

Philosophy & Ethics of Artificial Intelligence

Master in Philosophy of Science

Philippe van Basshuysen

28 October 2019

Worksheet 2 – Agency and Responsibility

1. **Leftover from last week**

*Define the following kinds of AI:*

Narrow AI:

Artificial General Intelligence (AGI):

Artificial Superintelligence:

1. **Agency**

When should something count as an *intentional agent*?

Discuss whether the following AI systems are intentional agents:

* self-driving cars
* image recognition system that learns to recognise the faces of people (*supervised machine learning*)

1. **Responsibility**

Are intentional agents responsible for the consequences of their acts?

What is a Responsibility Gap, and when might it occur?

If Responsibility Gaps arise in the context of AI systems, can they be closed by assigning responsibility to the systems? Can AI systems be “fit to be held responsible”?

1. **Rights for AI systems?**

Distinguish legal vs. moral rights, deontic vs. status rights, and non-derivative vs. derivative rights.

Under what circumstances (if at all) should we grant AI systems non-derivative rights?

1. **Next Week: Epistemology of Machine Learning**

*Readings:*

Will Knight (2017). The Dark Secret at the Heart of AI. <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>

Emily Sullivan (forthcoming). Understanding from Machine Learning Models. *British Journal for the Philosophy of Science*.

*Optional readings:*

Alex Hern (2015). Yes, androids do dream of electric sheep. <https://www.theguardian.com/technology/2015/jun/18/google-image-recognition-neural-network-androids-dream-electric-sheep>

Gregory Wheeler (2016). Machine Epistemology and Big Data. In: McIntyre, Lee; Rosenberg, Alex (eds.), *The Routledge Companion to The Philosophy of Social Science*, Routledge.