PhD project
Explaining data-driven decisions with legal, ethical or social impact to end users

To be carried out at the Department of Information and Computing Sciences of the faculty of Science, Utrecht University. This is a preliminary announcement; information on the application procedure and deadlines will be posted when available.

What will you do?

This project aims to explain outcomes of data-driven machine-learning applications that support decision-making procedures to end users of such applications, such as lawyers, business people or ordinary citizens. The techniques should apply in contexts where a human decision maker is informed by data-driven algorithms and where the decisions have ethical, legal or societal implications. They should generate explanations for outputs for specific inputs. The generated explanations should be such that the reasons for the output can be understood and critically examined on their quality. The project will especially focus on explaining 'black-box' applications in that it will focus on model-agnostic methods, assuming only access to the training data and the possibility to evaluate a model's output given input data. This will make the explanation methods independent of a model's internal structure. This is important since in many real-life applications the learned models will not be interpretable or accessible, for instance, when the model is learned by deep learning or when the application is proprietary.

Why is it important for Hybrid Intelligence?

The 'big-data' revolution in AI has deep implications for many aspects of society. Companies increasingly use big data for making predictions (e.g. will this customer leave us for a new business?) or taking decisions (e.g. hiring personnel) that are relevant to their business. In the US criminal courts increasingly use machine-learning algorithms that predict the chance of recidivism by suspects requesting bail. Machine-learning applications that predict outcomes of legal cases are increasingly used by law firms. Government agencies increasingly use big data algorithms for monitoring citizens (e.g. for fraud detection) or in taking decisions about them.

While the use of big data in society, the law and by governments may have many benefits, such as more consistent and more informed decision making, a major concern is the non-transparent, black-box nature of many big-data algorithms for decision support, which prevents a critical examination of the legal or moral quality of predictions and decisions. This concern is felt in many areas of AI and has triggered a call for 'explainable' AI. In the law the problem is particularly urgent, since receiving a justification of a legally relevant decision is a fundamental right of citizens, embodied in many legal procedures and regulatory frameworks.

How will you do it?

The project will explore the use of existing and novel techniques of formal and computational argumentation. It will in particular study techniques for case-based argumentation, which model how decision-makers draw analogies to past cases (precedents) and discuss their relevant similarities and differences. A case-based approach is natural since the training data of (supervised) machine-learning applications can be seen as decided past cases, i.e., as precedents. An argumentation-based approach is important since the focus is not just on *understanding* outputs but also on critically examining their quality.

Three aspects of increasing levels of complexity will be studied:

1. How can argumentation-based explanations be generated in terms of only the training data?
2. How can such explanations be enriched by utilising available domain knowledge (e.g. provided by knowledge engineers when designing the explanation system or by users when interacting with it)?
3. How can explanation be modelled as a two-way interaction between the system and its user?

The project will apply a method suitable for design sciences. It will iteratively go through the following phases:

1. Develop algorithms based on the literature and ideas developed in the project.
2. Design evaluation criteria for their performance.
3. Explore their properties with simulation experiments and (if possible) formal proof
4. Validate their explanatory quality with user experiments

We are looking for….

… candidates with an Msc in AI, computer science, mathematics, data science or a related field, and with a strong interest in interdisciplinary AI research that combines the AI subfields of machine learning, data science, knowledge representation & reasoning and human-computer interaction. Moreover, the candidate should have a commitment to developing computational tools and techniques for helping people make better decisions. The candidate should be able to integrate various research methods and tools, such as formal methods, designing and implementing algorithms and experimental evaluation.

For more information….

… please contact Prof.dr. H. (Henry) Prakken, email: h.prakken@uu.nl