SORS: "Can We Rely On AI?"

Objectives

Abstract: Over the last decade, adversarial attack algorithms have revealed instabilities in deep learning tools. These algorithms raise issues regarding safety, reliability and interpretability in artificial intelligence (AI); especially in high risk settings. At the heart of these issues is the concept of instability: extreme sensitivity of the output to changes in the input. 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 in image classification and optical character recognition. I will also outline recent results on the overarching question of whether, under reasonable assumptions, it is inevitable that AI tools will be vulnerable to attack.

Desmond Higham

Short Bio: Professor of Numerical Analysis at the University of Edinburgh. He has been elected a Fellow of the Royal Society of Edinburgh and a Fellow of the Society of Industrial and Applied Mathematics (SIAM). He received the Shephard Prize from the London Mathematical Society and the Dahlquist Prize from SIAM. He served as editor-in-chief of SIAM Review from 2016--2023. He has research interests in scientific computation, data science and more recently in mathematical aspects of AI.

Speakers

Speaker: Desmond Higham. Professor of Numerical Analysis at the University of Edinburgh.
Host: Natasa Przulj, Integrative Computational Network Biology Leading Researcher

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---