Content

1. Course aims and objectives
2. Seminar format
3. Assessment
4. Reading

1. Course aims and objectives

Research on artificial intelligence (AI) is progressing rapidly, and AI plays an increasing role in our lives. These trends are likely to continue, raising new ethical questions about our interaction with AI. These questions are intellectually interesting in themselves, but as AI develops and is more widely used they are also of pressing practical importance. One of the main aims of this course is that you gain familiarity with some of these questions and their potential answers. In particular, you will become familiar with the following questions:

- How do we align the aims of autonomous AI systems with our own?
- Does the future of AI pose an existential threat to humanity?
- How do we prevent learning algorithms from acquiring morally objectionable biases?
- Should autonomous AI be used to kill in warfare?
- How should AI systems be embedded in our social relations? Is it permissible to fall in love with an AI system?
- What sort of ethical rules should an AI such as a self-driving car use?
- Can AI systems suffer moral harms? And if so, of what kinds?
- Can AI systems be moral agents? If so, how should we hold them accountable?

Another aim of this course is that you develop the skills needed to think about these questions with other people. For this reason, there will be no formal lecturing by me on this course, and I will play somewhat a limited role in determining the focus of discussion. Instead, students will present the material for discussion each week, and we will work out which questions we would like to discuss together. We will work on these questions in small groups, and then present our findings to the main group for general discussion (there is more detail on how seminars will run below). In the process, you will:

- Improve your presentational skills.
- Improve your ability to pick out and formulate productive questions.
- Improve your ability to work on a problem as part of a small group.

2. Seminar format

Teaching on this course takes the form of a two hour seminar, held once per week. Here is an explanation of the seminar format, and some explanation of why the seminars have that format.
Before the seminar

Before each seminar you must (a) do the required reading (and watch the video if there is one), and (b) prepare a question that you would like to ask about that reading. Asking a good question is not easy, and you should expect to spend some time on this. The aim is to come up with a question that can be expressed in one or two sentences, and that someone who has read the material for the week will be able to understand. You may ask a question that is intended to make an idea or argument under discussion clearer. For example, you might ask a question of the form:

‘x claims that p, but I’m unclear about what p means. In particular, I don’t understand this word/phrase…’

‘x says that p, but I don’t understand why he says that p. Is p part of x’s view? Or is p part of an argument for his view?’

‘why does x spend so long talking about this? How is it relevant to x’s main claim/argument?’

You may also ask questions that are aimed at assessing ideas and their consequences. For example, you might ask questions of the form:

‘x says that p, but isn’t it a counterexample to p that…’

‘x says that p entails q, but surely p does not entail q because…’

‘x says that p. If p is true, then doesn’t this have the welcome/unwelcome consequence that…’

When trying to come up with a question, try to think of the question that will best enable you to understand and assess an idea from the reading that you think matters or that you find interesting. We don’t get many opportunities to ask questions, and our aim here is to get better at making the most of these opportunities.

Note that you are required to bring a question to the class for your participation mark. If there is a good reason that you are unable to bring a question to class please let me know, and you will not lose your participation mark for that week.

Student presentation

At the start of each seminar, two students will summarise the material for that week in a 10-15 minute presentation. These presentations are intended as a basis for the following discussion. As such, they should summarise the main points of the material, rather than recapping the material in detail. Giving a good summary of this sort is very difficult, but if you want to discuss a philosophical question (or any other kind of question for that matter) it’s crucial that you be able to communicate the gist of an idea clearly and quickly, without getting bogged down in the details. I strongly encourage you to email me to make an appointment when you are due to present, and I will help you to prepare your presentation. I’m happy to talk to you about what to put into the presentation, what to leave out, how to structure it, and about how to understand the ideas that you are presenting. Note that you don’t need to have a complete understanding of what you are presenting, and it’s a good thing to talk about why you don’t understand certain aspects of the material in your presentation.

Presenters should provide either a handout or slides. Please send these to me after class and I will make them available on Learn.
If you have trouble with presenting due to anxiety or for other reasons then you don’t need to present. Please get in contact with me at the start of the course and we will make other arrangements.

