questo tipo di interazione, dato che le norme che li regolano sono state concepite in un'epoca in cui le lesioni e gli errori erano causati da umani o da macchine gestite direttamente dagli umani. Poiché l'integrazione di sistemi di IA può prevedere scenari in cui la lesione avviene senza alcun intervento umano, tali modelli devono essere adeguati. Infatti, così come l'utente è responsabile per un utilizzo improprio di un sistema dotato di AI, i programmatori e i costruttori sono responsabili per dei problemi legati alla progettazione o all'implementazione. Più generalmente, l'obiettivo è quello di riuscire a distribuire in maniera corretta ed equa le responsabilità al punto da garantire la sicurezza e l'efficacia dell'IA, ma allo stesso tempo non agire da barriera all'innovazione e al progresso tecnologico (Parikh et al.2022).
Con lo scopo di garantire un uso etico dell'IA, l'Unione Europea ha sviluppato e pubblicato una serie di linee guida sotto forma di codice etico che includono raccomandazioni per lo sviluppo e l'implementazione di sistemi abilitati all'IA. Successivamente, nel 2021 ha proposto una legge chiamata "*AI Act*", all'interno della quale fornisce una definizione di IA, e presenta un approccio basato sul rischio che divide gli impieghi dell'IA in quattro livelli[17]:

- Rischio minimo o nessun rischio: impiego consentito senza restrizioni, come filtri *anti-spam* o videogiochi dotati di IA;
- Rischio limitato: impiego consentito ma soggetto a obblighi di informazione/trasparenza, come *chatbot*;
- Rischio alto: impiego consentito a condizione che siano rispettati determinati requisiti e si sia ottenuta la valutazione di conformità *ex-ante*, come veicoli a guida autonoma o dispositivi medici;
- Rischio inaccettabile: proibito, come sistemi di *social scoring*.

Tale regolamentazione contiene, inoltre, alcune norme relative all'IA forte, anche conosciuta come *Artificial General Intelligence* (AGI) e rappresenta "la prima iniziativa, a livello mondiale, che fornisce un quadro giuridico per l'Intelligenza Artificiale (IA)"[18]. L'*AI Act* costituisce quindi una solida base per garantire che lo sviluppo dell'IA nell'UE sia etico, legale, socialmente equo e sostenibile dal punto di vista ambientale con lo scopo di contribuire al miglioramento dell'economia, della società e dell'ambiente.
Un altro aspetto sociale, particolarmente importante, e precedentemente accennato, riguarda la possibilità che alcuni sistemi dotati di IA possano sostituire i lavoratori e quindi eliminare numerosi posti di lavoro. Sebbene questa pratica sia comune già dal periodo della prima

[17] https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/.
[18] https://edps.europa.eu/press-publications/press-news/press-releases/2021/artificial-intelligence-act-welcomed-initiative.

dell'industrializzazione e quindi dal momento in cui le macchine hanno permesso di automatizzare alcuni processi, oggi, grazie all'avvento dell'IA, il processo si sta estendendo anche al settore dei servizi, dove ad esempio chatbot e assistenti virtuali riescono a sostituire i "colleghi umani" del servizio clienti (Frank et al. 2019). Questo scenario, rende abbastanza evidente il fatto che la perdita del lavoro non abbia un impatto esclusivamente finanziario sulla popolazione. Spesso, infatti, ciò ha delle conseguenze significative anche sulla salute mentale e sul benessere di una persona, perché l'incertezza, che potrebbe derivare anche dalla possibile difficoltà di reinserimento nel mondo del lavoro, può portare a fenomeni di *stress*, ansia e depressione.

In questo contesto, i governi hanno l'onere di sostenere i lavoratori tramite programmi di sostegno e di riqualificazione, e attraverso la regolamentazione dell'uso dell'IA nei settori in cui la perdita di posti di lavoro potrebbe essere più grave. Infine, anche i singoli individui possono adottare misure proattive per proteggere il proprio futuro professionale, ad esempio sviluppando competenze molto richieste, che possano aiutare il loro reinserimento nel mondo del lavoro.

In questo contesto di forte espansione dell'IA, nel 2023, oltre 50 mila persone hanno firmato una lettera aperta[19] con la quale hanno chiesto una pausa allo sviluppo di sistemi di IA più potenti di GPT-4 - il modello successivo di GPT-3 e, dichiarato da OpenAI molto più potente della versione precedente - al fine di consentire la definizione di regole che garantiscano il loro controllo. Tra i firmatari, sono presenti alcune delle figure di spicco del mondo digitale come il premio *Touring YoshuaBengio*, l'imprenditore ElonMusk e il co-fondatore di *Apple Steve Wozniak*. L'appello esorta tutti i laboratori di IA a fermare lo sviluppo di tali sistemi per almeno sei mesi e chiede ai governi di imporre una moratoria per le aziende che rifiutano di adeguarsi.

Durante questo periodo, gli sforzi dovrebbero essere concentrati sullo sviluppo di protocolli di sicurezza e sistemi di governance dell'IA per garantire che i sistemi di IA siano accurati, sicuri, affidabili ed equi.

[19] https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Conclusioni

Il crescente impiego di sistemi dotati di IA solleva importanti questioni etiche che riguardano alcuni fondamentali valori sociali come la *privacy*, la sicurezza, la responsabilità e l'equità. Infatti, sebbene i benefici che si riescono a trarre dall'integrazione di tale tecnologia nei vari ambiti - dalla medicina alla formazione - siano molteplici, poiché le sue applicazioni possono avere delle conseguenze anche a lungo termine sulla popolazione, è necessario puntare i riflettori sulle modalità di sviluppo e di utilizzo di questi sistemi.

In un contesto simile, infatti gioca un ruolo importante la consapevolezza degli utilizzatori di tali piattaforme poiché, soprattutto in algoritmi che fanno uso di *continuouslearning*, sono responsabili dei *feedback* forniti al sistema. Risulta, oltretutto, essenziale curare con la massima attenzione i processi di raccolta dei dati, al fine di ridurre o eliminare *bias* e discriminazioni di alcun tipo considerando però, allo stesso tempo, fondamentale tutelare gli utenti e i loro dati personali.

È quindi importante che i ricercatori, gli sviluppatori e gli utenti di sistemi dotati di IA collaborino per assicurare che questa tecnologia sia utilizzata in modo responsabile ed etico. Questo richiede, oltre che una maggiore coscienza, anche il continuo miglioramento e il rispetto di regolamenti e normative che permettano di garantire la sicurezza e l'efficacia dell'IA senza però agire da ostacolo al progresso tecnologico.

