Special Session on Explainable Artificial Intelligence: Theory and Applications

Alexander Gegov

School of Computing, University of Portsmouth, UK

alexander.gegov@port.ac.uk

This special session will present recent research results in Explainable Artificial Intelligence. The latter has attracted significant attention over the last few years as part of a serious drive towards making Artificial Intelligence more trustworthy to users in terms of its recommendations and decisions. In this context, interpretability of Machine Learning has been at the heart of this drive.

The special session will explore theoretical developments in a wide range of explanation methods for machine learning models such as model specific and model agnostic methods. It will also explore the application of these methods to a wide range of machine learning models such as interpretable and black box models. The models will be evaluated in terms of both accuracy and interpretability.

The special session will feature a discussion on how explanation methods for machine learning models can facilitate a transition towards more advanced Artificial Intelligence. In this respect, topics of interest will be about how explanations can be used for detecting bias and errors in the data as well as changing feature values in machine learning models to achieve more desirable outcomes.