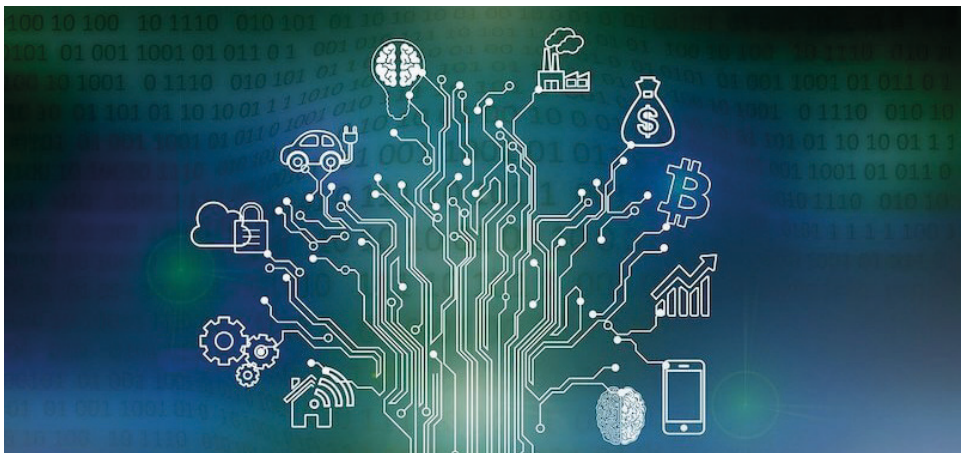
Dobbiamo temere l'Intelligenza Artificiale?

Francesco Marcelloni

The advent of chatGPT has further fueled discussions about the possible dangers of Artificial Intelligence. While everyone recognizes the immense potential of this technology, many also highlight the potential risks it poses to individuals and society. The article explains that some risks are inherent to the machine learning algorithms that underlie the major successes of current Artificial Intelligence, while others arise from its improper use. What are governments doing to protect us? In Europe, the European Parliament has just approved the AI Act, which regulates the development and use of AI-based systems. This is an important initial step that could have significant implications, but only if scientific research develops adequate technical support to ensure that systems comply with the regulations.
Keywords: *Artificial Intelligence*, *Machine Learning*, *EU AI Act*, *Federate Learning*

Il termine "intelligenza artificiale" (IA) venne coniato per la prima volta nel 1956, durante una storica conferenza tenuta al Dartmouth College, negli Stati Uniti[1]. Questa conferenza è universalmente riconosciuta come l'inizio ufficiale dell'IA come disciplina accademica. Gli studiosi che vi parteciparono, tra cui gli organizzatori John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon, sono spesso considerati i pionieri dell'IA. Per anni l'IA è rimasta confinata nell'ambito scientifico, dove diversi ricercatori hanno lavorato allo sviluppo di nuovi algoritmi, metodologie e tecnologie, senza avere la risonanza mediatica a cui stiamo assistendo recentemente: il grande pubblico aveva potuto conoscere l'IA solo attraverso qualche film di fantascienza, dove erano state immaginate le sue enormi potenzialità ma anche evidenziati i suoi possibili rischi, spesso catastrofici. Oggi l'IA è applicata in molteplici settori nella vita quotidiana, quali ad esempio sanità, automazione industriale, assistenza alla guida, riconoscimento vocale e del linguaggio naturale, finanza, e-commerce, sicurezza informatica, insegnamento. In molte di queste applicazioni, l'IA ha ricadute positive sugli individui e sulla società, che tutti possono apprezzare. Per esempio, l'IA è stata utilizzata con successo per migliorare la diagnosi medica, è alla base degli assistenti vocali, dei sistemi di traduzione automatica del testo, dei sistemi di raccomandazione personalizzata, e di molti dei sistemi che vengono utiliz-



1. McCarthy J, Minsk ML, Rochester N, Shannon CE, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955, AI Magazine, volume 27, 4:12-14, 2016, DOI: 10.1609/aimag.v27i4.1904.