On the use of smart systems: ethical impacts and liability of artificial intelligence

Nocera G., Aiello G. F., De Vita F.

Key Words: Artificial Intelligence – Ethics – Responsibility

Citation: Nocera G., Aiello G. F., De Vita F. (2023). *On the use of smart systems: ethical impacts and liability of artificial intelligence*

doi.medra.org: 10.48237/LIJ_073

30/06/2023

Licenza Creative Commons Attribuzione 4.0 Internazionale

Funding: None

Competing interests: None

Corresponding author: Nocera G., Aiello G. F., De Vita F.

Abstract: Artificial Intelligence (AI) represents a technological resource that is revolutionizing the way we interact with the surrounding world. Nowadays, there are numerous services that make use of machine learning algorithms to simplify some of the operations we perform daily. In these systems, data plays a crucial role, especially in the training process of any AI model, to the point that it can influence its functioning. By virtue of this, it is essential to pay particular attention to the collection phase of data, in order to build datasets that are free from biases and that ensure a fair and correct functioning of a system. In fact, since such tools can have consequences - even in the long term - on people's lives, the use of distorted data could raise important ethical issues. In such a context, the construction and use of regulations and norms that increase the awareness of the end user are fundamental, ensuring the safety and effectiveness of AI without acting as an obstacle to technological progress. This article outlines the importance of analyzing ethical aspects in the design and development phases of AI-powered systems, with particular attention to the social implications that this technology can have.

Introduction

Artificial Intelligence (AI) is becoming an increasingly important technological resource for humanity. Just in the last decade, thanks to its continuous integration with existing systems, it has revolutionized the world we live in, with the intent to improve people's quality of life. In fact,

[1] Department: Development and Engineering
Company: nabla. Email: gnocera@nabla.studio Orcid: https://orcid.org/0009-0007-9871-0316
[2] Department: Operations Società: nabla Email: giulio@nabla.studio
[3] Department of Engineering. University of Messina. Email: fdevita@unime.it
Orcid: https://orcid.org/0000-0002-6709-8001

despite not always being obvious, many of the applications and services used daily by millions of users rely on AI algorithms.
Search engines, such as Google, use this technology to provide users with a better search experience by personalizing results in real-time; virtual assistants like Siri, Alexa, and Google Assistant use machine learning models to understand human language and make several daily activities more natural, such as finding directions or interacting with home appliances; in the entertainment industry, companies like Netflix and Spotify use AI to identify and suggest suitable content for each of their users. Some car manufacturers, such as Tesla and Volvo, use machine learning techniques to improve the safety of transportation and develop self-driving cars. Moreover, AI has proven to be a valuable ally in the medical field, where machine learning algorithms can speed up the diagnosis of some diseases and provide significant support in the search for new treatments. These examples of application demonstrate how AI permeates our lives, often in imperceptible ways, thus improving the quality of services and technologies.
However, the increasing use of AI in everyday services raises important ethical issues since it has a direct impact on people's lives. For example, AI-enabled services could be used to create content that should be the result of human labor, such as research activities that lead to the writing of scientific articles or the resolution of school and university exam exercises.
Furthermore, AI could also be used to make decisions regarding personnel hiring, candidate selection for loan programs, or the detention of individuals suspected of crimes, i.e., in decision-making processes that could have long-term effects on human life. In these latter cases, if the programming rules of AI do not take into account diversity, justice, and non-discrimination, ethical problems may arise. If the model is influenced by prejudices and stereotypes that derive from the dataset used during the learning or training phase, this could lead to unfair and discriminatory implementation. Therefore, researchers and professionals in the field are working to ensure that AI is developed and used ethically and responsibly.
The goal of this work is to deepen the relationship between ethics and AI. In particular, Chapter 2 provides a brief overview of some of the most significant events that have marked the history of AI, from its inception to the present day. Chapter 3 outlines the importance of the ethics of systems that make use of AI, highlighting some cases where the use of such intelligent systems has caused unethical behavior. Subsequently,
in chapter 4, the importance of data in all AI-equipped systems is analyzed, highlighting the problems that a poorly representative dataset can cause if used during training phases. Chapter 5 focuses on how the responsibilities of actions involving an AI system are managed, considering the state of the art and what the current and future needs are. Finally, in chapter 6, a summary of the topics covered is presented, and a possible path to pursue for improving the increasingly present coexistence relationship between humans and so-called intelligent machines is highlighted.

The Artificial Intelligence

In the 1940s, mathematicians and scientists such as Shannon, Wiener, and McCulloch, delving into the study of information theory and cybernetics, marked the beginning of a period of great scientific ferment that prepared the community for the emergence of AI as a scientific discipline in the Fifties.

Since 1950, a series of events related to the development of intelligent machines took place, the main ones of which are given below.

Just in 1950, Claude Shannon described the first computer program capable of playing chess (Shannon 1988) and presented Theseus[1]*, a mechanical mouse capable of exploring a maze, finding the way out, and using the acquired experience to solve cases in which the maze had been modified. In the same year, Isaac Asimov proposed his "I, Robot" (Asimov 2013), in which he discussed the possibility of peaceful coexistence between human beings and autonomous robots through the help of his three laws of robotics, providing the basis for future discussions on the ethics of robots and AI. Alan Turing - considered one of the fathers of modern computing - presented, in his famous article "Computing machinery and intelligence" (Turing 1950), a test according to which a machine could be considered intelligent if its behavior was indistinguishable from that of a person when observed by a human being.*

In the years following Turing's test, the first fields of research related to AI and thinking machines began to take shape, and in 1956 McCarthy introduced the term "artificial intelligence" during the Dartmouth Conference (McCarthy et al. 1955). Some of the most eminent scientists, engineers, mathematicians, and psychologists of the time participated in this conference. Among these, Allen Newell, Cliff Shaw, and Herbert Simon presented the "Logic Theorist" (Newell et al. 1956), the first AI program designed to imitate the problem-solving abilities of a human being. Unlike modern approaches to AI, which are based on the so-called learning from experience, the Logic Theorist was based on "symbolic AI," the dominant branch of AI between the 1950s and 1970s. This approach, based on the manipulation of symbols and rules where knowledge is explicitly encoded, has limitations related to the management of uncertain or imprecise situations and in automatic learning.

