

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

Philippe van Basshuysen

06 January 2019

Worksheet 9 – Designing Ethical Algorithms?

1. **Should machines be preference utilitarians?**

State Russell’s 3 principles for beneficial machines in your own words.

Will machines that obey the three principles be beneficial? What could go wrong?

Why is it problematic to programme an AGI to obey human commands? How does Russell try to solve the problem?

Why is it important to integrate uncertainty into the machine’s objective function?

1. **Next Week: AI and the Future of Work**

Readings:

Ernst et al. (2018). The economics of artificial intelligence: Implications for the future of work. *ilo future of work research paper series*

Stuart Russell (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane. Pp. 113-124.

Optional Readings:

Derek Thompson (2015). A World Without Work. <https://www.theatlantic.com/magazine/archive/2015/07/world-without-work/395294/>