Title: Aspects of Trustworthy Machine Learning

Abstract: Machine learning technologies, in particular deep learning, offer state-of-art tools for various domains including computer vision, natural language processing, or data-driven control and robotics. Yet their use in increasingly complex, open, and partially underspecified real-world environments carries severe risks: models become vulnerable to unexpected behavior and system failures in situations which have not been encountered during training; models are prone to attacks by adversaries, leading to surprising errors known as adversarial attacks; models show predictions or functions with a severe bias which humans would regard as unfair, e.g., discriminating persons with specific ethnical background or gender.

Such behavior has led to a quest for trustworthy AI, i.e., models which show safe behavior and which are compatible to human expectations in the real world. The tutorial course, will focus on three important aspects associated to trustworthy AI:

1. **Explainable AI**: How to enhance possibly black-box AI models by aspects which enable humans insight into their behavior and possible causes of failures? This part will introduce a taxonomy of different ways of explaining AI models. It will delve into two approached, so-called counterfactual explanations and feature relevance determination methods in more detail, highlighting aspects such as its computational complexity or uniqueness, and it will also touch on the question how to access the success of approaches of 'explainability'.
2. **Fairness**: How to guarantee that AI models do not discriminate specific groups or persons, e.g., by a systematic increase of the model error for specific groups? In this part, an overview of different notions how to define fairness in the first place will be given, and the impossibility of fulfilling all such criteria will be discussed. We will also have a look at approaches to arrive at fair models, by either addressing the data sets or the model training itself.
3. **Robustness**: Topics covered are diverse forms of adversarial attacks which lead to errors and system failures. This covers the general type of attacks (e.g. white box versus black box attack, universal attack, attacks in the physical world) as well as the specific numeric how to arrive at such attacks. Approaches how to robustify models against attacks are discussed, in particular robust training and classification with reject option.