Abstract

Diploma thesis *„The failure modes of Large Language Models“* focuses on addressing failure modes of Large Language Models (LLMs) from the ethical, moral and security point of view. The method of the empirical analysis is document analysis that defines the existing study, and the process by which failure modes are selected from it and analysed further. It looks closely at OpenAI's Generative Pre-trained Transformer 3 (GPT-3) and its improved successor Instruct Generative Pre-trained Transformer (IGPT). The thesis initially investigates model bias, privacy violations and fake news as the main failure modes of GPT-3. Consequently, it utilizes the concept of technological determinism as an ideology to evaluate whether IGPT has been effectively designed to address all the aforementioned concerns. The core argument of the thesis is that the utopic and dystopic view of technological determinism need to be combined with the additional aspect of human control. LLMs are in need of human involvement to help machines better understand context, mitigate failure modes, and of course, to ground them in reality. Therefore, contextualist view is portrayed as the most accurate lens through which to look at LLMs as it argues they depend on the responsibilities, positions, and agency of involved human actors. The positive element of IGPT is its improved processes that include human control through human-in-the-loop systems. However, IGPT is still in its infancy and needs improvement by looking at human agents more systematically. There indeed is a difficult human compliance journey ahead.

Author: Bc. Soňa Milová

Supervisor: Mgr. Petr Špelda, Ph. D.

Study programme: Security Studies

Academic Year: 2022/2023