THEME: Cybersecurity: Technology and Ethics

PROJECT: Building Interpretable and De-biased AI for use in the legal system

Lead Supervisor: Dr Niall McLaughlin

Supervisory Team:
Dr Niall McLaughlin,
Prof John Morison,
Dr Jésus Martinez-del-Rincon

Primary Location: CSIT

AI and machine learning are increasingly used to aid human decision making, or in some cases to almost entirely replace human decision makers in the legal context. If AI systems are to be tasked with making important decisions, we need to understand those decision-making processes better and develop better tools to enable us to reproduce where possible the behaviour of such systems and understand the circumstances under which they can be trusted.

The most common way to create an AI system today is through a process called supervised learning, where a large corpus of data is used by an algorithm to learn associations between input-output pairs.

There are several problems with this approach to AI:

Firstly, the decision-making process is often opaque to the end-user. From the end-user's perspective the system simply produces a decision, but it is not capable of explaining the reasoning process used to reach that decision. Such ideas are inimical to any judicial process and ways of making the decision-making more visible need to be explored.

Secondly, due to inadequacies in the training dataset, combined with the ability of AI algorithms to find correlations, the decisions made by an AI system may take into account factors that should not be used in the legal decision-making process. This may be reflected in biased decisions made by the system or it at the very least it will be a closed system where new problems can only be addressed on the basis of previous solutions. This can lead to a feedback loop where the AI system reproduces and reinforces biases societal, historical and judicial biases.

In order to use AI in the context of legal decision making both these challenges should be addressed.

This PhD will have two key stages:

- Develop novel methods for explaining the decisions made by a modern AI system based on Deep-Learning neural networks. We will examine the links between interpretable AI and adversarial examples in order to generate human-understandable explanations of the AI's behaviour.

- Examine ways to produce more dynamic legal datasets. This will lead to the development of a novel method for representing data, such that AI systems can use the data for decision making in ways that replicate socially grounded systems with any biases removed, so that a more impartial decision making can be reproduced.

Primary Academic Discipline: Computer Science / Machine Learning