COMPUTER AND CONTROL ENGINEERING

Istituto AI4I - Safe by Design Deep Learning: From Uncertainty Quantification to Large Language Models

Funded By	Istituto Italiano sull'Intelligenza Artificiale per l'Industria [P.iva/ CF:13130030011]
Supervisor	PASTOR ELIANA - eliana.pastor@polito.it
Contact	LUCA LAURENTI (luca.laurenti@ai4i.it)
Context of the research activity	The goal of this PhD project is to develop techniques to quantify the robustness of deep learning models and to use these techniques to design state-of-the-art deep learning models that not only are accurate, but also come with quantifiable correctness guarantees. To achieve this, the PhD will explore and devise techniques at the intersection of machine learning, probability theory, control theory, and formal methods.
Objectives	Modern deep learning models have achieved remarkable performance across many domains, from computer vision to natural language processing. However, these models can fail unpredictably when faced with small perturbations in input data and lack robustness. This lack of robustness is preventing their use in various applications, such as autonomous cars or medical robots, where safety is paramount and a failure of the model can have fatal or costly consequences. The safe-by-design paradigm seeks to address these limitations by embedding safety and reliability directly into the model's architecture and training process. The overall goal of this project is to contribute to addressing this problem by developing safe-by-design machine learning models that not only are accurate but also come with quantifiable safety and robustness guarantees. To achieve this ambitious goal, the problem will be executed in three main phases. During the first year, the candidate will conduct a study of uncertainty quantification and robustness techniques, exploring their applicability to quantify the safety of modern deep learning architectures. This phase will include reviewing techniques such as randomised smoothing and Bayesian neural networks, and how their use can be employed to design robust models. The second year will focus on the design of machine learning models that are safe-by-design. The candidate will develop new training algorithms and architectures that integrate uncertainty-based constraints directly into the

	learning process, using the techniques explored in the first phase. The proposed models and architectures will be evaluated on standard benchmarks for safety-critical AI tasks. In the third year, the candidate will explore how these safe-by-design techniques can be employed in the context of large language models (LLMs) and/or state-of-the-art deep learning architectures. During this phase, there will also be an opportunity to investigate the use of the developed techniques in industrial applications. It is expected that the scientific results of the project will be reported at top machine learning and artificial intelligence conferences (e.g., NeurIPS, ICML, ICLR, UAI, AISTATS, IJCAI, AAAI) and international journals (e.g., IEEE TKDE, ACM TKDD, ACM TIST). The research program will be conducted in collaboration with the Italian Institute of Artificial Intelligence for Industry (AI4I) and with the Delft Institute of Technology (TU Delft).

Skills and competencies for the development of the activity	• Good programming skills and proficiency in programming languages. • Strong background in either machine learning, deep learning, or applied mathematics. • Strong analytical skills and an ability to work at the intersection of several research domains. • Be fluent in English, both written and oral.