**Collation of questions**

After the presentation, we will spend 10-15 minutes going through the questions that people want to discuss. We will talk together about whether some of the questions that people have bought are the same or similar to other people’s questions, and about how to clarify certain questions. At the end of this process we will draw up a list of 5-8 questions. Note that although I will always try to ensure that everyone’s original question is relevant to something on this list, I cannot guarantee it.

**Small group discussion of questions**

We will then spend half an hour working on these questions in small groups. Each group will start by nominating someone to keep notes and present the findings of the group to the class. In some cases it will be possible to arrive at fairly definite answers to questions, while in other cases this will not be possible; indeed, it may well be that there is no known answer to a question. In this case, the job of the group is to note that this is so, and to outline some potential answers along with their pros and cons.

**General group discussion**

After a short break the rest of the class will consist of general group discussion of the answers arrived at by each group. Each group will give a short summary of their response to the question they were assigned, followed by a short discussion of that answer by the entire group.

3. **Assessment**

Your grade for this course will be made up of the following:

- 5% participation grade.
- 20% short writing assignment (500 words).
- 20% short writing assignment (500 words).
- 55% end-of-semester essay (2,000 words).

Word counts include footnotes, but do not include essay titles and bibliographies.

**Participation grade**

For each week that you do not turn up to class with a prepared question you will lose 1% of your grade. You cannot lose more than 5% of your grade in this way.

**Short writing assignments**

For each short writing assignment you will pick a question that you have worked on in a small group in class and present a response to that question in 500 words or less. Note that your ability to pick and formulate a good question is part of what is being assessed here, and I encourage you to get in contact with me to discuss which question you would like to address. I am happy for you to re-word the original question, and I encourage you to draw on both your small group discussion and the wider class discussion of the question.

In these assignments I am looking for a clear and concise response to the question. I do not want you to try to summarise everything that was said about the question in class. Note that I am asking you
to present a *response* rather than an *answer* to the question. You might present what you think is the answer to the question, or you might present a potential answer and explain why it does not ultimately work. Or, you might compare the pros and cons of two different answers. We will devote some class time to discussion of how to do the short writing assignments before they are released.

*End-of-semester essay*

For the end-of-semester essay you may pick a question from the list of essay questions posted on Learn, or you may make up your own question. If you do the latter, please email me or come and see me to check whether your question is ok. If you pick an overly ambitious/narrow, off topic, or unclear question this can seriously affect your mark.

You may pick essay questions that overlap with your short writing assignments, and it is ok for there to be some overlap in your answers. However, you may not simply paste material from your short writing assignments into your essays. Any good use of material from your short writing assignments here will involve adapting and developing it so that it fits into your essay. The less you change it, the less successful your use of material from short writing assignments in an essay is likely to be.

Regarding reading for essays, the essential and secondary readings should be your first port of call. But you should not restrict yourself to these. Follow up and read relevant cited papers in the bibliographies of papers from the reading list, use Google and Google Scholar’s useful ‘cited by’ feature to explore further responses to the papers you read, and visit the websites and journals listed below to discover other relevant articles for your essay. Ethics of AI is a fast moving field and a relevant article may appear during the course of the semester.

If you haven’t written a philosophy essay before (and perhaps even if you have), you may want to look at the following advice:

- Jim Pryor: [Guidelines on Writing a Philosophy Paper](#)
- Amy Kind: [How to Write a Philosophy Paper](#)
- Philosophy Faculty, Cambridge: [Tackling the Philosophy Essay](#)
- Harvard Writing Center: [A Brief Guide to Writing the Philosophy Paper](#)
- Peter Lipton: [Writing Philosophy](#)
- Peter Smith: [Writing Style](#)
- Mark Sprevak’s document, ‘Strategy for writing a first class essay’, in the Course Info folder on Learn.

*4. Reading*

*Background reading*

If you are completely new to this topic, this podcast introduces some of issues we discuss:

- [Living with Robots: A Conversation with Kate Darling](#), *Waking Up* podcast, 1 March 2017

I recommend that you read these two books. They will help you understand many of the topics.