Due to the importance of the topics discussed, the Dartmouth seminar is considered the founding event in the history of AI (Crevier 1993). At this conference, in fact, many concepts that are still the subject of research today, such as neural networks and natural language processing, were discussed. In the same period, Arthur Samuel presented the first software capable of playing checkers (Samuel 2000), learning, and adapting its own strategy, thus introducing the concept of reinforcement learning (Sutton 1997). In the early 1960s, the first artificial neural networks were also developed, including Rosenblatt and Pitt's Perceptron algorithm (McCulloch et al. 1943), capable of recognizing images and learning the difference between geometric shapes.

The exciting successes achieved in the 1950s and 1960s gathered huge investments but, at the same time, caused a wave of excessive confidence and unrealistic expectations. And precisely these exaggerated predictions and the gap with the disappointing results obtained in the 1970s and 1980s led to what was subsequently called the "AI Winter" (Crevier 1993). The AI Winter was a period of disappointment and lack of interest that caused a sharp decline in research activities until the 1990s. During these twenty years, there were alternating periods of excitement and disappointment that led researchers to a paradigm shift that occurred towards the end of the 1980s. In particular, the focus was on the fact that humans also learn through trial and error, i.e., through experience, and based on these considerations, interest in artificial neural networks was rekindled. Significant contributions were made by the development of the backpropagation algorithm by Rumelhart, Hinton, and Williams

[1] https://www.youtube.com/watch?v=_9_AEVQ_p74.

(Rumelhart et al. 1986), which allowed the optimization of the training of such networks; the presentation of convolutional neural networks by Le Cun (LeCun et al. 1989), among whose first implementations it is possible to identify the recognition of characters and numbers[2]; and the studies on Deep Learning, which began in 2010 and thanks to Bengio, Hinton, and Le Cun (LeCun et al. 2015), led to algorithms capable of solving increasingly complex problems. At the same time, the advance of miniaturization, the development of more energy-efficient electronic devices, and the expansion of the Internet network, ensured the evolution of what is now called the Internet of Things (IoT), configured as a large distribution of devices capable of collecting and processing data. Starting from the 2000s, AI has experienced exponential growth thanks to advancements in computing power, represented in part by Cloud technology and the availability of large amounts of data in what we call the era of big data. This context has allowed for the development of applications such as virtual assistants, autonomous driving cars, and predictive analysis, enabling AI to be successfully applied in many areas, including medicine, agriculture, finance, and cybersecurity.

For example, in 2011, Google launched a project called Google Brain[3], which uses a multi-layer artificial neural network to improve the capabilities of the eponymous search engine. That same year, IBM's AI system Watson participated in the TV quiz show Jeopardy! (Best 2013), competing against some of the best players of the program and winning, demonstrating that AI could be used not only for research but also to effectively answer complex questions. In 2014, Facebook launched an AI research project called DeepFace, producing a facial recognition algorithm that could identify individuals with 97.35%[4] accuracy. However, current AI is still mainly defined as weak, meaning it is limited to performing specific and well-defined tasks without showing any form of autonomy or generalized intelligence.

At the end of 2015, a group of scientists and entrepreneurs including Elon Musk and Sam Altman founded OpenAI[5], a non-profit company based in San Francisco. Its goal is to collaborate with various organizations and companies, such as Microsoft and Amazon, to promote AI research and development, paying particular attention to the existential risks arising from so-called general artificial intelligence - also known as strong AI, which should be able to emulate human reasoning to solve general problems - making its patents and research open to the public.

Following OpenAI's transformation from a non-profit company to a capped-profit company to attract new capital and collaborations, between 2021 and 2022, applications capable of generating images and entertaining conversations were presented to the public. Specifically, research on conversational technologies led OpenAI to create ChatGPT[6], a tool that uses deep learning to generate responses similarly to how a human would. This model was trained on huge amounts of text data, becoming one of the most advanced conversation-based AI models.

[2] https://www.thats-ai.org/en-GB/units/a-brief-history-of-ai.
[3] https://www.geeksforgeeks.org/what-is-google-brain/.
[4] https://research.facebook.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/.
[5] https://openai.com/.
[6] https://openai.com/blog/chatgpt.

Despite the computing power achieved, many experts believe that strong AI, that is a system capable of reaching and surpassing human intelligence, is still a distant goal. However, as demonstrated by the latest versions of applications like ChatGPT, research and development of new automatic learning techniques are pushing AI closer and closer to this goal.

The AI Ethics

The Treccani Encyclopedia defines ethics as the branch of philosophy that studies "the will and actions of man as a free and rational being, capable of giving himself a set of values and norms to respect [...] and does not only concern the subjective sphere of personal choices but also affects collective life"[7]. In other words, ethics evaluates human behavior and decisions based on fundamental values such as justice, truth, freedom, respect for human dignity, the well-being of individuals and communities, distinguishing them as good, just, and lawful, in contrast to behaviors considered unjust, unlawful, inappropriate, or bad. In general, ethics seeks to answer questions such as "What is right or wrong?" "What should I do?" "How should I behave?" providing a framework of reference to help people make more conscious and responsible decisions.

For this reason, and because AI is a technology that can have significant and lasting impacts on society and human rights, it is important to consider the ethical aspects that concern its relationship with humans. In this context, the so-called "AI ethics" is defined, which focuses on how to design, develop, and use AI in a responsible, fair, and transparent manner, also considering issues related to responsibility and management of AI itself, including:

- *fairness and justice, as AI could, in some cases and through some decisions, perpetuate and amplify social and economic inequalities;*
- *transparency of the choices made by machines, as these decisions are often the result of the execution of complex algorithms;*
- *responsibility for the actions of machines equipped with intelligent algorithms and the definition of liability in the event of damage or errors;*
- *impact on jobs, as on the one hand, mechanical activities are increasingly automated, on the other hand, new sectors requiring advanced technical and management skills are emerging.*

Furthermore, since this technology is becoming more widespread in many sectors, from medicine to finance, to personnel selection activities, it is important to consider the specific circumstance in which AI is used and therefore, the different perspectives and interests at stake. For example, the use of intelligent algorithms in the medical field can raise different ethical issues than those that emerge in finance or recruitment. In fact, it is precisely from the use of AI in everyday decision-making processes that over time, unethical behaviors and, in some cases, ethically incorrect behaviors resulting from intelligent algorithm operations have been identified.

