Fairness in Neural Networks

Seminario a cura di

Mattia Cerrato

Ricercatore presso

Johannes Gutenberg-Universität di Mainz

---

7 maggio 2021 dalle 12:00 alle 13:00 | **EVENTO ONLINE**

Il **Seminario AI | Fairness in Neural Networks**, tenuto da **Mattia Cerrato** - *Ricercatore presso Johannes Gutenberg-Universität di Mainz* si inserisce in una serie di appuntamenti organizzati **dall'Artificial Intelligence Lab** di **Intesa Sanpaolo Innovation Center** che hanno l'obiettivo di illustrare e diffondere le evoluzioni delle ricerche condotte nell'ambito dell'Intelligenza Artificiale.
Il seminario, a cura di **Intesa Sanpaolo Innovation Center**, è previsto tramite collegamento da remoto (riferimenti e istruzioni a fondo pagina) **venerdì 7 maggio ore 12.00-13.00**.

Abstract: Sono anni tumultuosi per l'Apprendimento Automatico (AA) e l'Intelligenza Artificiale. L'approccio connessionista che prende il nome di "**Reti Neurali Profonde**" ha vissuto una rinascita che lo ha portato a stabilire **nuovi record** nello stato dell'arte di una moltitudine di applicazioni e, sebbene le performance mostrate siano impressionanti, questi modelli sono difficilmente spiegabili sotto diversi punti di vista.

Nell'impossibilità di produrre modelli pienamente spiegabili, la ricerca in AA ha focalizzato i propri sforzi su modelli "non discriminanti". Nuove metodologie e nuove funzioni obiettivo sono state sviluppate per vincolare l'addestramento delle reti in modo da forzare **un comportamento più equo**.

In questo seminario verranno illustrati alcuni approcci allo stato dell'arte in questo campo. Vedremo come assicurare che la rappresentazione "hidden" appresa dalle reti neurali sia invariante rispetto ad alcuni attributi selezionati e ritenuti sensibili (permettendo

quindi di creare modelli che sono fair by design) e illustreremo due modi per **migliorare la spiegabilità di questi modelli** senza comprometterne l'equità.

Bio: **Mattia Cerrato** - *Ricercatore presso l'University of Mainz*, ha conseguito il M.Sc. in informatica presso l'Università degli Studi di Torino nel 2014. Attualmente è un **Ph.D. candidate in computer science** presso l'Università degli Studi di Torino con una tesi dal titolo "Invariance in neural representations with applications in fairness and debiasing ". Durante il dottorato di ricerca è entrato a far parte del gruppo Machine Learning **dell'Università di Torino**. Attualmente è ricercatore presso l'University of Mainz, in Germania, dove continua il proprio lavoro sulla fairness e l'interpretability. È autore di pubblicazioni in prestigiose conferenze sul **Machine Learning** come ad esempio IEEE DSAA e ACM SAC. È **membro del Program Committee** per la Conferenza Europea sul machine learning.

PER PARTECIPARE

COLLEGATI ALLA RIUNIONE WEBEX