Giuseppe Primiero
University of Milan, Italy

Logics for trustworthy AI

The opacity of AI systems based on machine learning methods is well known, and largely discussed in the literature. Making ML systems more transparent is the current task of Explainable AI. While a variety of tools are being developed to this aim, logical methods are playing an increasingly important role, as by their nature they aid in reconstructing models of computations which are surveyable and checkable.

In this lecture I will offer a light survey of some work recently co-authored on logical models for trustworthy AI, which allow to express additional properties such as fairness and accuracy. Such formal methods allow for the post-hoc reconstruction of models for probabilistically behaving computational systems, and as such are feasible for use in the analysis of ML-based systems.

In doing so, I will argue for two main theses: first, trustworthy AI is strongly tied with the understanding of fairness and accuracy; second, the aim of increasing transparency and explainability of AIs essentially requires the combination of data intensive statistical methods with logical tools to help understanding their beahaviours. To conclude, I will draw some considerations on what such analysis tells us about explanation.