One of the first known cases in which AI showed racial discrimination dates back to 2015 when Google faced criticism for its intelligent image recognition algorithm that "erroneously tagged a black couple as gorillas"[8]. This problem raised numerous concerns about racism in AI and led Google - the

[7] https://www.treccani.it/enciclopedia/etica_(Enciclopedia-dei-ragazzi).
[8] https://www.bbc.com/news/technology-33347866.

development company that created and provided the service - to publicly apologize as the responsible party, as it was the company that had created and published the software. Specifically, the cause of the problem was related to the model's training operations, i.e., the lack of diversity in the data set used to train the machine learning algorithm because the dataset used contained a higher quantity of images of light-skinned people compared to those of people of color.

Another news story of discrimination caused by AI algorithms was reported by Reuters, which stated that Amazon decided to stop a pilot project on a pre-selection system for personnel that made use of AI algorithms due to problems related to discrimination against women and minorities (Vincent 2018). The algorithm, trained on historical data of employees - which were mainly composed of men - tended to systematically downgrade the CVs of female candidates and those with typical ethnic minority names, thus showing a bias in the selection process. However, cases of gender or race discrimination due to AI are not limited to the field of personnel selection and recruitment or image classification. In the financial sector, for example, the launch of the Apple Card - a virtual credit card, integrated with the iPhone's Wallet app, usable for purchases worldwide - using an AI-equipped credit scoring algorithm, resulted in some women receiving lower credit limits than men with the same level of income and credit.

In recent years, moreover, the use of AI algorithms in legal systems is becoming increasingly common. An example is its use in decision-making processes concerning pre-trial detention, which raises numerous concerns about the impartiality of algorithms and the possibility that they may discriminate against certain categories of people. For example, a 2016 study (Mattu et al. 2016) demonstrated how the COMPAS algorithm, adopted to predict the risk of recidivism among detainees, although not using race per se, used many race-related factors and returned higher risk scores for black defendants (Vyas et al. 2020). Furthermore, it is important to note that some AI algorithms work in "black-box" mode, i.e., they produce results without the logic and decision-making process behind the choices being completely clear and transparent. Obviously, this characteristic becomes particularly relevant in the field of justice (Bathaee 2018), where the use of these algorithms has raised concerns about transparency, impartiality, and human rights violations.

In 2016, Microsoft launched Tay, a chatbot - a program that uses AI to simulate a conversation with a user through a chat - on Twitter, whose goal was to learn from user conversations through a technique called continuous learning (Hancell 2023). However, Tay began sending offensive and inappropriate messages due to the influence of some malicious users. This led Microsoft to deactivate the chatbot within 24 hours (Perez 2016). This problem generated concerns about the ability of AI to learn and imitate human behavior, including offensive and discriminatory behavior, highlighting the need for greater regulation and ethical attention in the implementation of AI. In this case, in fact, the model had not been programmed to distinguish acceptable from unacceptable behavior.

Samantha Cole, on the other hand, published an article (Cole 2017) on Motherboard in 2017 highlighting some cases of AI usage for creating well-known fake videos called "Deepfakes". In this AI technique, machine learning procedures based on neural networks are used to create manipulated videos or images that appear real. Specifically, the author explained how the technology was used by a Reddit user called "Deepfakes", who selected the faces of celebrities like Gal Gadot, Maisie Williams, and Taylor Swift and inserted them into the bodies of various porn actresses, resulting in several

obscene videos that appear to be performed by the celebrities. In a similar context, however, the fact that the user only used his computer and a machine learning algorithm, which anyone can download from the internet, made it even more dangerous (Cerdan et al. 2019). This ease of creation raised considerable concerns, as videos created using the "Deepfake" technique are not only linked to the porn industry, but over time, have also been used to create fake political speeches and spread fake news, which can have serious consequences for those involved.

In the medical field, the use of AI algorithms has raised ethical concerns as well. An example is represented by the research work carried out by Raj et al. (Raj et al. 2019), in which the objective was to create machine learning algorithms capable of predicting the mortality of patients in intensive care after traumatic brain injury. The system created was able to provide results with an accuracy of over 80% in predicting mortality at 30 days. In this case, although the algorithm was developed with the objective of improving patient care, this discovery raised concerns about the use of AI technologies for end-of-life decision-making, as such predictions could influence the choice of doctors and patients regarding the care to be provided.

In the social sphere, the spread of intelligent algorithms has elicited different reactions, ranging from the enthusiasm of those who see it as a potential tool to increase efficiency and innovation, to the concern of those who are worried about the possible impact on job loss. In particular, in 2023, the University of Queensland and KPMG Australia conducted a survey (Gillespie et al. 2023) interviewing over 17,000 people from 17 different countries. From here, although it was found that over two-thirds of those interviewed feel optimistic about the benefits that AI can bring to society, 42% (about two out of five people) are concerned that AI could replace jobs in their sector. Today, the number of platforms powered by AI algorithms available online is growing very quickly, and tools like ChatGPT are able to provide substantial support in numerous activities, such as data collection and analysis, source finding, and support in copywriting activities.

However, even though OpenAI CEO Sam Altman has state (Altman 2022) that relying on ChatGPT for important matters is not advisable, as it represents only an embryonic form of progress and there are still many aspects to be developed to ensure its solidity and truthfulness, several users have shown themselves to be extremely enthusiastic about the results already obtained. Some have described on various social networks the steps they have taken to develop smartphone applications without any experience in specific programming languages (Pigford 2023). At the same time, ResumeBuilder.com conducted a survey[9] on a sample of 1,000 American companies, in which about half claimed to have replaced some of their employees with ChatGPT.

It was designed with the goal of conversing simply with all human beings, and in fact, one of its strengths lies in its ease of use and interaction, that is, through a chat where the user can ask the model to answer questions, write code or texts, or even solve debugging problems.

The Generative Pretrained Transformer 3 (GPT-3) - the AI model developed by OpenAI, as the basis of ChatGPT - effectiveness in producing results for most of these tasks allows us to understand the enthusiasm of the community. In addition, the ability to remember past conversations, and its way of correcting what is highlighted as an error, increases the user's temptation to suppose that the machine understands what he or she is writing. However, as pointed out by Floridi - a full professor of

[9] https://www.resumebuilder.com/1-in-4-companies-have-already-replaced-workers-with-chatgpt/.

philosophy and ethics of information at the University of Oxford - in an interview (Marchetti 2023), obviously the tool does not have real intelligence and above all is not capable of understanding but is "simply" able to produce sentences that are probable given the set of information provided during the training phase[10]. It is precisely its operation that is the reason why its outputs can sometimes be inaccurate, untruthful, and sometimes misleading[11]. Furthermore, since its training phase ended at the end of 2021, it is not able to know everything that has happened since. In particular, the model, equipped with 175 billion parameters, was trained on large amounts of data from the Internet and written by human beings, and that is why the answers it provides can seem human.