zati per aumentare la sicurezza delle auto, solo per citare alcune delle applicazioni di uso comune.
Alla base dell'IA, ci sono gli algoritmi di apprendimento automatico che consentono ai computer di apprendere in modo autonomo dai dati: invece di seguire istruzioni dettagliate per svolgere specifici compiti, questi algoritmi utilizzano i dati storici come esempi per estrarre conoscenza, creare modelli e trarre conclusioni, al fine di fare previsioni o prendere decisioni in situazioni simili a quelle incontrate nel processo di addestramento. Un esempio di apprendimento automatico è lo sviluppo di un sistema per il riconoscimento di oggetti in immagini, che oggi viene tipicamente eseguito con reti neurali convoluzionali (CNN)[2]. Le CNN sono un tipo di architettura di rete neurale artificiale ispirata al modo in cui il cervello umano analizza e identifica le immagini. La CNN viene addestrata con un vasto insieme di immagini contenenti oggetti delle varie categorie, come gatti, cani, automobili, persone, e impara dai dati di addestramento a riconoscere i pattern e le caratteristiche distintive che definiscono ciascuna categoria di oggetti. Una volta addestrata, la CNN può essere utilizzata per riconoscere oggetti in nuove immagini. Ad esempio, quando le viene mostrata un'immagine di un cane, la CNN utilizza ciò che ha imparato durante l'addestramento per identificare l'oggetto come un cane.
Viene spontaneo chiedersi perché solo recentemente, dopo quasi 70 anni dalla sua nascita, l'IA sia diventata così popolare da essere argomento dibattuto non solo tra gli esperti e da spingere diversi paesi nel mondo ad interventi legislativi mirati. Le ragioni sono principalmente cinque:
1) *Lo sviluppo tecnologico*: oggi abbiamo a disposizione grande potenza di calcolo, anche grazie allo sviluppo di architetture hardware specializzate per l'IA, e quasi infinita capacità di memorizzazione che permettono di eseguire algoritmi di apprendimento automatico sempre più complessi utilizzando enormi quantità di esempi;
2) *Nuovi algoritmi di apprendimento automatico*: gli algoritmi utilizzati oggi sono più evoluti e sofisticati rispetto al passato, e sono in grado di addestrare modelli sempre più complessi raggiungendo risultati sempre più precisi;
3) *Disponibilità di grandi moli di dati*: grazie all'esplosione dei social network, delle piattaforme digitali, dell'Internet delle cose che ci permette di raccogliere dati da sensori dislocati ovunque, ecc., oggi possiamo acquisire e memorizzare grandi moli di dati che consentono di eseguire i nuovi algoritmi di apprendimento automatico in modo efficace;
4) *Impatto nella vita quotidiana*: l'IA oggi non è una tecnologia confinata in pochi laboratori di ricerca ma viene usata quotidianamente con un notevole impatto nella vita di tutti i giorni in vari settori: possiamo dire che ormai l'IA è pervasiva;
5) *Grandi investimenti*: l'IA è universalmente considerata una tecnologia dirompente e sta ricevendo investimenti significativi sia nella ricerca che nel trasferimento tecnologico in tutto il mondo da parte di aziende, governi e istituzioni accademiche: questi investimenti stanno permettendo di coinvolgere sempre più ricercatori nello sviluppo delle sue potenzialità.
L'enorme crescita dell'IA e il suo impatto sugli individui e sulla società ha portato diverse nazioni a dotarsi di regolamenti appositi per cercare di limitare i possibili rischi legati al suo utilizzo. Tutti questi regolamenti riconoscono da un lato la rilevanza dell'IA e le sue ricadute positive in molteplici settori applicativi, dall'altro cercano di limitare gli eventuali rischi che possono essere generati dal suo uso. Ma quali possono essere questi rischi? I rischi possono essere intrinseci

2. Lecun Y, Bottou L, Bengio Y, Haffner P, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, volume 86, 11:2278-2324, 1998, DOI: 10.1109/5.726791.

agli algoritmi che vengono utilizzati e\o generati dall'uso improprio dei sistemi di IA.

