Offre n°2025-09201

Post-Doctoral Research Visit F/M From AI audits to AI security: an information gain hierarchy

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD

Niveau de diplôme exigé : Thèse ou équivalent

Fonction : Post-Doctorant

A propos du centre ou de la direction fonctionnelle

The Inria Centre at Rennes University is one of Inria's nine centres and has more than thirty research teams. The Inria Centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Contexte et atouts du poste

AI-based models are now core to a wide range of applications, including highly critical ones. The stakes are considerable for companies or institutions deploying them, as their training amounts to up to a billion dollars (e.g. for the training of ChatGPT). This clearly calls for defending them against attacks, like copy/extraction. In parallel, institutions such as state regulators have to ensure that these models operate according to law, in particular with regards to possible discrimination [1]. Researchers are then tasked to provide algorithms to auditors for assessing important metrics regarding deployed AI-based models. In such black-box audits, where an auditor has no access to the remotely operated model's internals, the goal is to stealthily (i.e. with a few queries only) estimate some metrics, such as fairness [6]. Interestingly, and this has not yet been mentioned in the literature, this audit setup is very close to offensive information gain, from a conceptual standpoint. Indeed, potential attackers are incentivized to try and leak information out of deployed models [4,10]. Motivations range from economic intelligence, obtaining implementation details, to simply avoiding development costs by copying deployed models. An auditor is interested in stealthy model observation [3], to avoid

disrupting the audited model by using too many queries. Identically, an attacker also desire stealthiness, here to avoid being detected and cut off. In particular, auditors generally want to obtain precise property estimation, yet confined to a single feature (e.g. male/female fairness), while attackers aim at having a global picture of the model (for basic copy, or evading some parameter set). Thus, there is an avenue to devise offensive methods in between stealthy audits and global attacks, to try and leak novel model characteristics. The ambition of our group is to bridge the gap between these two critical setups: legal auditing and offensive security, in the domain of modern deployed AI models. From this unique standpoint, and from the body of work in the field of AI auditing, we expect to find new insights for attacking and defending deployed AI models, by finding novel angles. For instance, we proposed a unified way to approach model fingerprinting [2] that is of interest for an auditor to guess which model she is observing on a platform; we conjecture that leveraging such an approach to measure the evolution in time of such a model (does the model changes due to updates?) is of core interest for an attacker, as she can derive what is at play at the company hosting this model. This could provide ground for the attacker for economic intelligence, while leaking some precious information that has to be defended by the attacked company.

Mission confiée

- Research
- Working with Ph.D. students from the group

Principales activités

A striking remark when looking at the current types of attacks on AI models is their quantity and apparent independence (see [10] Fig. 3): each is treated as a separate domain. In addition to this list of attacks, we claim that an audit may be viewed as the leak of a feature from a production model, and must be considered as a potential threat. In that light, clarifications in the relation between these attacks might come from a systematic study of how they relate with regards to the setup they operate in, versus the information gain they permit. We propose to work on a hierarchy of attacks, that will uncover the smallest attacks (in terms of assumptions and scope) and how they might be composed into larger attacks, and so on. This hierarchy will reveal unexplored configurations, where several simple attacks will be combined to build richer attacks. This hierarchy will provide the missing link between audits and AI security, bridging the two in a formal way. The postdoc candidate will leverage algorithmic background, to devise a hierarchy, in a parallel to the Herlihy hierarchy in algorithms. We intend to use the notion of "distinguishability" [14] as a hierarchy backbone (to assess if an attack leaks data permitting strong or weak distinguishability of

models). In particular, the field of "property testing" will be related to this hierarchy.

References

[1] Le Merrer, E., Pons, R., & Tredan, G. (2024). Algorithmic audits of algorithms, and the law. AI and Ethics, 4(4), 1365-1375.
[2] Godinot, A., Le Merrer, E., Penzo, C., Taïani, F., & Tredan, G. (2025). Queries, Representation & Detection: The Next 100 Model Fingerprinting Schemes. In AAAI.
[3] Le Merrer, E., & Tredan, G. (2020) Remote explainability faces the bouncer problem. Nature machine intelligence, 2(9), 529-539.
[4] Maho, T., Furon, T., & Le Merrer, E. (2021). Surfree: a fast surrogate-free black-box attack. In CVPR.
[5] Godinot, A., Le Merrer, E., Tredan, G., Penzo, C., & Taïani, F. (2024). Under manipulations, are some AI models harder to audit?. In IEEE Conference on Secure and Trustworthy Machine Learning.
[6] de Vos, M., Dhasade, A., Garcia Bourrée, J., Kermarrec, A. M., Le Merrer, E., Rottembourg, B., & Tredan, G. (2024). Fairness auditing with multi-agent collaboration. In ECAI.
[7] Le Merrer, E., Perez, P., & Tredan, G. (2020). Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications, 32(13), 9233-9244.
[8] Le Merrer, E., Morgan, B., & Tredan, G. (2021). Setting the record straighter on shadow banning. In INFOCOM.
[9] Maho, T., Furon, T., & Le Merrer, E. (2022). Randomized smoothing under attack: How good is it in practice?. In ICASSP.
[10]Ma et al., « Safety at Scale: A Comprehensive Survey of Large Model Safety». arXiv:2502.05206v3
[11]Yan, T., & Zhang, C. (2022). Active fairness auditing. In ICML.
[12]Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. (2023). “Real attackers don't compute gradients”: bridging the gap between adversarial ml research and practice. In 2023 IEEE conference on secure and trustworthy machine learning.
[13]Fukuchi, K., Hara, S., & Maehara, T. (2020). Faking fairness via stealthily biased sampling. In AAAI.
[14]Attiya, H., & Rajsbaum, S. (2020). Indistinguishability. Communications of the ACM, 63(5), 90-99.
[15]ANSSI (2024). Security recommandations for a generative AI system. ANSSI-PA-102.

Compétences

- Advanced machine learning background, and theory of machine learning
- Python coding skills for experiments (if required)
- A good publication track record is mandatory
- Fluency in English is mandatory

Avantages

- 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

Rémunération

Monthly gross salary amounting to 2788 euros

Informations générales

- **Thème/Domaine :** Optimisation, apprentissage et méthodes statistiques Statistiques (Big data) (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2025-10-01
- **Durée de contrat :** 1 an, 5 mois
- **Date limite pour postuler :** 2025-09-21

Contacts

- **Équipe Inria :** ARTISHAU
- **Recruteur :**
  Le Merrer Erwan / erwan.le-merrer@inria.fr

A propos d'Inria

Inria est l’institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Please submit online : your resume, cover letter and letters of recommendation eventually

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.