Offer #2024-08308

PhD Position F/M Auditing online AI models

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria center at the University of Rennes is one of eight Inria centers and has more than thirty research teams. The Inria center is a major and recognized player in the field of digital sciences. It is at the heart of a rich ecosystem of R&D and innovation, including highly innovative SMEs, large industrial groups, competitiveness clusters, research and higher education institutions, centers of excellence, and technological research institutes.

Assignment

Context
The widespread use of black-box machine learning models in decision-making systems has created a need for transparency and accountability. However, existing regulations and auditing techniques are insufficient to address the challenges posed by these opaque models.
These challenges include the need for efficient auditing (e.g. the computational tractability of these audits), and the ability to handle continuously evolving models. While active auditing techniques show promise [3], they are limited to low-capacity models [1]. Thus, there is a need for more efficient and scalable auditing techniques to address the challenges posed by these modern black-box models. In addition, the auditing of remote online platforms' models presents further challenges due to their constant evolution. Auditing algorithms that can handle dynamic models and provide practical guidance for regulators are dearly missing.

Main activities

Research questions and objectives
Although passive auditing offers a straightforward approach, its efficiency is limited. This is because it treats each query independently, without considering the potential insights gained from previous responses. This can lead to redundant queries and less efficient use of the audit budget. Active auditing, on the other hand, leverages the information gained from previous queries to select the most informative subsequent queries. This can significantly improve efficiency by focusing on areas of the input space that are most likely to reveal insights about the model's behavior. Active auditing offers potential advantages, but also faces several challenges. It can be computationally expensive, especially for large-scale models, limiting its practical applicability. Additionally, it often requires knowledge of the model's hypothesis space, which may not always be available or accurate. Finally, models can be manipulated to mislead auditors [2], making it difficult to obtain accurate results.
To address the challenges faced by active auditing, future research could explore hybrid approaches that combine passive and active auditing to balance efficiency and robustness.
More efficient active learning algorithms could be developed to reduce their computational

costs. Investigating methods to make active auditing more resistant to model manipulation
is another area of potential research. Furthermore, leveraging model explanations to guide
the selection of queries and improve the interpretability of audit results could also enhance
the effectiveness of auditing techniques for black-box models. By addressing these challenges
and exploring novel combinations, we plan to develop more effective and practical auditing.

Figure 1 provides an overview of the PACMAM project. On the left, the auditor is tasked with auditing the target platform on the right. This platform exploits a model $h_\square$, which belongs to a given hypothesis space H. To perform her audit, the auditor sends requests to the model (blue arrows) and gathers the corresponding model responses (red arrows). The three red locks represent the identified scientific challenges: efficiency, tractability, and
dynamism. Furthermore, the concepts used to address these challenges are associated with
the work packages in which they intervene.

First, we will focus on a parameter that is paramount to audit difficulty: the capacity of the audited model. While for the auditing of low capacity models the state-of-the-art has shown that active approaches offer exponential budget improvements, we have shown in a
work entitled ”Under manipulations, are there AI models harder to audit?” [1] that when auditing high capacity models, active and passive methods perform equally well in terms of
budget. Furthermore, we will investigate the situations between those two extreme cases,
namely by finely relating the target model capacity and its audit difficulty (defined as the audit accuracy achievable under a fixed request budget).

We will then focus on devising tractable active auditing algorithms that exhibit an acceptable computational complexity (supported by the auditor to find good inputs for active auditing). Known active auditing algorithms are so computationally expensive that their current practical applicability is strictly limited to the simplest models. We will propose
three directions for improvements: 1) relaxing optimality request for tractability, 2) changing
assumptions to allow active audits to perform with an acceptable complexity, and 3) making
active audits robust to fair-washing attacks, i.e., a platform might expose a tweaked model
to the auditor (H’), instead of the one in production, to game the audit.

Finally, we will turn our attention to evolving models. Although the state-of-the-art typically considers the audited model as static, regulators need to ensure a constant surveil-
lance of platforms that continuously update their models. In this setting, active auditing approaches would leverage previous audits to efficiently request the model and detect the
evolution of the audited metric. Beyond simple change detection, we seek to devise algo-rithms that can indicate a direction (be it “good” or “bad”) in the evolution of these models,
with regard to reference models hosted by the auditor. Importantly, this work package will
propose a prototype to track the evolution of a chosen modern model in production (in vivo
auditing).

[1] Augustin Godinot, Erwan Le Merrer, Gilles Tr´edan, Camilla Penzo, and Franois Ta¨ıani.
Under manipulations, are some ai models harder to audit? In 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pages 644–664. IEEE, 2024.
[2] Erwan Le Merrer and Gilles Tr´edan. Remote explainability faces the bouncer problem.
In Nature Machine Intelligence 2, 529–539, 2020.
[3] Tom Yan and Chicheng Zhang. Active fairness auditing. In Proceedings of the 39th International Conference on Machine Learning, pages 24929–24962. PMLR.

Skills

- understanding of theoretical foundations of machine learning
- Python coding skills

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 (after 6 months of employment) 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

2100

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-12-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-11-30

Contacts

- **Inria Team :** ARTISHAU
- **PhD Supervisor :**
  Le Merrer Erwan / erwan.le-merrer@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.