La prima categoria di rischi è una conseguenza del modo in cui gli algoritmi di apprendimento automatico operano. Come spiegato precedentemente, questi algoritmi apprendono dai dati. Se i dati utilizzati, per esempio, sono influenzati da pregiudizi o distorsioni, ovviamente questi si trasmettono al modello che viene appreso automaticamente. Nella letteratura sono stati riportati vari casi in cui questo problema è stato riscontrato. Nel seguito, vengono discussi due dei casi più famosi. Nel 2018, venne rivelato che il modello sviluppato da Amazon utilizzando tecniche di apprendimento automatico per il reclutamento del personale mostrava un forte pregiudizio di genere[3]. A causa dei dati storici sulle assunzioni precedenti, in cui l'azienda aveva tendenza a favorire candidati di genere maschile per alcuni ruoli tecnici, il modello generato dagli algoritmi di apprendimento automatico aveva acquisito questo pregiudizio e continuava a discriminare le candidature di genere femminile.

Nel 2019 un articolo comparso su Science metteva in evidenza come un algoritmo ampiamente utilizzato dal sistema sanitario statunitense per identificare e aiutare i pazienti con esigenze di salute complesse presentasse una significativa discriminazione razziale: nonostante lo stesso livello di rischio fosse stato assegnato dall'algoritmo, nella realtà i pazienti di colore risultavano notevolmente più malati dei pazienti bianchi[4]. Correggere questa disparità avrebbe aumentato la percentuale di pazienti di colore che avrebbero ricevuto ulteriore assistenza dal 17,7% al 46,5%. Il pregiudizio si manifestava perché l'algoritmo prediceva i costi delle cure anziché la malattia, ma l'accesso disuguale alle cure risultava in meno denaro speso per la cura dei pazienti di colore rispetto a quello speso per i pazienti bianchi, così portando alla discriminazione razziale.

I dati usati nell'apprendimento automatico potrebbero inoltre non rappresentare adeguatamente tutti i possibili scenari in cui il sistema si troverà ad operare. Il sistema potrebbe quindi essere non particolarmente robusto e non rispondere adeguatamente quando l'input che gli viene fornito è molto diverso dagli esempi utilizzati per l'addestramento. Prendiamo in considerazione l'ambito della guida autonoma dove vengono utilizzati sistemi di riconoscimento dei cartelli stradali per adattare la velocità dei veicoli ai limiti imposti nella strada di percorrenza. Questi sistemi sono basati su algoritmi di apprendimento automatico e utilizzano varie immagini di questi cartelli durante la fase di apprendimento. In una ricerca sperimentale pubblicata nel 2020[5] è stato mostrato come questi sistemi possano essere tratti in inganno quando i cartelli sono stati rovinati dall'usura del tempo o manomessi appositamente. Negli esempi mostrati in fig. 1 il limite di velocità è riconosciuto come 85 miglia l'ora invece che 35, che è il limite di velocità imposto nei percorsi urbani negli Stati Uniti. Trovandosi in

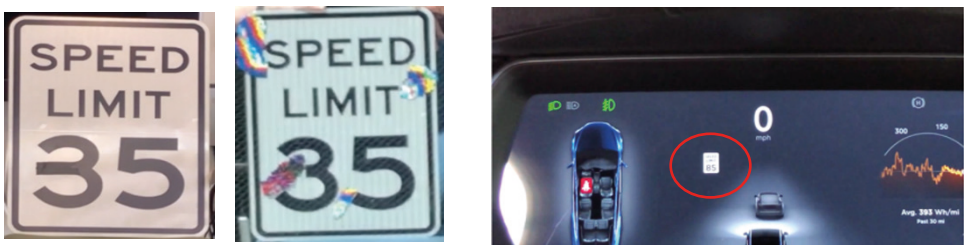


Figura 1. Esempi di cartelli stradali manomessi o usurati e riconoscimento improprio del sistema di visione (la velocità riconosciuta è evidenziata da un cerchio rosso) Gli esempi sono tratti dal riferimento a nota 5.

