Contact: Epaminondas (Epi) Kapetanios - e.kapetanios@herts.ac.uk

Explainable AI (XAI) and Interpretable ML (IML)

Background: Bridging the gap between symbolic computation and cognitive skills of machines, such as learning and understanding, which are traditionally innate to human intelligence, has always been a big challenge in computer and data sciences, particularly when it comes to communicating with, explaining and justifying computational artefacts to end-users. This has been witnessed by attempts to infuse semantics into symbolic computation, with varying successes and failures in the past, since the early stages of the AI emergence (e.g., expert systems), database querying and search (e.g., natural language-based querying), as well as the evolutionary trajectory of the Web (e.g., the Semantic Web).

The resurgence of Artificial Intelligence (AI) and Machine Learning (ML) and their applications in critical domains such as health care, criminal justice, economics, did not tackle this challenge: it has been exacerbated. AI/ML based results and decisions are difficult to be explained and interpreted leading to an increased mistrust of such applications. More details about this topic can be found by our special research topic at the journal Frontiers in AI:

https://www.frontiersin.org/research-topics/20958/explanation-in-human-ai-systems

Within this context, we would welcome Ph.D. applications seeking to contribute to the following research activities. This is not an exhaustive list of proposals.

Topic 1: Explanations via Machine Learning on Source Code

Machine Learning on Source Code is an emerging and exciting domain of research which stands at the crossroad of deep learning, natural language processing, software engineering, and programming languages. Like the notion of Big Data applications, large repositories of programs (e.g., open-source code in GitHub, Bitbucket…) emerged ("Big Code") many of which are being used for machine learning applications. Current Interpretable Machine Learning (IML) approaches, however, are focussing on the significance of features (e.g., attribution problem), at global or local level, in decision making and results. Explaining and justifying results without understanding of the source code does not provide contextual explanations about the quality of the code, the data being used, the algorithms being implemented.

Topic 2: Mining Bias and Unfairness in Data for Decision Support

Bias and fairness have been identified as two key issues in machine learning applications. Despite the fact that the data science and machine learning communities have already identified the problem and offered approaches to mitigate bias and unfairness, e.g., the aequitas audit toolkit, Carnegie Mellon University, http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/, one of the core underlying causes for unfairness is bias in training data. Especially, bias in data is not yet a central topic in data engineering and management research. With an ever-increasing volume of, often heterogeneous, data, e.g., astronomical data, pandemic data, crowdsourcing activities, mining bias and unfairness in Big Data raises new challenges for bias and fairness metrics and identification as well as mitigation methods.

Topic 3: Explainable Recommendations and Search

This research targets natural language-based generation of explanations behind recommendations and search results. Currently, voice-based search interfaces, e.g., Alexa, Siri, Google Assistant, are capable of producing and returning results, however, there is no way to engage in a conversational mode of voice-based interactions with the end-user for explaining search results and recommendations.