Fabio Roli

University of Cagliari

Some Thoughts on Safety of Machine Learning

During the last forty years, machine learning research has been focused mainly on accuracy of algorithms, and occasionally on speed and scalability. Issues that are very important in mature engineering fields, like safety, reliability, testing, have been basically neglected. But something changed over the last five years. The rise of the machine learning paradigm based on “big data + deep learning + GPUs” and the market expectation of products for high-stakes applications, triggered the interest of academia, and the worries of stakeholders and media, on neglected issues, like the safety of a self-driving vehicle operating in an adversarial environment “out of its training set”. This talk does some reflections on the safety of modern machine learning, with a mention to a few issues related to safety, like liability and transparency of algorithms. We will use concepts and examples coming from recent works in computer science, dealing with safety of machine learning algorithms in hostile environments and against adversarial inputs, but keeping a language appropriate for an interdisciplinary audience.