# Fairness and Privacy in Machine Learning

Machine Learning is becoming ubiquitous in our everyday lives. It is now used in digital assistants, for medical diagnosis, for autonomous vehicles, .... Its success can be explained by the good performances of learned models, sometimes reaching human-level capabilities. However, simply being accurate is not sufficient if these models are to be largely deployed. Hence, the notion of trustworthiness has to be considered as soon as human are involved in the loop. For example, a model used for medical diagnosis should not be biased against sub-groups of the population and should not leak any personal information about the patients. Among all the trustworthiness notions that have been proposed, fairness and privacy seem to be particularly important as regulations have been proposed in several countries. For example, the so-called 80% rule has been established in the US to avoid hiring practices that could have negative effects on sub-groups of the population. Similarly, the GDPR law in the EU protects the privacy of the citizens by preventing unsolicited collection and processing of personal information.

Fairness and Privacy have been extensively studied as individual constraints. Hence, there exists numerous algorithms to enforce fairness constraints in the literature (Caton and Haas, 2020; Mehrabi et al., 2021). Similarly, many approaches have been proposed to learn models that preserve privacy (Dwork et al., 2014; Liu et al., 2021). However, while structural links between fairness and privacy have been identified early on (Dwork et al., 2012), it is only very recently that the question of obtaining fair and privacy preserving models has gained some traction (Ekstrand et al., 2018; Datta et al., 2018) and that specific approaches able to learn models that are bot fair and privacy preserving have been proposed (Kilbertus et al., 2018; Jagielski et al., 2019; Xu et al., 2019; Alabi, 2019; Mozannar et al., 2020; Tran et al., 2020; Mahdi Khalili et al., 2020). Unfortunately, these approaches are often limited to simple problems (such as binary classification) or might sometime lack theoretical guarantees ensuring that the learned models are trustworthy beyond the training examples.

**Objectives:** The goal of this 6 months internship is to jointly study Fairness and Privacy in Machine Learning. The main objectives are (i) to review some of the existing literature in the field, (ii) to design new algorithms to learn fair, private, and accurate models and (iii) to derive theoretical guarantees on the fairness, privacy, and utility levels of the obtained models.

**Requirements:** Successful candidates should have a solid background in Machine Learning with an interest in Fairness and Privacy. Some knowledge in Statistical Learning Theory would be a plus but is not mandatory. A good understanding of the Python programming language is expected. Finally, proficiency in English is required as it will be one of the main working language.

**Keywords:** Machine Learning, Fairness, Privacy, Learning Theory

**Contact:** The recruited student will be based in the INRIA Lille - Nord Europe research center and will be supervised by Michaël Perrot. The interested students should send an e-mail (in english) with a CV and a Motivation Letter to *michael.perrot@inria.fr*.

## References

Alabi, D. (2019). The cost of a reductions approach to private fair optimization. *arXiv preprint arXiv:1906.09613.*

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053.*

Datta, A., Sen, S., and Tschantz, M. C. (2018). Correspondences between privacy and nondiscrimination: why they should be studied together. *arXiv preprint arXiv:1808.01735.*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Ekstrand, M. D., Joshaghani, R., and Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47. PMLR.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR.

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., and Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Mahdi Khalili, M., Zhang, X., Abroshan, M., and Sojoudi, S. (2020). Improving fairness and privacy in selection problems. *arXiv e-prints*, pages arXiv–2012.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mozannar, H., Ohannessian, M., and Srebro, N. (2020). Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR.

Tran, C., Fioretto, F., and Van Hentenryck, P. (2020). Differentially private and fair deep learning: A lagrangian dual approach. *arXiv preprint arXiv:2009.12562.*

Xu, D., Yuan, S., and Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599.