When AI Goes Awry

Des Higham, University of Edinburgh

Over the last decade, adversarial attack algorithms have revealed instabilities in artificial intelligence (AI) tools. These algorithms raise issues regarding safety, reliability and interpretability; especially in high-risk settings. Mathematics is at the heart of this landscape, with ideas from numerical analysis, optimization, and high dimensional stochastic analysis playing key roles. From a practical perspective, there has been a war of escalation between those developing attack and defence strategies. At a more theoretical level, researchers have also studied bigger picture questions concerning the existence and computability of successful attacks. I will present examples of attack algorithms for neural networks in image classification, for transformer models in optical character recognition and for large language models. I will also show how recent generative diffusion models can be used adversarially. From a more theoretical perspective, I will outline recent results on the overarching question of whether, under reasonable assumptions, it is inevitable that AI tools will be vulnerable to attack.