The Importance of Data in AI

The training of complex AI models, which use deep learning algorithms - an AI technology based on "deep" neural networks, i.e., with multiple layers, and autonomous learning - takes place through the administration of large amounts of data and the exploitation of significant computing power. An example is GPT-3, which has been trained on over 570GB of data (Brown et al. 2020), after filtering a starting dataset containing over 45TB of data. This dataset is composed of data from various sources, such as books, newspapers, scientific articles, patents, online conversations, and so on.

The purpose of the training phase of an AI model is to use the collected data to structure a mathematical model so that it can predict the correct outputs for new inputs that the algorithm has never seen before. In the specific case of GPT-3, the data was used to proceed with supervised training - a type of machine learning approach in which an algorithm is trained on a set of input data for which the corresponding desired output is also provided - whose goal is to allow the model to predict the next word in a given sentence or to generate text autonomously. During training, the model explored the existing relationships between words within the input data, seeking to identify patterns and regularities in the text, which allowed it to learn a wide range of linguistic and cultural knowledge.

Some other examples are deep learning models based on convolutional neural networks and used for image recognition, such as "Amazon Rekognition," developed by Amazon, or "Inception-v3," developed by Google. These algorithms are capable of classifying objects present in photos, detecting human faces or analyzing facial expressions. In this case too, the models' generalization skills - i.e., the ability to accurately recognize a wide range of objects - derive from the fact that they are trained on a large dataset of images, called ImageNet, which contains over 14 million labeled images in more than 20,000 categories, which subsequently allow the model to recognize patterns or elements in input images that the model had not yet seen.

In general, it is precisely the principle of operation of machine learning that requires deep learning models to use large datasets during the training phase. Whether they are used to recognize plant diseases through photos of leaves (De Vita et al. 2020), evaluate a user's creditworthiness (Wang e Xiao 2022), or translate between multiple languages (Johnson et al. 2017), in all these models, data plays a crucial role, so much so that Google's Director of Research, Peter Norvig, wrote in 2009 (Halevy et al. 2009) that "simple machine learning models trained on large amounts of data tend to have higher accuracy than more complex models trained on a smaller amount of data".

[10] https://www.alphagalileo.org/en-gb/Item-Display/ItemId/230890.
[11] https://help.openai.com/en/articles/6783457-chatgpt-general-faq.

However, the quantity of data alone is not enough to achieve good results. Another critical factor for machine learning models to be able to provide excellent results is the quality of the data. For this reason, for an AI model to have a good chance of learning the associations correctly and making precise and generalized predictions, it is essential that the data be of high quality, representative of the problem to be solved, and correctly labeled. In fact, when a model is trained with low-quality data, for example, containing errors, incomplete information, or not representative of the problem to be solved, it presents limitations in creating links between input and output variables, which can lead to inaccurate or even incorrect predictions.

For example, in cases where the data used to train models are collected in a non-uniform manner, contain factors that may be subject to biases, such as race, or cannot be considered a representative sample of society, there is a risk that the results provided by the model will be distorted. Over time, in fact, it has been shown(Grote et al. 2022) how some machine learning models used for medical diagnoses have lower performance when used to predict outcomes that affect women or ethnic minorities. In the research conducted by Obermeyer et al. (Obermeyer et al. 2019), the authors estimated that a calculation error in the system had reduced the number of African American patients who should have been able to participate in healthcare programs by more than half.

Therefore, in order to reduce - or even better avoid - the risk of bias in producing results or generating incorrect or discriminatory responses, the data used for training AI models must be accurate, complete, sufficient in number, and free of biases. This is precisely why a very important phase during the development of AI algorithms is represented by the data analysis, research, or construction phase. In fact, although the number of datasets available online[12] is continuously increasing, depending on the specific implementation to be realized, it may be necessary to build an ad hoc one that takes into account all the peculiarities of the system.

In these cases, an initial analysis of the problem is carried out and the definition of the inputs and outputs of the system is defined (for example, input images and an output number corresponding to a classification, or a group of input numbers and a numerical value output). Subsequently, data collection and selection are carried out that will be provided to the system and, in the event that supervised training is to be carried out, it is essential to "label" the selected data, i.e., associate with these, the corresponding output value that one would like to obtain from the model.

Considering the phases of creation and modeling of the dataset, one of the key aspects of the process is represented by the selection of data sources. In fact, it is appropriate to verify if it is possible to draw on public data already present or if it is necessary to collect new samples. In the latter case, it is necessary to be able to collect the data in reasonable times. In addition, depending on the type of data collected, it is fundamental to emphasize the importance of privacy regarding the information collected, as it is necessary to ensure that such data are used ethically and respecting the rights of users.

An example among all is represented by the healthcare sector. Here, the fact that health data is rich in sensitive information has highlighted, over time, the need to regulate its use to limit the potential risks related to possible "lack of privacy, confidentiality, and data protection for patients and citizens" (Directorate-General for Parliamentary Research Services (European Parliament) et al. 2022). In fact, although anonymization practices allow sensitive information to be removed from health data,

[12] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research.

compromises often have to be managed: if the data are anonymized to the point of not providing any useful information about patients, there is a risk of losing fundamental information for the model; conversely, when the usefulness of the data - and therefore the information contained within them - is high, the risk of re identification increases (Sepas et al. 2022).

The issue of privacy, however, does not only concern healthcare data. Another case that has raised significant controversy regarding the collection and management of data for AI algorithm training activities has to do with Bard, the AI chatbot created by Google. This model, like ChatGPT, is equipped with generative AI - that is, AI that uses machine learning models to create completely new results from a training dataset - and, just like its competitor developed by OpenAI, although it is able to replicate results in a "realistic" manner, it often makes mistakes in the content of the results provided. The controversy was sparked in 2023 by a tweet (Crawford 2023) from researcher Kate Crawford, who reported the model's response to the question "Where do Bard's datasets come from?" The software, in fact, indicated that some of the data sources included Google products, such as Gmail - one of the most popular email clients in the world. The response, which was not well-received by various Gmail users, was immediately debunked by the official Twitter profile of Google Workspace, which also emphasized that Bard is still an experiment and, as such, may make mistakes.

