AI Safety Project

MATTHIAS SCHOTT, TIMO SAALA

The Project

- BMBF funded project
- Cooperation of University of Bonn and RWTH Aachen (formerly also University of Mainz)

- Three large HEP experiments included
  - CMS (Aachen)
  - ATLAS (Bonn)
  - IceCube (Aachen)

- Additionally also have connections to Lamarr (CAISA Lab)
  - Main focus here: Deep Learning, specifically NLP

The Problem

- Main problem we are concerned about:
  - How can we correctly quantify systematic uncertainties from deep neural networks

- Possibly something AI Safety can help with answering

- Specifically so far:
  - Study different concepts from AI Safety on HEP models
  - E.g. varying Adversarial Attacks, as well as Defenses

The „Side“-Benefits

- Defenses from AI Safety can additionally increase robustness + generalization capabilities

- Attacks can present vulnerabilities / weaknesses in established models
  - To some extent can be used to „explain“ what happens in Deep Learning models

- Might also present some further interesting characteristics of physics data and models
  - E.g. maybe also applicable for transfer learning?

Current Work

- Constructed a pipeline, taking CMS Open Data (ROOT, 2012 Run), filtering it, and saving it as pandas DataFrames

- Re-Created multiple established HEP models using CMS Open Data
  - For further studies

- Constructed a novel adversarial attack, optimized to minimize the change in 1D variable distributions (as opposed to change on a per-input basis)

Future Work

- Construct more physics-motivated Adversarial Attacks
- Study effects of these Attacks in the context of Adversarial Defense techniques
- Test and establish new ways in which to increase the robustness of HEP Deep Neural Networks
- **Establish ways in which to better quantify the systematic uncertainties of Deep Neural Networks**