Verifiable Artificial Intelligence

Prof. Pallab Dasgupta

IIT Kharagpur

An important limitation of many AI and ML techniques is in the lack of interpretability of their output in a form that is legible and verifiable by the end user. Interpretability and verifiability are closely related terms in the context of AI and ML, but the aim of the latter is to guarantee that the output is safe. The proof of safety may leverage explainable artefacts, and may also in turn serve as an explanation. Formal methods for verification are based on logical correctness and formal deduction, whereas the science of machine learning is based on statistical exploitation of patterns. Human cognition works by combining pattern recognition with deduction and inferencing – and the challenge lies in developing this convergence in AI.

The focus of this talk will be to primarily present the scientific challenges in this domain, without adherence to any specific school of thought or any specific line of work.