Philosophy & Ethics of Artificial Intelligence
Master in Philosophy of Science
Philippe van Basshuysen

28 December 2019

Syllabus & Reading List

NOTE: THE READING LIST WILL BE UPDATED AS THE TERM MOVES ON.

Topics and Aims

In this seminar, we will reflect on philosophical problems that recent advances in artificial intelligence (AI) have brought to light, as well as problems that AI is likely to pose in the future. After completing the course, students will be able to make informed judgements about what AI is and what it can achieve; how it impacts our lives and society; and whether and how research and use of AI should be regulated.

We will introduce and discuss five major topics:

AI: Philosophical Questions. We will kick off the seminar with basic questions of how to define AI and how to classify different types, such as narrow AI and artificial general intelligence (AGI). We will discuss when an AI system should count as an agent and how an answer to this question relates to responsibility for the system's actions. We will also examine epistemological questions relating to the black boxes that machine learning generates.

Risk and Potential of AI. As the capacities of AI increase, so do the associated risks of unintended and undesirable consequences. It has been argued that superintelligence – AI that surpasses human cognitive performance – may bring about socially disastrous outcomes. We will become acquainted with and evaluate the arguments for the possibility of superintelligence and its associated risks.

Ethical Challenges of Narrow AI. Narrow AI poses ethical problems that are visible already today. We will classify these issues. We will also examine some case studies in which the application of algorithms appears to be ethically problematic, with an emphasis on algorithmic bias.

Towards Ethical Algorithms? We will discuss possible solutions to the ethical concerns encountered in previous sections. These include learning how to make fair decisions while using algorithmic recommendations as input, as well as designing "ethical algorithms". The limitations and drawbacks of these proposals will also be examined.

AI and the Economy. Advances in AI have wide-ranging economic impacts. We will examine the mechanisms through which AI may increase efficiency but which at the same time may put jobs at risk. We will discuss how we might reorganise society as a response to such developments. We will deliberate about whether the generation of data should be treated as labour. Finally, we will consider possible wide-ranging transformations of markets and the possibility of an "AI socialism".

Organisational Issues

- Email: philippe.van.basshuysen@philos.uni-hannover.de
- Office hours: Mondays 12-1 pm, Room B407 (4. OG)
- Please register for the course on Stud.IP (https://studip.uni-hannover.de/dispatch.php/course/details?sem_id=b9a390950828917411a7ee49b3e003d5)
- In order to get credits for this course, participants are asked to give a **short presentation** (10-15 minutes) about one of the required readings, and write a **term paper** (up to 5000 words – strict word limit including headings, references, footnotes, etc.).
- Please come to my office hour one or two weeks prior to your presentation.
- You can submit a critical analysis exercise (up to 1500 words) as a useful starting point for your term paper. This is a short analysis and evaluation of one of the main readings of your choice. You will get feedback (not a grade) if you submit before the Christmas break (22 December).

Reading List

AI: Philosophical Questions

Week 1 (21/10/2019): Defining AI

No required readings, but Chapter 1 of Stuart J. Russell & Peter Norvig (2010). *Artificial Intelligence: A Modern Approach*. Pearson (3rd edition) is a good introduction to the topic.

If you want to dive a bit deeper into the techniques used in AI, I recommend the free online course *Elements of AI* (https://www.elementsofai.com/), which is designed by Reaktor and the University of Helsinki. Some exercises used in the seminars are taken from this course.

Week 2 (28/10/2019): Agency and Responsibility

Required Readings:

Andreas Matthias (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, 175–183.

Christian List (2019). Group Agency and Artificial Intelligence. *Discussion Paper*

Additional, optional readings:

Migle Laukyte (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology* 19, 1–17.

Robert Sparrow (2007). Killer Robots. *Journal of Applied Philosophy* 24(1), 62-77.

Janosch Delcker (2018). Europe divided over robot 'personhood'. Politico, https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/

Week 3 (04/11/2019): 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.

Carlos Zednik (forthcoming). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence

Risk and Potential of AI

Week 4 (11/11/2019): The Possibility of Superintelligence

Readings:

Stuart Russell (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane. Chapter 3: HOW MIGHT AI PROGRESS IN THE FUTURE? And Chapter 5: OVERLY INTELLIGENT AI

Kevin Kelly (2017). The Myth of a Superhuman AI. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/

Optional readings:

Nick Bostrom (2014). Superintelligence. Oxford: Oxford University Press. Chapter 4: THE KINETICS OF AN INTELLIGENCE EXPLOSION

Week 5 (18/11/2019): 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.

Week 6 (25/11/2019): Additional Topic (Public Climate School Week): AI and Climate Change.

Open discussion – no required readings.

Ethical Challenges of Narrow AI

Week 7 (02/12/2019): Ethics of Algorithms

Readings:

Brent Daniel Mittelstadt et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 1-21

Cathy O'Neill (2016). *Weapons of Math Destruction*. New York: Broadway Books. CHAPTER 7: "Sweating Bullets: On the Job"

Optional readings:

Alexander Nill & Robert J. Aalberts: Legal and Ethical Challenges of Online Behavioral Targeting in Advertising, http://www.tandfonline.com/doi/abs/10.1080/10641734.2014.899529

Duhigg: How Companies learn your secrets. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?mcubz=1

Week 8 (09/12/2019): Algorithmic Bias and Fairness

Readings:

Reuben Binns (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of Machine Learning Research 81:1-11, Conference on Fairness, Accountability, and Transparency

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Optional readings:

David Danks & Alex J. London (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (IJCAI 2017)

Anne Washington (2019). How to Argue with an Algorithm: Lessons from the Compas-ProPublica Debate. *Colo.Tech.L.J.* 17(1), 131-160.

Sigal Samuel (2019). This AI makes you look like a masterpiece — while teaching you about its own bias. https://www.vox.com/future-perfect/2019/7/25/20708589/ai-portraits-art-bias-classical-painting

Rambachan, Roth (2019) – Bias In, Bias Out? Evaluating the Folk Wisdom. Draft

Towards Ethical Algorithms

Week 9 (16/12/2019): Ethical Decision-making with Algorithms

Readings:

Alexandra Chouldecheva (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056v1*. NOTE: This is a more technical paper. Try to read pp. 1-7.

Optional readings:

Kirsten Martin (2018). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 1-16.

Week 10 (06/01/2020): Designing Ethical Algorithms?

Readings:

Stuart Russell (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane. Chapter 7: AI: A DIFFERENT APPROACH and Chapter 8: PROVABLY BENEFICIAL AI

Optional readings:

Peter Eckersley (2019). Impossibility and Uncertainty Theorems in AI Value Alignment. *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*

Nick Bostrom (2014). Superintelligence. Oxford: Oxford University Press. Chapter 12: ACQUIRING VALUES

AI and the Economy.

Week 11 (13/01/2020): 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/

Week 12 (20/01/2020): Data Regulation

Readings:

Eric Posner and Glen Weyl (2018). Radical Markets: Uprooting Capitalism and Democracy for a Just Society. Princeton and Oxford: Princeton University Press. CHAPTER 5: “Data as Labor”

Spiekermann et al. (forthcoming). Big Data Justice: A Case for Regulating the Global Information Commons

Week 13 (27/01/2020): The Future of the Market Economy

Readings:

Eric A. Posner and E. Glen Weyl (2018). Radical Markets: Uprooting Capitalism and Democracy for a Just Society. Princeton and Oxford: Princeton University Press. CHAPTER EPILOGUE: “After Markets?”