Affidabilità e Sicurezza dell'Intelligenza Artificiale

Annalisa Appice

Università degli Studi di Bari Aldo Moro

l termine "Intelligenza Artificiale" (IA) fu coniato nel 1955 da John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude E. Shannon per descrivere il progetto pionieristico, poi presentato nella Conferenza di Dartmouth del 1956, di costruire programmi capaci di apprendere, ragionare e risolvere problemi come gli esseri umani. Da allora, dopo un iniziale entusiasmo, l'IA attraversò una fase di oblio negli anni 70 e 80 con una drastica riduzione dei finanziamenti e dell'interesse per la ricerca in IA. La fiducia verso le potenzialità dell'IA si rinnovò tra gli anni 90 e 2000, quando l'apprendimento automatico (machine learning) iniziò a conquistare un interesse sempre crescente e attrarre finanziamenti anche industriali per lo sviluppo di algoritmi capaci di imparare dai dati. La diffusione di Internet, la disponibilità dei Big Data e lo sviluppo di una potenza di calcolo crescente e sempre più accessibile ha nell'ultimo decennio inaugurato l'alba della IA moderna guidata dagli algoritmi di apprendimento profondo (deep learning) che ha permesso all'IA di entrare nella vita di ogni giorno. Oggi l'IA è indiscutibilmente una componente determinate della nostra vita quotidiana. Questa accelerazione esponenziale dell'uso dell'IA nella moderna società sta infatti contribuendo a ridefinire gli equilibri sociali, culturali ed economici dell'umanità con una rapidità senza precedenti. In particolare, la velocità e l'impatto del cambiamento introdotto dalla presenza ubiqua dell'IA è tale da rendere necessaria una riflessione sulle effettive conseguenze a lungo termine dell'uso massiccio dell'IA nella società moderna.

L'IA moderna è principalmente predittiva e messa a punto per supportare i sistemi di decisione in diversi ambiti sociali ed economici. Esempi di IA predittiva sono i sistemi sviluppati per il monitoraggio della superficie terrestre, che grazie anche alla libera disponibilità di immagini satellitari ottenute con il programma Copernicus, sono oggi in grado di realizzare un monitoraggio continuo e in larga scala della superficie terrestre. Tecniche di machine learning e deep learning coniugate con la disponibilità di conoscenza pregressa, tipicamente di origine umana, delle etichette die dati collezionati per l'addestramento contribuiscono all'apprendimento di modelli di IA predittiva accurati che permettono di visionare in maniera automatica e in tempo quasi reale ampie zone della superfice terreste alla ricerca di cambiamenti e/o minacce. Tuttavia, l'accuratezza dei modelli di IA predittiva dipende no solo dalla attenzione dedicata all'addestramento del modello di predizione, ma anche dalla cura del dato raccolto per l'addestramento. Tecniche di IA predittiva per il monitoraggio dello stato d salute delle foreste tramite l'analisi di immagini satellitari sono state di recente sintetizzate da un gruppo di ricercatori dell'Università di Bari Aldo Moro nell'ambito del progetto Europeo SWIFTT finanziato dalla comunità Europea con Grant Agreement 101082732. Applicazioni di IA predittiva sono anche realizzate nella cybersicurezza per difendere le infrastrutture digitali pubbliche e private da cyber-attacchi, oltre che scoprire e mitigare in tempo reale cyber-crimini. L'ambito sanitario è un altro importante dominio di applicazione dove l'IA è divenuta strumento tecnologico imprescindibile per lo sviluppo di nuovi sistemi di supporto alla diagnostica medica e pianificazione dei servizi di cura, in grado di coadiuvare, e mai sostituire, il personale

tecnico-sanitario grazie alla capacità di elaborare grandi volumi di dati comprensivi di immagini e note in linguaggio naturale.

In tutti i casi descritti, l'IA non è mai intesa a sostituzione dell'umano, ma come strumento di co-intelligenza teso a una collaborazione attiva tra intelligenza umana e intelligenza macchina, superando il paradigma dell'autonomia totale. Al fine di sviluppare una co-intelligenza, la decisione dell'IA predittiva deve essere spiegata all'umano che ne valuta i gradi di affidabilità e incertezza. La decisione dell'IA deve essere valida e, eventualmente, modificata con una prospettiva di IA Simbiotica secondo cui gli agenti digitali smettono di essere meri esecutori per diventare compagni cognitivi, in grado di pianificare, adattarsi e interagire efficacemente con l'essere umano. Di recente questa visione ha trovato concreta attuazione nello Spoke 6, portato avanti dall'Università degli Studi di Bari Aldo Moro, del progetto FAIR – Future Artificial Intelligence Research, parte del PNRR italiano. Lo Spoke ha, nell'ultimo triennio, portato avanti le principali sfide di ricerca legate a questa nuova forma di collaborazione, evidenziando le condizioni necessarie per garantirne lo sviluppo, l'accettabilità, la sostenibilità e le applicazioni nei contesti reali e limitare l'uso indiscriminato e rischioso dell'IA. La visione di IA Simbiotica si allinea, anche, alla necessità di regolamentare l'uso indiscriminato o rischioso di dati e IA che ha le sue prime reazioni nel Data Act and AI Act, quali strumenti di regolamentazione ma anche presa di coscienza da parte della società in merito ai problemi legati alla affidabilità della IA.

In riferimento, alla affidabilità dell'IA, è divenuto imprescindibile per i sistemi di IA moderna trattare l'incertezza del dato oltre che l'incertezza della decisione. Da qui emerge la necessità di una IA moderna per la quale sia misurabile non solo l'accuratezza e affidabilità della decisione, ma anche la protezione, sicurezza e resilienza, responsabilità e trasparenza, comprensione e modellazione, riservatezza, imparzialità, attraverso la mitigazione dei pregiudizi dannosi. Garantire la sicurezza dell'IA è oggi, anche per effetto del proliferare dell'IA generativa, uno dei pilastri per lo sviluppo di una IA predittiva affidabile e robusta. Se da un lato l'IA generativa fornisce un servizio creativo, dall'altro il suo uso ubiquo nella vita quotidiana rende necessario lo sviluppo di sistemi in grado di riconosce contenuti generati artificialmente e la loro integrazione nella IA predittiva per consentire di proteggere i sistemi di addestramento da casi di avvelenamento delle fonti dati o permettere ai sistemi di IA predittiva di identificare contenuti creati con fine malevole che, se misconosciuti dai sistemi di IA, possono avere conseguenze critiche per la società (per esempio, fake news o malware).

Pertanto, la raccomandazione è oggi quella di puntare verso la creazione di una IA affidabile che è pienamente possibile con una visione di IA Simbiotica. È, quindi, necessario continuare sulla strada della ricerca e sviluppo che, con una mentalità multidisciplinare, porti allo sviluppo di metodi, best practices e linee guida per la creazione di una IA affidabile che permetta di conseguire in maniera sicura il massimo potenziale delle aspettative previste dall'IA nella società di oggi e domani.