3. Dastin J, *Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters, 11 Ottobre 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

4. Ziad O, Powers B, Vogeli C, Mullainathan S, *Dissecting racial bias in an algorithm used to manage the health of populations,* Science, Volume 336: 447-453, 2019, DOI:10.1126/science.aax2342

5. Povolny S, *Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles*, McAfee, 2020, https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

ambiente urbano, si può immaginare quali danni possa causare un veicolo a guida autonoma che interpreta non correttamente il limite di velocità. Possiamo evitare che questi problemi si presentino? Da un punto di vista tecnico, ogni sistema di IA viene valutato con degli esempi di test non utilizzati nell'apprendimento per misurare la sua accuratezza in termini di probabilità di commettere un errore. Conoscere questa probabilità, comunque, non ci permette di sapere né quando questo errore si verificherà né quali danni potrà causare. In sistemi critici, quali la guida autonoma, questo può limitare moltissimo l'impiego di sistemi basati sull'apprendimento automatico e richiede controlli adeguati. Normalmente, per poter validare la risposta prodotta da questi sistemi si utilizzano o altri sistemi basati su sorgenti di informazione diverse o la supervisione umana. Nell'esempio precedente, un sistema che utilizza la posizione del veicolo e le mappe con i limiti di velocità può confermare con certezza il risultato prodotto dal sistema di riconoscimento dei cartelli stradali.

La seconda categoria di rischi non è imputabile direttamente agli algoritmi di apprendimento automatico ma dipende da come l'IA viene utilizzata. Le potenzialità dell'IA, che già si sono evidenziate in questi ultimi anni e che nel futuro saranno ancora maggiori con l'evoluzione degli algoritmi e delle infrastrutture di calcolo, aprono scenari che possono effettivamente diventare pericolosi per gli individui e per la società. Per esempio, l'IA può essere utilizzata per analizzare i dati di utenti sui social media e creare contenuti personalizzati o annunci mirati per influenzare e manipolare il comportamento degli utenti. Oppure sistemi di sorveglianza basati sull'IA possono essere utilizzati per monitorare e tracciare le attività delle persone senza il loro consenso, minando la privacy individuale e i diritti umani. Su questi aspetti non può che intervenire la regolamentazione per permettere un uso dell'IA che sia rispettoso degli esseri umani e dei principi che regolano la nostra società. La ricerca scientifica, tuttavia, continua ad avere un ruolo fondamentale anche nella regolamentazione: per riuscire a verificare che i sistemi di IA rispettino le regole, i sistemi devono essere progettati in modo tale che siano trasparenti, cioè siano capaci di spiegare chiaramente come prendono decisioni o producono determinati risultati.

I paesi dell'Unione Europea, come spesso accade quando si parla di diritti, si sono mossi in anticipo rispetto agli altri paesi nel mondo. Il 14 giugno 2023 il Parlamento Europeo ha approvato il regolamento (AI Act) che stabilisce regole armonizzate sull'IA[6]. Se da un lato nella relazione che accompagnava il regolamento viene ribadito che l'uso dell'IA può contribuire al conseguimento di risultati vantaggiosi dal punto di vista sociale e ambientale nonché fornire vantaggi competitivi fondamentali alle imprese e all'economia europea, dall'altro si afferma che il regolamento è particolarmente necessario in settori ad alto impatto, tra i quali figurano quelli dei cambiamenti climatici, dell'ambiente e della sanità, il settore pubblico, la finanza, la mobilità, gli affari interni e l'agricoltura. Gli obiettivi specifici che il quadro normativo proposto si prefigge di raggiungere sono: "1) assicurare che i sistemi di IA immessi sul mercato dell'Unione e utilizzati siano sicuri e rispettino la normativa vigente in materia di diritti fondamentali e valori dell'Unione; 2) assicurare la certezza del diritto per facilitare gli investimenti e l'innovazione nell'IA; 3) migliorare la governance e l'applicazione effettiva della normativa esistente in materia di diritti fondamentali e requisiti di sicurezza applicabili ai sistemi di IA; 4) facilitare lo sviluppo di un mercato unico per applicazioni di IA lecite, sicure e affidabili

6. Parlamento Europeo, AI ACT, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_IT.html

nonché prevenire la frammentazione del mercato". Il Regolamento è l'atto conclusivo di un percorso iniziato qualche anno prima e che aveva visto un passo fondamentale l'8 aprile 2019 con la pubblicazione del documento "Orientamenti Etici per un'IA affidabile" da parte del gruppo di esperti sull'IA istituito dalla Commissione Europea nel 2018[7]. Questo documento raccoglieva più di 500 commenti ricevuti attraverso una consultazione aperta sulla prima versione pubblicata nel 2018.

