Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220304

Explainable Machine Learning for Trustworthy AI

Fosca GIANNOTTI [1]
Scuola Normale Superiore, Pisa (Italy)
Information Science and Technology Institute "A. Faedo" of the National Research Council, Pisa (Italy)

Keywords. Explainable AI, Trustworthy AI, Transparency, Machine Learning, Symbolic AI

Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic not only for the lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. The future of AI lies in enabling people to collaborate with machines to solve complex problems. Like any efficient collaboration, this requires good communication, trust, clarity and understanding.

Explainable AI addresses such challenges and for years different AI communities have studied such topic, leading to different definitions, evaluation protocols, motivations, and results. This lecture provides a reasoned introduction to the work of Explainable AI (XAI) to date, and surveys the literature with a focus on machine learning and symbolic AI related approaches. We motivate the needs of XAI in real-world and large-scale application, while presenting state-of-the-art techniques and best practices, as well as discussing the many open challenges.

[1]Corresponding Author: Fosca Giannotti, fosca.giannotti@isti.cnr.it