Trustworthy and Secure Artificial Intelligence

Contact
Varun Ojha, varun.ojha@newcastle.ac.uk

Research project
This project will be aligned with EPSRC funded National Edge AI Hub. It will investigate AI model quality for its trustworthiness and security. AI trustworthiness has been identified as a major challenge worldwide, especially after the emergence of large models. Multiple countries have established research centres and institutions to study and investigate the safety of AI, and it is the most emerging research direction in AI. With three PhD graduates in AI security research in my research group and several researchers working on this topic currently, I invite motivated AI security researchers to investigate and develop AI safety algorithms and methods under this project. Especially to *investigate various known and unknown attack scenarios and security challenges of a set of AI models* such as Deep Neural Networks and Transformers, to *develop novel adversarial robust defense mechanisms* to protect AI models against these attacks and to *recommend standards and practices for AI model security* based on common features of attack across a varied range of AI models. I will also help explore scholarship opportunity if available.

Applicant skills/background
This project requires skills in programming in python and machine learning research skills. The project welcomes researchers from disciplines of mathematics, engineering, physics, computer science, electronics.

References
Pravin C, Martino I, Nicosia G, Ojha V (2024), Fragility, Robustness and Antifragility in Deep Learning, Artificial Intelligence, Elsevier.

Liu et al. (2024), Dynamic Label Adversarial Training for Deep Learning Robustness Against Adversarial Attacks, 31st International Conference on Neural Information Processing.

Ojha et al. (2022), Backpropagation Neural Tree, Neural Networks, Elsevier.