Explainable and Fair Artificial Intelligence (AI) for End-Users with Complex and Social Care Health Needs

Contact person: Dr Anastasia Angelopoulou, agelopa@westminster.ac.uk

The growth of Artificial Intelligence (AI) and Machine Learning (ML) has shown promising performance for many practical problems such as driver assistance, but often regarded by the public with caution, prejudice, cognitive bias and strong scepticism. This is exacerbated by the lack of transparency and understandability of these technologies in the way they function and behave, particularly in unexpected situations (e.g., solvency check, medical diagnosis). Hence, they are regarded as black boxes.

The aim of this PhD is to open up these black boxes (i.e., making decisions clear from a technical point of view) and to make the predictions interpretable and explainable to humans (even for those not having a technical background). To this end, the PhD should address two different scenarios with typically technically inexperienced users, namely, (a) automatic diagnostic for identifying dementia in ageing deaf users of British Sign Language (BSL), and (b) automatic classification of medical data. Besides the fact that progress in both areas is beneficial from a society point of view, these also show typical scenarios, where people need to deal with technology they cannot fully understand.

This PhD should tackle both problems in parallel by building and evaluating a prototypical explanation tool specifically targeting Deep Learning ML-based AI. The tool will provide explanations about the behaviour of deep neural network learning, allowing us to understand and explain the output and, in particular, cases of failure (*e.g., analyse the learning behaviour over time of the hidden parameters such as weights and activation functions*).