

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

Philippe van Basshuysen

11 November 2019

Worksheet 4 – The Possibility of Superintelligence

1. **How will AI Develop in the Future?**

What are the present obstacles to the development of a superintelligence that Stuart Russell identifies?

Are these obstacles likely to be overcome in the medium or long run?

Why does Kevin Kelly believe that superhuman AI is a “myth”?

1. **The Risks of Superhuman AI**

Why might it be problematic if there are machines substantially more intelligent than humans?

Why is it not a solution to this problem to programme machines in a way that aligns with human purposes?

If superintelligent machines pursue objectives that we don’t like, why can’t we simply pull the plug?

1. **Next Week: The Implications of Superintelligence**

Readings:

Nick Bostrom (2014). Superintelligence. Oxford: Oxford University Press. Chapter 6: COGNITIVE SUPERPOWERS, Chapter 7: THE SUPERINTELLIGENT WILL and Chapter 8: IS THE DEFAULT OUTCOME DOOM?

Maciej Ceglowski (2016). Superintelligence: The Idea That Eats Smart People. <https://idlewords.com/talks/superintelligence.htm>

Optional readings:

Nick Bostrom & Eliezer Yudkowsky (2015). The Ethics of Artificial Intelligence. In *Cambridge Handbook of Artificial Intelligence* (Keith Frankish and William Ramsey, eds.), New York: Cambridge University Press.