If you would like an introduction to ethics, these are good starting points:
If you have not studied philosophy before, then these are excellent resources for filling in gaps in knowledge:

- The Stanford Encyclopedia of Philosophy
- The Internet Encyclopedia of Philosophy
- Routledge Encyclopedia of Philosophy Online
- Philosophy Compass

Class reading

For each week, the readings (along with other useful resources like videos and tutorials) are listed below. These resources are divided into essential and secondary. Essential readings and videos are the material that it is your responsibility to read before each class.

Please also delve into the secondary readings and videos. They can help you develop your thoughts about the essential reading and help you to come up with a focused question for class.

All the essential readings and as many as possible of the secondary readings are available on Learn: go to Readings, and then Resource List.

Before class, read the essential reading carefully. You may find a paper challenging or difficult – persist! If you do not understand something, read it again, think about it, try to make sense of it in your own words. If after multiple attempts to make sense, you still cannot, then there is a good chance that you have identified a real problem in the article – a perfect point for your question, or to form the basis of an excellent essay!

Jim Pryor has some wonderful practical advice for reading philosophy (as he says, ‘you should expect to read a philosophy article more than once’).

Week 1 – What is ethics of AI?

Essential reading:

- Sections 1, 2 and 3 of this course handbook.

Secondary:

- ‘Benefits & risks of artificial intelligence’, Future of Life Institute

Week 2 – The singularity

Essential video:
• Harris, S. ‘Can we build AI without losing control over it?’, TED talk, October 2016.

Essential reading:

Secondary:
• 20 papers responding to Chalmers’s paper in 2 special issues of Journal of Consciousness Studies (these can be downloaded as PDFs from a computer on the University of Edinburgh network):
  – http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/f0020001
  – http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/f0020007
• Bostrom, N. (2014), Superintelligence: Paths, Dangers, Strategies, Oxford University Press, Chapters 2–6
• E. Yudkowsky. ‘Three Major Singularity Schools’, blogpost on Machine Intelligence Research Institute, September 2007
• The Singularity Film has some nice interviews with experts.

Week 3 – The value alignment problem

Essential video:
• Bostrom, N. ‘What happens when our computers get smarter than we are?’, TED talk, April 2015.

Essential reading:

Secondary:
- E. Yudkowsky and S. Harris ‘AI: Racing Toward the Brink’, interview and podcast on Machine Intelligence Research Institute, 28 February 2018
- E. Yudkowsky. ‘There’s No Fire Alarm for Artificial General Intelligence’, blogpost on Machine Intelligence Research Institute, 13 October 2017
- S. J. Russell. ‘Q & A: The future of artificial intelligence’
- S. J. Russell. ‘3 principles for creating safer AI’, TED talk, April 2017
- *Robot & Frank* nicely explores some of difficulties of a machine learning human values

**Week 4 – Racist AI**

Essential video:

Essential reading:

Secondary:
- The AI Now Institute, *Algorithmic Accountability Policy Toolkit*, posted on their website, 1 October 2018
- R. Ghani. ‘You Say You Want Transparency and Interpretability?’, blogpost on 29 April, 2016
- *Algorithmic bias: From discrimination discovery to fairness-aware data mining* recorded tutorial with lots of extra resources

**Week 5 – Autonomous weapons**
Essential video:


Essential reading:


Secondary:


Week 6 – Falling in love with AI

Essential video:

- Radiolab (2018). ‘More or less Human’, Podcast, May 2018

Essential reading:


Secondary:

- Devlin, K. ‘In defence of sex machines: why trying to ban sex robots is wrong’, The Conversation, 17 September, 2015
- Essays in this edited collection:
• Sharkey, A. (2016) Should we welcome robot teachers?, Ethics and Information Technology 18, pp. 283–297
• The Verge, (2018), ‘Sony’s Aibo is a very good robot dog’, news report, 9 January 2018
• Here is a rather good film that explores some of these ideas
• Chiang, T. (2010) The Lifecycle of Software Objects Subterranean Press. This is a short story that also explores some of these ideas.