In this context, there are regulations available in various countries to guarantee users' privacy. In Europe, for example, the General Data Protection Regulation (GDPR) is in force, which regulates the management of personal data of European Union citizens. In the specific case of Italy, there is the Privacy Guarantor, an administrative authority whose responsibilities include ensuring compliance with the regulations on the processing of personal data. This authority oversees platforms, services, and large companies that deal with large amounts of personal data and, if it detects unclear aspects, it can decide to decree the limitation of the processing of Italian users' data towards one or more companies, as happened with ChatGPT on March 31, 2023[13]. This decision led OpenAI to temporarily suspend the service for all users connecting from Italy, thus dividing the digital community into two groups: those who agree with the Guarantor's decision, opposing an opaque and unregulated use of their personal data by OpenAI(Bernieri 2023), and those who are hostile to this solution, who would have preferred to continue using the service and see this block as "a block to Italy's progress, which is excluded from using this technology" (Cremonesi 2023).

The EthicalResponsibility of AI

As previously discussed, AI can be used to improve people's quality of life, protect the environment, and promote social justice. It can provide the necessary support for the development of new technologies that can be used in fields ranging from entertainment to the automotive industry, passing through applications in the healthcare field. However, a wrong implementation of AI could perpetuate inequalities, cause privacy violations, allow for the manipulation of opinions, cause damage, and in some cases even death.

Indeed, although a study (Luttrell et al. 2015) estimated that if 90% of cars in the United States were autonomous, 25,000 lives could be saved every year, there are several cases of fatal accidents involving such vehicles. Specifically, the authority that regulates transport in the United States, the

[13] https://www.agi.it/innovazione/news/2023-04-01/chatgpt_stop_in_italia_openai_garante_privacy-20770966/.

National Highway Traffic Safety Administration (NHTSA), has published a report[14] indicating that 392 accidents that occurred between July 1, 2021, and May 15, 2022, involved cars with advanced driver assistance technology, and among them, there were 6 deaths.

The first attested case of a similar incident occurred on March 18, 2018, in Arizona when a self-driving car hit and killed a woman who was crossing a four-lane road (Levin et al. 2018). In this specific circumstance, Rafaela Vasquez, the backup driver - the person in the driver's seat with the ability to regain control of the vehicle in case of an emergency to ensure safety during the testing phase of a self-driving car - was charged with negligent homicide for her "negligence in the performance of her duties in intervening in a dangerous situation[15]", as it was documented that the woman was watching the program "The Voice" in the minutes before the accident, according to a police report.

With the aim of standardizing autonomous car solutions and defining responsibilities in various scenarios, the Society of Automotive Engineers (SAE) – a standardization entity in the aerospace, automotive, and vehicular industries – has created a scale[16] of six levels that specifies and defines the degrees of vehicle autonomy, where the last two correspond to completely autonomous systems, in which human presence is not necessary. In situations where machines can make autonomous choices, it may not be possible to find a human figure directly responsible for the accident. In these cases, if there are problems related to the physical realization of the car, it will be possible to contest the responsibility of the defects directly to the manufacturer. Similarly, responsibility for damage caused by programming errors that did not anticipate what should have been anticipated is attributable to the programmer (Cappellini 2019). But what happens in cases not directly related to product damage?

Indeed, since during its operation, an autonomous car may find itself in unexpected situations, it must be able to act by trying to minimize the damage. To do this, these systems must proceed with choices that are the result of evaluations and settings provided to the machine by designers.

With these issues in mind and to better explore such a "moral dilemma" in the field of autonomous driving, a group of researchers from the Massachusetts Institute of Technology (MIT), Harvard, and other universities published a paper entitled "The Moral Machine Experiment" in the journal Nature (Awad et al. 2018). In this work, Awad et al. described the result of a survey conducted on a platform through which more than 40 million responses were collected from millions of people living in 233 different countries and territories. This survey presented various scenarios in which users were forced to make a moral choice about whom to save in case of an accident involving different groups of people, such as between a group of pedestrians and a car driver, between a group of elderly people and a group of children, or between a group of obese people and a group of thin people, the authors found that people generally preferred to save as many lives as possible, in any context, even at the cost of sacrificing the driver. Furthermore, the study highlighted some cultural and individual differences in ethical preferences, such as age, gender, and education level.

[14] https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-06/ADAS-L2-SGO-Report-June-2022.pdf.

[15] https://www.azcentral.com/story/news/local/tempe-breaking/2018/06/21/uber-self-driving-car-crash-tempe-police-elaine-herzberg/724344002.

[16] https://www.sae.org/site/blog/sae-j3016-update.

It is, therefore, quite evident that the massive adoption of AI systems puts traditional concepts of responsibility to the test. Consider a scenario where a patient undergoes surgery and suffers harm due to a robot equipped with an AI algorithm working in black-box mode. Over time, various ideas have been proposed to address this issue. Among these is the granting of legal status to the device and reserving a heritage to enable it to compensate for the damages caused, or a partial legal recognition of the AI device so that it takes direct responsibility for harmful actions while continuing to consider programmers and users indirectly responsible based on specific damage.
Another option is to implement a mandatory insurance system similar to liability insurance (Colletti 2021). In any case, these proposals share the fact that current models of responsibility management are not prepared for this type of interaction, since the norms that govern them were conceived at a time when injuries and errors were caused by humans or machines directly operated by humans. Since the integration of AI systems can foresee scenarios where injury occurs without any human intervention, such models must be adapted. In fact, just as the user is responsible for the improper use of an AI-enabled system, programmers and builders are responsible for problems related to design or implementation. More generally, the goal is to be able to distribute responsibilities correctly and fairly to the point of guaranteeing the safety and effectiveness of AI while not acting as a barrier to innovation and technological progress (Parikh et al. 2022).

