Contextual Attacks against AI-based Inference

AI-based inference algorithms are increasingly used to automate decision-making processes in various types of computing systems. The process typically involves the training of an AI-based model that will make inferences based on which decisions are made. Models are trained based on input data that capture the relevant aspects of the inference task at hand. Since collecting training data for training accurate inference models can be time-consuming and laborious, many applications *use pre-trained models* for particular tasks.

However, when such inference models are trained, often a particular contextual setting from which the training data are harvested is *implicitly assumed.* These contextual factors or constraints are, however, not explicitly incorporated in the properties or specification of the resulting inference models. This opens up an opportunity for malicious players to stage attacks against AI-based inference systems by fooling systems to apply seemingly benign AI models in *contextually incompatible* inference tasks or contextual settings.

The topic of this thesis would be to examine such contextual attacks and investigate approaches for mitigating them. The tasks to be solved would include:

- Study of relevant literature regarding contextual integrity and context-based attacks
- Implement a study case demonstrating a successful contextual attack against AI-based inference and analyse its key characteristics
- Design and implement a defence approach for mitigating the contextual attack

Note: This represents only a preliminary high-level description of the topic. Details and actual scoping of the thesis will be agreed on with separately with the supervisor.

Thesis supervisor:

Prof. Dr.-Ing. Markus Miettinen
markus.miettinen@fb2.fra-uas.de
+49 69 1533 3969

Secondary supervisor: Prof. Dr.-Ing. Peter Ebinger

Thesis language: English or German