Il documento promuove un'IA affidabile basata su tre componenti che dovrebbero essere presenti durante l'intero ciclo di vita del sistema: a) legalità, b) eticità e c) robustezza. Nel documento, sono presentate indicazioni per sviluppare sistemi di IA affidabile, proponendo sette requisiti sistemici, individuali e sociali, che dovrebbero essere soddisfatti (fig. 2): 1) intervento e sorveglianza umani; 2) robustezza tecnica e sicurezza; 3) riservatezza e governance dei dati; 4) trasparenza; 5) diversità, non discriminazione ed equità; 6) benessere sociale e ambientale; 7) verificabilità. Gli esempi di sistemi presentati precedentemente ovviamente non avrebbero soddisfatto questi requisiti.

Come sta rispondendo la comunità scientifica italiana alla necessità da un lato di essere all'avanguardia nella ricerca sull'IA e dall'altro di produrre sistemi conformi con l'IA Act e con gli orientamenti etici?

Il PNRR ha finanziato per tre anni a partire dall'inizio del 2023 con 122 milioni di euro il progetto nazionale "Future Artificial Intelligence Research (FAIR)", coordinato dal CNR[8]. FAIR si pone l'obiettivo di "contribuire ad affrontare le domande di ricerca, le metodologie, i modelli, le tecnologie e anche le regole etiche e legali per costruire sistemi di Intelligenza Artificiale capaci di interagire e collaborare con gli umani, di percepire ed agire all'interno di contesti in continua evoluzione, di essere coscienti dei propri limiti e capaci di adattarsi a nuove situazioni, di essere consapevoli dei perimetri di sicurezza e fiducia, e di essere attenti all'impatto ambientale e sociale che la loro realizzazione ed esecuzione può comportare", come descritto nel sito del progetto. L'università di Pisa, cui afferisce l'autore, coordina una (l'unità 1) delle 10 unità operative (denotate spoke) del progetto. In particolare, lo spoke 1 si sta concentrando sulla IA centrata sull'essere umano, che ha l'obiettivo di sviluppare e implementare soluzioni di IA che siano etiche, trasparenti, comprensibili e che migliorino la vita delle persone, senza compromettere i loro diritti, privacy o sicurezza. Lo spoke 1 ha una forte focalizzazione multidisciplinare, con sinergie tra intelligenza artificiale, interazione uomo-computer, scienze cognitive, sistemi complessi, matematica, etica, diritto e scienze sociali. All'interno dello spoke il

Figura 2. I sette requisiti che dovrebbero essere soddisfatti da un sistema di AI secondo il gruppo di esperti sull'IA istituito dalla Commissione Europea nel 2018.

7. Commissione Europea, *Ethics guidelines for trustworthy AI*, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

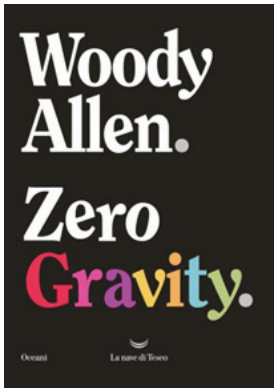
8. *Future Artificial Intelligence Research*, Progetto finanziato dalla Commissione Europea nel programma NextGeneration EU, https://future-ai-research.it/

