Intelligenza Artificiale, Etica e Studi di Genere

Francesca Alessandra Lisi, Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari

Introduzione

L'Intelligenza Artificiale (IA) è l'ultima evoluzione tecnologica che, svolgendo un ruolo fondamentale nella "Quarta Rivoluzione Industriale", sta trasformando non solo l'economia globale ma anche la nostra società. In quanto tale, porta con sé tecnologie cosiddette dirompenti (*disruptive technologies*), ovvero in grado di sostituire un processo, un prodotto o un'abitudine più datati. Per svelare come queste tecnologie possano avere un impatto trasformativo a livello socio-economico e per identificare quali svantaggi questa trasformazione potrebbe comportare, è necessario adottare un approccio responsabile nella ricerca e nell'innovazione relative ad esse. Tale approccio, promosso dalla Commissione Europea ed indicato come *Responsible Research and Innovation* (RRI)[1], si articola in 5 elementi tematici: *Ethics*, *Gender*, *Public Engagement*, *Education* e *Open Access*. In particolare, i temi dell'etica e del genere hanno improntato buona parte della mia indagine scientifica negli ultimi anni, quella che si è concentrata sui seguenti due filoni di ricerca: *machine ethics* e *feminist AI*.

***Machine ethics*, ovvero come riprodurre l'etica nelle macchine**

L'etica nelle macchine si focalizza sulla progettazione di sistemi di IA in grado di prendere decisioni etiche e di comportarsi in modo moralmente accettabile. Qui si va oltre la semplice garanzia della sicurezza, mirando invece ad integrare, implementare e modellare il ragionamento morale all'interno di agenti autonomi, ed assicurando che essi agiscano in conformità con i valori umani.

Gli aspetti etici sono particolarmente sfidanti quando la relazione fra agenti artificiali ed esseri umani è simbiotica. Nel progetto Future AI Research (FAIR)[2], abbiamo intravisto nella robotica assistiva un campo interessante e stimolante di indagine sull'accettabilità della simbiosi umano-IA. I robot assistivi condividono infatti un'area di lavoro e interagiscono direttamente con gli esseri umani. La loro caratteristica distintiva è la capacità di percepire l'ambiente circostante con l'ausilio di sensori e algoritmi di IA, di comunicare con le persone in modo multimodale, di navigare autonomamente e di prendere decisioni in modo indipendente.

Come caso di studio, abbiamo considerato la simbiosi fra un paziente umano adulto in grado di intendere e di volere ed un robot badante a cui è delegato il compito di monitorare l'aderenza alla terapia da parte del paziente. Questo compito si presta ad una analisi più approfondita in quanto solleva problemi etici. Negli scenari plausibili di interazione fra il paziente ed il robot badante si prospetta il seguente dilemma etico: *Il robot badante si rivolge al paziente per ricordargli di assumere una certa pillola all'ora indicata. Il paziente si rifiuta di assumerla. Come deve comportarsi il robot? Deve insistere affinché il paziente assuma la pillola oppure deve desistere, accettando la decisione del paziente?*

Tre dei quattro principi/doveri dell'etica biomedica sono suscettibili di essere soddisfatti o violati in dilemmi di questo tipo: il dovere del *rispetto dell'autonomia* (ovvero il diritto del paziente di prendere autonomamente delle decisioni informate circa la propria salute e le terapie), il dovere della *non maleficenza* (cioè l'obbligo morale da parte del robot di evitare di causare danni al paziente) e il dovere della *beneficenza* (ovvero l'obbligo morale da parte del robot di fare del bene o di agire nel migliore interesse del paziente). La decisione da parte del robot badante non può che essere guidata da una attenta ponderazione delle due opzioni in termini di rischio relativo a danni fisici e/o mentali o ad altri fattori come la violazione della privacy.

[1] RRI: https://europa.eu/eurobarometer/surveys/detail/1096
[2] FAIR: https://fondazione-fair.it/

La realizzazione di agenti artificiali etici come il robot badante comporta l'integrazione di numerose tecnologie di IA che vanno dall'elaborazione del linguaggio naturale all'apprendimento automatico di regole con eccezioni, oltre a strumenti per la valutazione del rischio. La conversazione con il paziente umano deve essere infatti codificata nel linguaggio interno del robot il quale potrà applicarvi le regole che disciplinano il suo comportamento etico sulla base del rischio stimato. Per garantire la trasparenza del suo processo decisionale, il robot deve anche essere in grado di generare una spiegazione comprensibile agli umani.

***Feminist AI*, ovvero come ripensare l'IA in chiave femminista**

L'IA femminista nasce dalla contaminazione dell'IA con gli studi di genere e le teorie femministe. Promuove un approccio alla progettazione, allo sviluppo e all'implementazione di un'IA che privilegia equità, inclusione ed equità, contrastando attivamente i pregiudizi sistemici legati a genere, razza e intersezionalità. Complessivamente ha obiettivi più ambiziosi che vanno ben oltre la soluzione tecnica, in quanto mira a correggere gli squilibri di potere, garantendo che gli algoritmi non perpetuino ingiustizie patriarcali o coloniali.

Recenti studi hanno evidenziato che svariati sistemi di IA utilizzati per supportare i processi decisionali umani (come il sistema COMPAS per la predizione di recidive criminali[3]) possono introiettare, replicare ed amplificare alcuni *bias* diffusi nella società. Un bias è una distorsione nel sistema di conoscenza condivisa nella società, pro o contro qualcosa o qualcuno/a, e meramente basata su stereotipi e pregiudizi. Gli algoritmi di *Machine Learning*, come le persone, sono vulnerabili a queste distorsioni. Tali algoritmi, infatti, vengono addestrati con moltissimi esempi di comportamento input-output in modo che possano generalizzare rispetto agli esempi forniti e sviluppare la capacità di prevedere un output in corrispondenza a un input qualunque.

Affrontare il problema del bias algoritmico attraverso la lente degli studi di genere fa emergere la necessità di un approccio socio-tecnico. I fenomeni discriminatori infatti nascono dall'intreccio di componenti umane e tecnologiche. Ecco perché le soluzioni tecniche non sono sufficienti a risolvere il problema. Abbiamo proposto di applicare la teoria della performatività di genere della filosofa Judith Butler, per illustrare come l'IA funzioni all'interno del tessuto sociale, manifestando relazioni di potere basate sul genere o su altre dimensioni della diversità (per es. la razza) anche in modo intersezionale. La discussione sull'equità algoritmica non dovrebbe limitarsi a considerare il genere (o la razza) come un attributo o una caratteristica sensibile da gestire, ma dovrebbe altresì analizzare anche l'ontologia sociale da cui esso emerge, diventa problematico e produce una relazione asimmetrica tra le persone.

Riferimenti

T.L. Beauchamp, J.F. Childress. Principles of Biomedical Ethics. Oxford University Press, Oxford, 2019.

A. Dyoub, F. A. Lisi, S. Costantini, I. Letteri, G. De Gasperis, C. Gallese: Towards Practical Ethics for AI. Intelligenza Artificiale 20(1): 145-176 (2026)

F. A. Lisi (2023). Ethics and Gender for Responsible Research and Innovation in AI (extended abstract). Proc. of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022): 1-8. https://ceur-ws.org/Vol-3319/invited1.pdf

G. Nino, F. A. Lisi (2024). Rethinking Bias and Fairness in AI through the Lens of Gender Studies. Proc. of the 3rd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming (BEWARE 2024): 6-22. https://ceur-ws.org/Vol-3881/paper3.pdf

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing