Project Sketch for Horizon-CL4-2021-Human-01-24

A big problem with AI-enabled decision-making is the complexity of two interrelated parts of a socio-technical system containing the AI: the difficulty of testing a technical solution with multiple moving parts, and the impossibility of predicting how human sub-groups will interact with it. Combined with the scale and speed of AI-based systems, there is justifiable apprehension that deploying AI-based solutions could have a deleterious impact on large sections of the population, especially those that are marginalized and are not usually parts of focus groups or requirements specifications. We propose a research project that evaluates both, the technical as well as social, aspects of any AI-based solution. The development team should be able to predict properties relating to fairness as well as possible interaction effects between individuals and social sub-groups that interact with it. This would allow the design and development teams to incorporate system fixes before it is ever used in the wild.

Our key areas of interest for HORIZON-CL4-2021-HUMAN-01-24 revolve around the development of robust methodologies for the “testing” of AI techniques for sociocultural and other biases. We would expect to develop a unit-testing-like methodology for the quantification and analysis of a wide range of biases in analytical pipelines. Thus, UCD’s main contribution would be in towards expectation 1 and 2 of the call text, i.e.

1. Develop technologies and algorithms to evaluate and address bias in AI-based systems; and
2. Develop standardized processes to assess and quantify the trustworthiness of the developed AI systems, in particular assessment of bias, diversity, non-discrimination and intersectionality – based on different types of bias measures.

Specifically, we would contribute towards the development of a scalable testing framework encompassing notions of “ethical”, “fair” and “inclusive” AI. This would take the role of stub-testing ML systems, i.e. before full design and development of the system. Thus, we would afford practitioners an early warning system which would flag need for redesign early in the system development phase and before actual injustice/harm is done. We would envisage this within a wider methodology of exploratory (fairness / bias) analysis of AI systems to enable practitioners to “stress test” their applications of these technologies; largely aligned with expected outcomes 1 and 4.

In terms of competencies, we offer a consortium the following skill set:

- Development of a systematic methods to predict potential impacts of AI/ML based system
- Create reproducible, explainable impact scenarios
- Instrumentation of ML (Evaluation) Platforms
- Parallel and distributed computational environments
- Multi-agent systems and agent-based simulation
- Expertise in a wide range of potential use cases, including Social Media, Computational Finance, Text Analytics.

Researcher Profiles:

Simon Caton - https://people.ucd.ie/simon.caton/ [simon.caton@ucd.ie]

https://scholar.google.com/citations?user=yr8skEQAAAAJ

Vivek Nallur - https://people.ucd.ie/vivek.nallur [vivek.nallur@ucd.ie]

https://scholar.google.com/citations?user=TeqJsjcAAAAJ