gruppo di ricerca sull'IA del Dipartimento di Ingegneria dell'Informazione, coordinato dall'autore, sta sviluppando algoritmi di apprendimento federato di modelli trasparenti di IA. L'apprendimento federato consente di addestrare modelli di apprendimento automatico su dispositivi locali o in nodi distribuiti, senza la necessità di inviare i dati grezzi a un server centrale, così preservando la privacy dei dati. L'approccio proposto soddisfa i requisiti 3 e 4 descritti negli Orientamenti Etici per un'IA affidabile e il suo sviluppo è iniziato all'interno del progetto Europeo Hexa-X[9]. Il gruppo è stato selezionato come Key Innovator dall'Innovation Radar della Commissione Europea per questo contributo innovativo. Il gruppo sta anche portando avanti la sua ricerca su modelli di IA trasparenti e apprendimento federato all'interno del progetto FoReLaB (Future-Oriented REsearch LABoratory), finanziato dal Ministero dell'Università e Ricerca (MUR) nell'ambito del bando *Dipartimenti di Eccellenza*[10]. Il progetto si pone l'obiettivo di sviluppare nuove metodologie, paradigmi e tecnologie abilitanti per il nuovo paradigma dell'Industria 5.0.
Concludendo, l'IA appare oggi una tecnologia di cui non possiamo più fare a meno visti gli enormi progressi che ha permesso di raggiungere in vari settori con impatti positivi nella nostra vita di tutti i giorni. Tuttavia, come è stato descritto, l'IA ha dei rischi e il suo utilizzo massivo richiede attenzione e controllo. Se da un lato l'Unione Europea si è mossa già verso la regolamentazione dei sistemi basati sull'IA, in modo da limitarne l'uso improprio, dall'altro la ricerca deve cercare di sviluppare nuovi algoritmi e metodi per consentire di rendere i modelli di IA sviluppati sempre più performanti ma anche più robusti, trasparenti e meno soggetti a pregiudizi e distorsioni presenti nei dati di apprendimento.

9. Merluzzi M et al., *The Hexa-X project vision on Artificial Intelligence and Machine Learning-driven Communication and Computation co-design for 6G*, IEEE Access, Volume 11: 65620-65648, DOI: 10.1109/ACCESS.2023.3287939

10. Progetto FoReLaB, finanziato dal Ministero dell'Università e Ricerca (MUR) nel bando *Dipartimenti* di Eccellenza, http://forelab.unipi.it

Le macchine intelligenti di Woody Allen



Woody Allen, *Zero Gravity*
La nave di Teseo, 2022

Lo scorso anno Woody Allen è tornato a deliziarci con una raccolta dei suoi spiazzanti raccontini. I maligni diranno che sta ripulendo vecchi cassetti riempiti di appunti in gioventù, per la serie "non si butta via niente", ma bisogna concedergli che, oltre a risultare comunque divertente, si tiene molto aggiornato. Un racconto, in particolare, *Quando sul cofano della macchina c'è Nietzsche*, ce lo mostra all'avanguardia in materia di tecnologia e di filosofia. In campo tecnologico il nostro vispo ottantasettenne si misura con l'intelligenza artificiale, niente di meno, applicata in questo caso alle automobili: "non solo le furbe scatolette di latta non hanno più bisogno di essere guidate da qualcuno, ma si parla ormai di vetture con cervelli in grado di compiere scelte morali". Ed eccoci quindi alla filosofia, con la quale Woody Allen non è mai stato tenero. Qui prende di mira l'ultima moda dell'"etica utilitaristica" declinata all'americana, che riduce l'idea di Bentham secondo cui un'azione è buona (cioè utile) se produce il più alto grado di felicità per il maggior numero di persone a un esercizio su bislacchi dilemmi (ammazzo un uomo grasso o due uomini magri?) o a riflessioni sui danni collaterali (non a caso questa materia viene insegnata nelle accademie militari) di improbabili comportamenti.
Così appunto ragiona l'automobile intelligente protagonista del racconto: tiro sotto la vecchietta che attraversa la strada o affronto il rischio di una sbandata che potrebbe danneggiare il mio proprietario? Investo un musicista che suona il violoncello in modo sublime o due balordi gonfi di birra?
Alla fine, destinata ormai alla rottamazione, la colta e intelligente automobile ci congeda con un suggerimento: "la prossima volta che comprate una macchina, diffidate di quelle in grado di discutere di Leibniz o Novalis. Scegliete quelle che hanno più spazio per le gambe e consumano meno".

Maria Turchetto