SIT799 Human Aligned Artificial Intelligence

Pass Task 4.1: Consequences of Adversarial Attacks on AI systems

Overview

During week 4, you have been introduced to: adversarial attacks against AI; Real-world examples of adversarial attacks on AI; Some solutions to fight against adversarial attacks on AI. To better understand adversarial attacks against AI, in this assignment, we will look at possible consequences of these attacks against AI systems.

To complete this assignment, you need to refer back to Week 4 lecture material.

Submission Details

For this task you need to report on consequences of adverse attacks against AI systems – for example, fooling a computer vision system or an NLP system for impersonation – and submit a report to OnTrack that outlines the following points:

- A brief discussion of the application of AI in the example (use case) you are describing.
- A discussion on why you think adverse attacks can be initiated against that example.
- A discussion on the consequences of such attacks against this application.
- A discussion on the possible solutions you can implement to prevent adverse attacks against the example you are describing.

Constraints

The submitted report should not exceed **600 words** in length. The report should have a **high-quality writing style**.