Week 7 – Humans need not apply

Essential video:

• CGP Grey, ‘Humans Need Not Apply’, video talk, 13 August 2014
• Autor, D. H., ‘Will automation take away our jobs?’, TED talk, September 2016

Essential reading:


Secondary:

• Collins, K. ‘A programmer automated their data-entry job. Now the question is whether to tell their employer’, Quartz, 30 June 2017
• Simon, M. (2018) The tale of the painting robot that didn’t steal anyone’s job, Wired Magazine, 8 February 2018
• Macquarie University, For Work / Against Work: Debates on the centrality of work has a useful online bibliography
Week 8 – Good and bad robots

Essential video:


Essential reading:


Secondary:

- Schwitzgebel, E. ‘Should You Defer to Ethical Experts?’, blogpost on 15 March 2019
- Wallach, W., Allen, C. (2008), Moral Machines, Oxford University Press

Week 9 – Who lives and who drives?

Essential video:

- Lin, P. ‘The ethical dilemma of self-driving cars’, TED talk, 8 December 2015
- Rahwan, I. ‘What moral decisions should driverless cars make?’, TED talk, 8 September 2017

Essential reading:

- Nyholm, S. ‘The ethics of crashes with self-driving cars: A roadmap, I’ Philosophy Compass 13, e12507
Secondary:

- The Moral Machine is a website with an interesting collection of moral dilemmas

Week 10 – Robot rights

Essential video:

- Cohen, G. *A.I. Ethics: Should We Grant Them Moral and Legal Personhood?*, video posted to YouTube, 23 September 2016

Essential reading:


Secondary:

• Kagan, S. (2016), ‘What’s Wrong with Speciesism?’ *Journal of Applied Philosophy* 33, 1–21 and responses in the same journal issue:

*Week 11 – Essay writing workshop*

In class this week we will talk about how to write a good essay, and work through some essay plans together. You may want to take a look at some of the essay writing advice articles listed in the *end-of-semester essay* section.

*Further resources*

These websites have useful material – blogposts, articles, videos, news items, links – relevant to this course. This is a fast moving area and the websites are updated regularly.

• [AI Now Institute](https://www.ali.research) at NYU has useful reports and publications.
• [Center for Human-Compatible AI](https://humancompatible.ai) has a superb bibliography.
• [Ethics + Emerging Sciences Group](https://ethicsemerging.org) often links to interesting news stories.
• [Ethics of AI conference at NYU](https://ethicsnyu.org) streamed video discussions from October 2016.
• [Future of Humanity Institute](https://fhi.ox.ac.uk) is the other big UK centre worth checking out.
• [Future of Life Institute](https://futurelife.org) has lots of useful resources and links.
• [LessWrong](https://lesswrong.com) has interesting posts on the alignment problem and related issues in decision making and a useful wiki
• [Leverhulme Centre for the Future of Intelligence](https://www.leverhulme.ac.uk) is a major UK centre on this topic and it’s worth checking out its events and news.
• [The Machine Intelligence Research Institute](https://mirlab.org) has many useful publications and blog posts

These journals regularly publish on the topics relevant to this course. If you are writing you essay and looking for extra reading on a topic, or a slightly different topic, then dip into recent issue.

• *AI & Society*
• *Ethical Theory and Moral Practice*
• *Ethics and Information Technology*
• *Minds & Machines*
• *Philosophy & Technology*
• *Science and Engineering Ethics*
• *arXiv (with sanity preserver)* – not a journal but most current research on AI is here

*Blogs.*

• [Amazon Machine Learning](https://amzn.to/328w3B7)
• [Andrej Karpathy on Medium](https://medium.com/@akek)
How to get started with AI or machine learning

You don’t need a lot of maths, but some you do need some. The key bits to get your head around are probability (especially Bayesian inference), calculus, and linear algebra. On the programming side, I’d advise starting with Python, which is an easy and fun language to learn.

Books:


Online resources and courses:

- [Beneficial AI Society (Edinburgh)](https://www.beneconfic.ai/)
- [Convolutional Neural Networks for Visual Recognition (Stanford)](https://cs231n.github.io)
- [Coursera — Machine Learning (Andrew Ng)](https://www.coursera.org/)
- [Reinforcement Learning (David Silver)](https://www.coursera.org/)