Käte Hamburger Kolleg: Cultures of Research
Lecture Series 2022 “Philosophy of AI_Optimits and Pessimist Views”

Kim Guldstrand Larsen (Aalborg University), 11.05.2022

Explainable and Verifiable Machine Learning. A Grand Challenge for Computer Science

Learning-based components in general and Neural Networks in particular are increasingly used in safety-critical systems, e.g. autonomous vehicles, smart energy grids, traffic control systems and several other complex and critical infrastructure systems. In such cyber-physical systems (CPS), which connect the physical and the digital world, the components do not only influence how the system perceive and interact with the environment but also makes decisions that influence its behavior. The rapid growth of machine-learning techniques in CPS leads to better products in terms of adaptability, performance, efficiency and usability. However, CPSs are often safety critical (e.g., self-driving cars and medical devices), and the need for measures against potential fatal accidents is self-evident and of key importance.

In fact, the trustworthy use of machine learning in safety-critical systems is a Grand Challenge. Importantly, the decisions made by a learning-based component should be explainable in terms that can be understood by a domain expert or end-user. Moreover, verification techniques must be developed that can provide absolute guarantees of a learning-based component, e.g. verified robustness of a Neural Network with respect to adversarial attacks or verified correctness when used in an application context.

The talk will highlight ongoing research towards explainable and verifiable learning-based CPS pursued within the context of the modelling and verification tool UPPAAL Stratego (www.uppaal.org). Here machine learning (reinforcement learning) are combined with symbolic synthesis (dating back to Alonso Church in 1957) in order to obtain provably safe, near-optimal and explainable control strategies. So far UPPAAL Stratego has been applied successfully within a number of application domains including energy systems, autonomous driving, storm-water detention ponds and traffic control.