"SCADNet - Structural Causal Analysis of Deep Neural Networks"

Causal explanations based on counterexample computation and counterfactual causal reasoning are a cornerstone of the design and engineering of (Safety-)Critical Software-Driven Systems (CSS). They are the foundation of system debugging during iterative system design cycles, form the basis of failure forensics in litigation and are central to the certification of system safety as well as societal acceptance.
In this project we focus on CSS that rely on decision-making components which are implemented using Machine Learning (ML) technology, in particular DeepLearning (DL).
As widely acknowledged, obtaining explanations for this type of CSS is fraught with significant challenges, largely due to the absence of easily interpretable logical code structure and, as a consequence, the non-applicability of existing software verification technology. We will develop novel structural abstraction and analysis techniques for Deep Neural Networks (DNNs) that allow for the computation of counterexamples in the context of specification-based safety verification. We will then address the elicitation of counterfactual evidence for safety violations and next turn to proposing automatic repairs for DNNs that violate specified safety properties, thus contributing to the explanation of the behavior of a DNN.
Finally, we will place these results in the context of a general theory of counterfactual causal analysis for ML-based CSS.