Offer #2024-07919

PhD Position F/M AI model audit

Contract type: Fixed-term contract
Level of qualifications required: Graduate degree or equivalent
Function: PhD Position
Level of experience: Recently graduated

About the research centre or Inria department

The Inria Saclay-Île-de-France Research Centre was established in 2008. It has developed as part of the Saclay site in partnership with Paris-Saclay University and with the Institut Polytechnique de Paris since 2021.

The centre has 39 project teams, 27 of which operate jointly with Paris-Saclay University and the Institut Polytechnique de Paris. Its activities occupy over 600 scientists and research and innovation support staff, including 54 different nationalities.

Context

Context: Rich and complex AI systems are increasingly used across multiple actors of society, from large language models to diagnostic models. Stakeholders generally ask for these systems to be audited, with initiatives such as the AI safety institute and many questions from regulatory bodies. However, there is a mismatch between the probabilistic objects of modern AI and the desired safety warrants: certitudes at the level of individual cases. From an engineering standpoint, in many complex settings, answers are best expressed with uncertainty quantification. From an audit standpoint, this quantification needs to be evaluated, both in its own right and insofar as it is linked to decision-making. This control of uncertainty is recognized as one of the main challenges of machine learning in high-stakes applications such as healthcare [11].

One challenge lies in the fact that individual probabilities are never directly observed; instead, only discrete labels are available. The machine-learning literature has predominantly focused on the concept of “calibration error” [5], which controls the error rate given a confidence score (i.e. a probabilistic output). A calibration error of zero implies that a predictor is neither over-confident nor under-confident. However, this measure being an average control applied across all individuals, it does not preclude the possibility of systematic over-confidence for some individuals and under-confidence for others. Likewise, conformal prediction methods come with certain guaranties on uncertainty, yet the strong results are marginal [9]. “Proper scoring rules” give finer a characterization: these functions fully control errors on individual probabilities via observed samples [4]. Yet, their value does not relate simply to an error rate that can be understood in application terms. Our recent research has delved into the decomposition of these into calibration, grouping, and irreducible errors. We have introduced an estimator for the grouping term, thereby completing existing estimators of calibration error. This approach allows for a comprehensive characterization of errors in probabilistic predictions [6]. Using these tools on large language models reveals cultural biases where the models’ uncertainty is more erroneous for answers about east-Asians than north Americans; biases that can then be partly corrected [2].

Environment: This PhD will take place at Inria Saclay, in the Soda team. Soda is a team of 25 people with 4 PIs doing computational and statistical research, both fundamental and applied, to harness large databases on health and society. Soda also develops core software tools such as scikit-learn. For the application partners of Soda and their corresponding application contexts, model validation and auditing are crucial, and the team has developed expertise on related topics. The two advisors have developed and applied crucial prior work for auditing predictors. Weekly meetings will be organized with the candidate in Inria.

We will also collaborate with stakeholders outside of mathematics and computer science, to make sure that we bring answers that are relevant to the broader society and that we run empirical studies close to important application scenarios. For this, we will work with the workforce of the PEPR Santé Numérique that considers evaluation of medical devices, some of them embarking AI.
Assignment

The goal of this PhD is to provide mathematical results and tools to guide auditing such systems: How to best interrogate a system? Given a set of results, inputs and outputs of the system, what can we extrapolate on other inputs?

The work will strive to move away from population level, enabling answers as tight as possible at the individual level. We will consider statistical results on estimating expected errors and bounds for a black-box AI system given observations of its inputs and outputs. We will also consider procedures to best audit a system. To give a meaningful quantification of the impact of errors, we will consider bridging to decision theory and notions of utility. Finally, we will loop back with stakeholders to define utilities and understand better the non-formal aspects of audits.

Main activities

* Finding and reading related bibliographic material
* Formalizing the problem
* Deriving mathematical results and related algorithms
* Numerical implementation and experimentation
* Writing corresponding publications
Skills

* Good written and spoken English proficiency
* Proficiency in Python, numpy, and scikit-learn
* Knowledge of mathematics, statistics, and computer-science fundamentals of machine learning
* Good statistical background, with machine learning knowledge.
* Curious mindset.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

Minimum remuneration: 2,100 € gross/month

Order of August 29, 2016 setting the remuneration of contract doctoral students

General Information

- Theme/Domain: Optimization, machine learning and statistical methods
  - Statistics (Big data) (BAP E)
- Town/city: Palaiseau
- Inria Center: Centre Inria de Saclay
- Starting date: 2024-09-01
- Duration of contract: 3 years
- Deadline to apply: 2024-08-31

Contacts

- Inria Team: SODA
- PhD Supervisor: Varoquaux Gael / Gael.Varoquaux@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security:
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy:
As part of its diversity policy, all Inria positions are accessible to people with disabilities.