In order to ensure ethical use of AI, the European Union has developed and published a series of guidelines in the form of an ethical code that includes recommendations for the development and implementation of AI-enabled systems. Subsequently, in 2021, it proposed a law called "AI Act," within which it provides a definition of AI, and presents a risk-based approach that divides the uses of AI into four levels[17]:

- *Minimal risk or no risk: allowed use without restrictions, such as spam filters or AI-enabled video games;*
- *Limited risk: allowed use subject to information/transparency obligations, such as chatbots;*
- *High risk: allowed use subject to certain requirements being met and obtaining ex-ante conformity assessment, such as autonomous driving vehicles or medical devices;*
- *Unacceptable risk: prohibited, such as social scoring systems.*

This regulation also contains some rules regarding strong AI, also known as Artificial General Intelligence (AGI), and represents "the first worldwide initiative that provides a legal framework for Artificial Intelligence (AI)[18]". The AI Act, therefore, constitutes a solid foundation for ensuring that the development of AI in the EU is ethical, legal, socially equitable, and environmentally sustainable, with the aim of contributing to the improvement of the economy, society, and the environment.
Another important social aspect, previously mentioned, concerns the possibility that some AI systems may replace workers and therefore eliminate numerous jobs. Although this practice has been common since the period of the first industrialization, when machines allowed some processes to be automated,

[17] https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/.
[18] https://edps.europa.eu/press-publications/press-news/press-releases/2021/artificial-intelligence-act-welcomed-initiative.

today, thanks to the advent of AI, the process is also extending to the service sector, where, for example, chatbots and virtual assistants can replace the "human colleagues" of customer service (Frank et al. 2019). This scenario makes it quite clear that job loss does not have a purely financial impact on the population. Often, in fact, this has significant consequences on a person's mental health and well-being, as uncertainty, which could also result from possible difficulty in reintegrating into the workforce, can lead to stress, anxiety, and depression.

In this context, governments have the responsibility to support workers through support and retraining programs, and through the regulation of the use of AI in sectors where job loss could be more severe. Finally, individuals can also take proactive measures to protect their future careers, such as developing highly sought-after skills that can help them reintegrate into the workforce.

In this era of strong AI expansion, over 50,000 people have signed an open letter[19] calling for a pause in the development of more powerful AI systems than GPT-4 - the next model after GPT-3 and declared by OpenAI to be much more powerful than the previous version - in order to allow the definition of rules that ensure their control. Among the signatories are some of the leading figures in the digital world, such as the Turing Award winner YoshuaBengio, entrepreneur Elon Musk, and Apple co-founder Steve Wozniak. The appeal urges all AI labs to halt the development of such systems for at least six months and calls on governments to impose a moratorium on companies that refuse to comply.

During this period, efforts should be focused on developing AI safety protocols and governance systems to ensure that AI systems are accurate, safe, reliable, and fair.

Conclusions

The growing use of AI-enabled systems raises important ethical questions concerning fundamental social values such as privacy, security, responsibility, and equity. Indeed, although the benefits that can be derived from the integration of such technology in various fields - from medicine to education - are manifold, since its applications can have long-term consequences on the population, it is necessary to focus on the development and use of these systems.

In such a context, the awareness of the users of these platforms plays an important role, especially in algorithms that make use of continuous learning, as they are responsible for providing feedback to the system. Moreover, it is essential to carefully manage the data collection processes, in order to reduce or eliminate biases and discrimination of any kind while at the same time ensuring the protection of users and their personal data.

It is therefore important that AI researchers, developers, and users work together to ensure that this technology is used in a responsible and ethical manner. This requires not only greater awareness, but also continuous improvement and compliance with regulations and norms that allow for the safety and effectiveness of AI without acting as an obstacle to technological progress.

[19] https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Bibliografia

Altman Sam [@sama]. (2022) «ChatGPT Is Incredibly Limited, but Good Enough at Some Things to Create a Misleading Impression of Greatness. It's a Mistake to Be Relying on It for Anything Important Right Now. It's a Preview of Progress; We Have Lots of Work to Do on Robustness and Truthfulness.» Tweet. *Twitter*. https://twitter.com/sama/status/1601731295792414720.

Asimov Isaac. (2013) *I, Robot*. This paperback edition published by HarperVoyager 2013. London: HarperVoyager.

Awad Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, e Iyad Rahwan. (2018) «The Moral Machine Experiment». *Nature* 563 (7729): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

BathaeeYavar. (2018) «The Artificial Intelligence Black Box and the Failure of Intent and Causation» 31.

Bernieri Christian - DPO [@prevenzione]. (2023) «Attenzione… qui abbiamo un problema: il @GPDP_IT fa il suo dovere e pone domande, chiede spiegazioni, vietando trattamenti illeciti di dati personali. Il tutto senza inventare nulla ma applicando leggi ben note a tutti. @OpenAI esclude l'Italia dall'accesso ai suoi servizi,…». Tweet. *Twitter*. https://twitter.com/prevenzione/status/1642034490410246147.

Best Jo. (2013) «IBM Watson: The inside Story of How the Jeopardy-Winning Supercomputer Was Born, and What It Wants to Do Next». TechRepublic. 9 settembre 2013. https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/.

Brown Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. (2020) «Language Models are Few-Shot Learners». arXiv. http://arxiv.org/abs/2005.14165.

Cappellini Alberto. (2019) «MACHINA DELINQUERE NON POTEST? * BREVI APPUNTI SU INTELLIGENZA ARTIFICIALE E RESPONSABILITÀ PENALE».

Cerdan Victor, e Graciela Padilla-Castillo. (2019) «Historia del fake audiovisual: deepfake y la mujer en un imaginario falsificado y perverso». *Historia y Comunicación Social* 24 (settembre). https://doi.org/10.5209/hics.66293.

Cole Samantha. (2017) «AI-Assisted Fake Porn Is Here and We're All Fucked». *Vice* (blog). 11 dicembre 2017. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

Colletti Elisa. (2021) «INTELLIGENZA ARTIFICIALE E ATTIVITÀ SANITARIA. PROFILI GIURIDICI DELL'UTILIZZO DELLA ROBOTICA IN MEDICINA».

Crawford Kate [@katecrawford]. (2023) «Umm, Anyone a Little Concerned That Bard Is Saying Its Training Dataset Includes… Gmail? I'm Assuming That's Flat out Wrong, Otherwise Google Is Crossing Some Serious Legal Boundaries. Https://T.Co/0muhrFeZEA». Tweet. *Twitter*. https://twitter.com/katecrawford/status/1638238113276346368.

Cremonesi Paolo. (2023) «Il Garante della privacy ha bloccato ChatGPT in Italia.» 2023. https://www.linkedin.com/posts/paolocremonesi_il-garante-della-privacy-ha-bloccato-chatgpt-activity-7047826026084626432-HXLO/?utm_source=share&utm_medium=member_desktop.

Crevier Daniel. (1993) *AI: The Tumultuous History of the Search for Artificial Intelligence*.

De Vita Fabrizio, Giorgio Nocera, Dario Bruneo, Valeria Tomaselli, Davide Giacalone, e Sajal K. Das. (2020) «Quantitative Analysis of Deep Leaf: a Plant Disease Detector on the Smart Edge». In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, 49–56. https://doi.org/10.1109/SMARTCOMP50058.2020.00027.

