

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

Philippe van Basshuysen

18 November 2019

Worksheet 5 – The Implications of Superintelligence

1. **Intelligence and motivation**

State the *orthogonality thesis* and the *instrumental convergence thesis* in your own words. Are the theses plausible?

What follows from the two theses about the likely behaviour of a superintelligent machine (even if we do not know its final goals)?

1. **Is the default outcome doom?**

What follows from the principles from (1.) and the “first-mover advantage” for the fate of humanity and other intelligent life?

How could we possibly prevent a superintelligent machine from “perversely instantiating” good goals?

1. **! Next Week: Public Climate School Week !**

There will be no presentation or worksheet. Instead, we will discuss topics at the intersection between AI and climate change, and we will possibly attend one of the climate school week events. In preparation of the seminar, please read and think about the following articles.

**Contact points of artificial intelligence and climate change**

1. AI can help tackle climate change by making the production of goods and services more energy-efficient.

Readings: <https://www.nationalgeographic.com/environment/2019/07/artificial-intelligence-climate-change/>

Rolnick et al. (2019), Tackling Climate Change with Machine Learning (<https://arxiv.org/pdf/1906.05433.pdf>) - read the introduction

1. Using AI and research into AI are themselves causes of significant amounts of carbon emissions.

Readings: <https://horizon-magazine.eu/article/ai-can-help-us-fight-climate-change-it-has-energy-problem-too.html>

1. Climate change research uses AI.

Readings: <https://www.climatechangenews.com/2019/09/19/climate-scientists-harness-artificial-intelligence-handle-big-data/>