Informatics Research
Institute of Toulouse

Seminar

Wednesday 4 October 2023
14h00 – 15h00
UT3 Paul Sabatier, IRIT, Auditorium J. Herbrand

Joao MARQUES SILVA
Team ADRIA, IRIT

A Glimpse of Formal Explainable AI and Refutation of Some XAI Myths

Asbtract: This seminar is being held in the wake of Joao Marques Silva's EurAI Fellow 2023 award:

Abstract: The advances in machine learning (ML) in recent years have been both impressive and far-reaching. However, the deployment of ML models is still impaired by lack of trust on how the best-performing ML models make predictions. The issue of lack of trust is even more acute in the uses of ML models in high-risk and safety-critical domains. eXplainable artificial intelligence (XAI) is at the core of ongoing efforts for delivering trustworthy AI. Unfortunately, XAI is riddled with critical misconceptions, that foster distrust instead of building trust. This talk provides a brief overview of the emerging field of formal explainable AI and highlights some of the most visible misconceptions of informal XAI. The talk also shows how formal methods have been used, both to disprove those misconceptions, but also to devise practically effective alternatives.

05 61 55 65 10
info@irit.fr

www.irit.fr

Organized by
Department IA - Team ADRIA