Felix Willen
An St. Magdalenen 6, 50678 Cologne
felix.willen@smail.th-koeln.de
07.11.2022

Exploration of the alignment problem in machine learning with the operation of simulated agents using a critical dissection of the prevailing anthropocentric approach

This is the bachelor thesis proposal for the study course "Code & Context" at the Technical University of Cologne.

- Student: Felix Willen
- Supervisor: Prof. Christian Faubel - Technische Hochschule Köln
- Term: Winter Semester 2022/2023
- Title: Exploration of the alignment problem in machine learning with the utilization of simulated agents using a critical dissection of the prevailing anthropocentric approach

Statement

The research paper will explore, how we tell machine learning algorithms, what to do. It is not an easy task to translate a human will on to a machine. There are many parts, where misalignment can happen. The so called Alignment Problem is raising in popularity and importance, because we are seeing machine learning algorithms being used in a wide range of technology. The prevailing anthroprocentric approach is deeply rooted in human history and results in a measurement of intelligence, according to our perception of it. This perception can be flawed and leads to distorted results, which can effect us as humans directly through our technology we use everyday.

Objectives

- Creating an enviroment in Unity, where agents train on specific tasks, in order to investigate the alignment problem
- Researching the anthropocentric approach in machine learning and its history
- Generating a techno-socio conclusion and bringing attention to the topic

Research Questions

- Is the alignment problem a human error or is the technology flawed?

- What is the human perception of intelligence and how did we translate it to machines and other beings?
- How are we influenced by the fast development of "Artifical Intelligence" in our everyday life?

Research approach

The practical approach of using ml-agents in Unity will directly explore how machine learning algorithms are used in order to follow human instructions. Communication is never easy, especially when the conversation partner is a computer and the language is code. No parts can not be overheard, but the interpretation can differ.

Related Work

James Bridle (2022) Ways of being: beyond human intelligence
Brian Christian (2020) The alignment problem: machine learning and human values
Adrian Daub (2020) What tech calls thinking
Hubert L. Dreyfus (1992) What Computers Still Can't do

Expected Results

The goal of this research is to develop a deep understanding of machine learning algorithms and how they learn and act. Further providing a sociological viewpoint of the technology influencing and affecting our lifes.

Schedule and Milestones

Date	Milestone	Plan
03.10.2022		Start of practical research part
	1	Research of the alignment problem
	2	Learning of ml-agents in Unity
	3	Train an agent in a simulated enviroment
07.11.2022		Exposé write up
11.11.2022		Submitting Bachelor Exposé and application documents
21.11.2022		Start of writing Bachelor Thesis
	4	Training agents in the final enviroment
	5	Evaluation of the reward function and the alignment of the agents
	6	Final writing
10.02.2023		End of Bachelor Thesis timeframe

Date	Milestone	Plan
before 26.02.2023		Thesis defense

Felix Willen
felix.willen@smail.th-koeln.de