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

Explainable AI (XAI) is not enough for novice end-users: *Towards Building Intelligible Explainers*

The rise of increasingly automated systems driven by AI/ML approaches also fuelled the resurgence of the notion of *Explainable* AI (XAI), as well as of *Interpretable* ML (IML), as a response of researchers and practitioners to increasing worries about lack of trust and transparency in intelligent agents and the decisions these may make. To this extent, several algorithmic approaches have been taken, e.g., *Permutation Importance, LIME, SHAP* values, to address this challenge and provide explanations or interpretations about decisions or predictions being made. Despite the recent resurgence of explanation and interpretability in AI, most of the research and practice in this area seems to use the researchers' intuitions of what constitutes a 'good' explanation. Explanation, however, as a concept has a long-standing tradition in *the Philosophy of Science* being closely associated with *Causation* and *Positivism* in *Natural Sciences* as well as with other frameworks of explanations in *Social and Cognitive Sciences*, where empirical studies and experiments are not always possible. The latter is significant in the context of Human – AI interaction, e.g., *Social Robots*, where humans follow different ways in generating and evaluating explanations.

Background: 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, which also seek to provide continuation and connect with a relatively recent experience with automated dementia recognition for British Sign Language users.

Topic 1: *Building Intelligible Explainers via Self-aware Intelligent Systems*

Building intelligible explainers for *intelligent* systems or *agents*, i.e., capable to be understood and comprehended in human-AI interactions, cannot be achieved by applying post-hoc data or behavioural analysis only. It is, therefore, of paramount importance to construct "*intelligent systems*", which are also self-aware and knowledgeable of all its interacting parts contributing to their behaviour and decisions. For instance, *explanandum* (e.g., data and algorithms provenance, training conditions, complexity), *explainee* context (e.g., age, impairments, intentions). ***Explainers***, in turn, should be designed and engineered in such a way that a) they tap into the knowledge base of a self-aware intelligent system to generate partial answers, b) engage in human-computer interactions to create dialogues and conversations as a meaningful thread of follow-up questions and partial answers in providing trustworthy, "good" explanations.

Topic 2: Modelling and evaluating "goodness" of explanations

Modelling and evaluating "goodness" of explanations provided by intelligible explainers has been an elusive task. Answering the question "*how good is an explanation*" depends on many factors such as quality and mode of the *explainer*, complexity of the *explanandum*, background and context of the *explainee*. Based on explanation theories and ontology engineering via the lenses of *critical realism*, a branch of philosophy that distinguishes between the 'real' world and the 'observable' world, we will aspire to set up a theoretical framework of "good explanations". We also aspire to set up a model at the convergence of all three constituent pillars, *explanandum, explainer, explainee*, with all qualitative and quantitative variables adhered to them. To this extent, the model should be a) capable of measuring "goodness" of an explanation or explainer, b) reinforced in its learning by end-user feedback as well as by the parameters about quality of the human (explainee)-computer interaction per se.