COMPUTER AND CONTROL ENGINEERING

DAUIN - Machine Unlearning

Funded By	Dipartimento DAUIN
Supervisor	BARALIS ELENA MARIA - elena.baralis@polito.it
Contact	GIOBERGIA FLAVIO - flavio.giobergia@polito.it
Context of the research activity	Machine Unlearning is the task of selectively erasing or modifying previously acquired knowledge from machine learning models. This is particularly relevant nowadays due to the increasing concerns surrounding privacy (e.g. the Right To Be Forgotten required by GDPR) and copyright infringements, as highlighted by recent cases involving Large Language Models. The key goal of this proposal is to propose novel architectures, algorithms and evaluation metrics for Machine Unlearning.
	In recent years, the rapid advancement of machine learning models, particularly Large Language Models (LLMs), has raised significant concerns regarding privacy and intellectual property rights. The need for responsible AI practices has become increasingly evident, driven by legal frameworks such as the General Data Protection Regulation (GDPR) that mandates the Right To Be Forgotten. Additionally, high-profile cases involving LLMs have highlighted the need to address issues related to the unintentional retention of sensitive information and potential copyright infringements. The proposed research activity on Machine Unlearning (MU) aims to tackle these challenges by developing novel techniques to selectively erase or modify previously acquired knowledge from machine learning models. The primary objectives of this research are twofold: first, to explore the current state of the art in MU, and second, to propose innovative architectures, algorithms, and evaluation metrics to enhance the efficacy of the unlearning process. Through these goals, the aim is to contribute to the establishment of ethical and responsible AI practices, ensuring compliance with legal requirements and mitigating the risks associated with unintentional information retention by machine learning models. The workplan for this PhD is structured to comprehensively address the multifaceted challenges of MU. The research will focus on proposing novel architectures and algorithms that facilitate effective unlearning while preserving the model's overall performance. Given the current lack of definitive metrics for MU, part of the research efforts will be focused toward trying to identify more suitable and comprehensive metrics. The research activity progresses from foundational research to the practical

Objectives	implementation, validation and application of MU techniques. An outline of the possible research plan is as follows. - First year The first year will be dedicated to literature review and conceptualization, leading to the formulation of the main research objectives for the rest of the doctorate. This initial phase involves an extensive study of the literature, identifying gaps and shortcomings ? leading to the definition of initial proposals for improvements over state-of-the-art techniques. - Second year Based on the areas of opportunity identified and the preliminary proposals made, the candidate will work on the ideation and implementation of novel architectures and algorithms for MU, with ongoing validation and refinement based on the feedback obtained from experiments and evaluations. -Third year The final year will focus on consolidating the findings and defining the applications of main interest for the output produced. During the second/third year, the candidate will have the opportunity to spend a period of time abroad in a leading research center. Publication venues for this research include leading conferences and journals in the fields of machine learning and artificial intelligence. Key conferences include the conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML), and the International Conference on Representation Learning (ICLR). Additionally, reputable journals such as the Journal of Machine Learning Research (JMLR) and the IEEE Transactions on Neural Networks and Learning Systems will be sought for in-depth dissemination of research contributions.

Skills and competencies for the development of the activity	The candidate should have a strong computer and data science background, in particular for what concerns: - Strong programming skills -- preferably in Python - Thorough understanding of theoretical and applied aspects of machine and deep learning - Fundamentals of Natural Language Processing