2018-00648 - Explainability and fairness of decision support algorithms [PhD Campaign]

Level of qualifications required: Graduate degree or equivalent

Fonction: PhD Position

About the research centre or Inria department

Inria, the French national institute for research in computer science and control, is dedicated to fundamental and applied research in information and communication science and technology (ICST). Inria has a workforce of 3,800 people working throughout its eight research centers established in seven regions of France.

Grenoble is the capital city of the French Alps. Combining the urban life-style of southern France with a unique mountain setting, it is ideally situated for outdoor activities. The Grenoble area is today an important centre of industry and science (second largest in France). Dedicated to an ambitious policy in the arts, the city is host to numerous cultural institutions. With 60,000 students (including 6,000 foreign students), Grenoble is the third largest student area in France.

Assignment

Decision support algorithms are increasingly used in many areas, including credit, healthcare, justice, police, information, e-commerce, with strong impacts on individuals, business and society. They can bring significant benefits and foster the development of a variety of new services but they also raise issues in terms of discrimination, fairness, loss of autonomy, manipulation, etc. Explainability is often presented as a way to address these issues but it is generally not defined precisely and difficult to implement.

Main activities

The goals of this PhD project are to:

- Identify the different types of explanations and their suitability in different situations (e.g. for professionals such as doctors or judges, or laypersons such as litigants, bank customers or users of internet services). Explanations can be either global (about the logic of the algorithm) or local (about specific results). They can take a variety of forms including simple models (decision trees, rules), annotations, typical examples, counterexamples, verification of specific properties (e.g. non-discrimination), etc.
- Define an interactive explanation framework allowing users to query algorithmic systems using an explanation language and successively refine their understanding of the issues of concern for them (e.g. why they have been denied a benefit, or what actions they should take to get better chances to get a loan, etc.).
- Implement and experiment the explanation system in specific contexts for professionals (e.g. in the medical sector, to better understand the factors influencing a diagnosis or treatment recommendation) and/or laypersons (e.g. to be in a better position to detect manipulations).

Références bibliographiques:

- https://www.fatml.org/resources/relevant-scholarship
- https://dl.acm.org/citation.cfm?id=230561

Keywords:

privacy
transparency
regulation
law
Skills
Some knowledge in algorithmics, statistics or artificial intelligence is required (not necessarily in all these areas)

Benefits package
- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Flexible working hours
- Sports facilities

Remuneration
Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included).