Directorate-General for Parliamentary Research Services (European Parliament), Karim Lekadir, Gianluca Quaglio, Anna Tselioudis Garmendia, e Catherine Gallin. (2022) *Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts*. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2861/568473.

Frank Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, et al. (2019) «Toward understanding the impact of artificial intelligence on labor». *Proceedings of the National Academy of Sciences* 116 (14): 6531–39. https://doi.org/10.1073/pnas.1900949116.

Gillespie Nicole, Steven Lockey, Caitlin Curtis, Javad Pool, e Ali Akbari. (2023) «Trust in Artificial Intelligence: A Global Study». Brisbane, Australia: The University of Queensland; KPMG Australia. https://doi.org/10.14264/00d3c94.

Grote Thomas, Geoff Keeling. (2022) «Enabling Fairness in Healthcare Through Machine Learning». *Ethics and Information Technology* 24 (3): 39. https://doi.org/10.1007/s10676-022-09658-7.

Halevy Alon, Peter Norvig, e Fernando Pereira. (2009) «The Unreasonable Effectiveness of Data». *IEEE Intelligent Systems* 24 (2): 8–12. https://doi.org/10.1109/MIS.2009.36

Hancell Greg. (2023) «(14) Continuous Learning is a must… | LinkedIn». 2023. https://www.linkedin.com/pulse/continuous-learning-must-greg-hancell/.

Johnson Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, et al. (2017) «Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation». arXiv. http://arxiv.org/abs/1611.04558.

LeCun Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, e L. D. Jackel. (1989) «Backpropagation Applied to Handwritten Zip Code Recognition». *Neural Computation* 1 (4): 541–51. https://doi.org/10.1162/neco.1989.1.4.541.

LeCun Yann, Yoshua Bengio, e Geoffrey Hinton. (2015) «Deep Learning». *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Levin Sam, e Julia Carrie Wong. (2018) «Self-Driving Uber Kills Arizona Woman in First Fatal Crash Involving Pedestrian». *The Guardian*, 19 marzo 2018, sez. Technology. https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

Luttrell Kevin, Michael Weaver, e Mitchel Harris. (2015) «The Effect of Autonomous Vehicles on Trauma and Health Care». *Journal of Trauma and Acute Care Surgery* 79 (4): 678–82. https://doi.org/10.1097/TA.0000000000000816.

Marchetti Dario, reg. (2023) *Il miracolo di ChatGPT è scrivere testi che per un umano richiederebbero intelligenza senza averne*. https://www.rainews.it/video/2023/02/il-miracolo-di-chatgpt—scrivere-testi-che-per-un-umano-richiederebbero-intelligenza-senza-averne-b4f858a2-8667-40ef-86ce-8cb4cabb0534.html.

Mattu Julia Angwin, Jeff Larson,Lauren Kirchner,Surya. (2016) «Machine Bias». ProPublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

McCarthy J, M L Minsky, N Rochester, e C E Shannon. (1955) «A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE», agosto.

McCulloch Warren S., e Walter Pitts. (1943) «A Logical Calculus of the Ideas Immanent in Nervous Activity». *The Bulletin of Mathematical Biophysics* 5 (4): 115–33. https://doi.org/10.1007/BF02478259

Newell A., e H. Simon. (1956) «The logic theory machine–A complex information processing system». *IRE Transactions on Information Theory* 2 (3): 61–79. https://doi.org/10.1109/TIT.1956.1056797.

Obermeyer Ziad, Brian Powers, Christine Vogeli, e Sendhil Mullainathan. (2019) «Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations». *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Parikh George Maliha, Ravi B. (2022) «Who Is Liable When AI Kills?» Scientific American. 2022. https://www.scientificamerican.com/article/who-is-liable-when-ai-kills/.

Perez Sarah. (2016) «Microsoft Silences Its New A.I. Bot Tay, after Twitter Users Teach It Racism [Updated]». *TechCrunch* (blog). 24 marzo 2016. https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/.

Pigford Josh [@Shpigford]. (2023) « AI Challenge: I'm Going to Build an IOS Game Using Only GPT-4 (and Probably Midjourney for Some Design Elements). I Have Basically Zero Experience Building IOS Apps. I've Tried to Learn the Traditional Way and Failed Every Time. May the Odds Be Ever in My Favor.» Tweet. *Twitter*. https://twitter.com/Shpigford/status/1637303300671275008.

Raj Rahul, Teemu Luostarinen, Eetu Pursiainen, Jussi P. Posti, Riikka S. K. Takala, Stepani Bendel, Teijo Konttila, e Miikka Korja. (2019) «Machine Learning-Based Dynamic Mortality Prediction after Traumatic Brain Injury». *Scientific Reports* 9 (1): 17672. https://doi.org/10.1038/s41598-019-53889-6.

Rumelhart David E., Geoffrey E. Hinton, e Ronald J. Williams. (1986) «Learning Representations by Back-Propagating Errors». *Nature* 323 (6088): 533–36. https://doi.org/10.1038/323533a0.

Samuel A. L. (2000) «Some studies in machine learning using the game of checkers». *IBM Journal of Research and Development* 44 (1.2): 206–26. https://doi.org/10.1147/rd.441.0206.

Sepas Ali, Ali Haider Bangash, Omar Alraoui, Khaled El Emam, e Alaa El-Hussuna. (2022) «Algorithms to anonymize structured medical and healthcare data: A systematic review». *Frontiers in Bioinformatics* 2. https://www.frontiersin.org/articles/10.3389/fbinf.2022.984807.

Shannon Claude E. (1988) «Programming a Computer for Playing Chess». In *Computer Chess Compendium*, a cura di David Levy, 2–13. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-1968-0_1.

Sutton Richard. (1997) «11.2 Samuel's Checkers Player». 30 maggio 1997. http://incompleteideas.net/book/11/node3.html.

Turing A. M. (1950) «I.—COMPUTING MACHINERY AND INTELLIGENCE». *Mind* LIX (236): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Vincent James. (2018) «Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women». The Verge. 10 ottobre 2018. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report.

Vyas Darshali A., Leo G. Eisenstein, e David S. Jones. (2020) «Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms». A cura di Debra Malina. *New England Journal of Medicine* 383 (9): 874–82. https://doi.org/10.1056/NEJMms2004740.

Wang Chongren, e Zhuoyi Xiao. (2022) «A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer». *Applied Sciences* 12 (21): 10995. https://doi.org/10.3390/app122110995.