**Ethics of Artificial Intelligence**

Course guide v1.04 for Phil10167

Dr. Mark Sprevak, University of Edinburgh

1 Course aims and objectives

Artificial intelligence (AI) is developing at an extremely rapid pace. We should expect to see significant changes in our society as AI systems become embedded in many aspects of our lives. This course will cover philosophical issues raised by current and future AI systems. Questions we consider include:

- 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 AI like 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?
- How should we live with and understand minds that are alien to our own?

By the end of the course, the student should be able to:

- Demonstrate knowledge of philosophical issues involved in ethics of artificial intelligence
- Demonstrate familiarity with relevant examples of AI systems
- Show ability to work in a small team
- Show ability to produce written work regularly to a deadline
- Acquire ability to express arguments clearly and concisely
- Gain skills in research, analysis and argumentation

2 Format of the course

This course likely differs from others you have taken, so please read this section carefully.

This course is taught in the flip-classroom format.\(^1\) In class, you will write, work in a team, present ideas, discuss the ideas of others, and engage in constructive and rational dialogue. The best way of learning philosophy is to do it, and that’s what we’ll do together in class.

Before each class, you are expected to: (i) have done the essential reading (and watch the video where relevant) and (ii) thought of at least one question about the essential reading to bring to class. In class we will work together to understand the reading and assess the claims and argument.

I will not assume that you have understood every aspect of the reading before class. It is fine to be puzzled by parts of the reading and not all of the readings are easy to read. But I will assume that that you have (a) done the reading and (b) made a serious attempt to understand it before class.

The plan for each class is as follows. We begin with a brief summary of the main points from the reading. Note that this is not a lecture – students will be expected to contribute to the summary. We will then collect questions from the class about the reading and select which we are going to pursue. During the remainder of the class, we will work through the questions, breaking them up into smaller questions where necessary. When we consider a question, each table in the classroom will be given a small task to help us answer that question. Different tables may be given different tasks. A table may be asked to argue for a particular case (pro or contra), or to assess the merits of a given view, or to think of counterexamples to a generalisation or fallacies with an argument, or to find out some information that is relevant to answering the question. This work is to be undertaken as a group: work together with others at your table to solve the task. We will then share the work of each table and discuss how it helps answer our question.

Some tasks for tables may be harder than others, and in some cases there may be no known solution. In such a case, it is your table’s job to explore the logical landscape, map out and carefully justify the options, and share these with the class. This is what the best philosophical work often consists in.

Week 11 is different. There will be no new readings or videos for week 11. Instead, it will be an opportunity to consolidate your work from the semester and write a report that captures the best ideas from your table from the semester. I will give a prize to the table that produces the best report.

2.1 Practicalities

Before the course starts:

1. You need a Google account if you don’t have one already: https://accounts.google.com/SignUp?
2. Log on to the course’s Google Drive web page to access all the materials – you can find a link to this web page on Learn

Before each class:

1. Read the essential reading (and watch video where relevant) before class for each week.
2. Come to class with at least 1 question prepared about the reading. We’ll go round and collect all the questions and decide which ones to discuss. Make sure your question can be expressed concisely and clearly. Try writing the question down in no more than 2 sentences before you come. Examples of possible questions include:
   - ‘I don’t understand X’, where X is some specific passage of the reading.
   - ‘I don’t see why X follows from Y’, where this is part of the argument of the reading.
   - ‘I’ve thought of a counterexample to X’, where X is some claim in the reading
   - ‘I’m not sure how X applies to Y’, where Y is relevant case not yet considered

In class:

1. Bring a copy (either electronic or hard copy) of the essential reading
2. You may use bring your own device (e.g. laptop, tablet, smartphone) or use the shared terminal on your table
3. Log on to the Google Drive page for the course, and open the ‘Workbook’ folder for your table
4. For each week, your table should start a new Google document in here in which you write down your questions, jot down the table's thoughts, and compose your solutions to tasks. This document will be shared with the class and your table can speak to it to introduce your table's ideas.

3 Assessment

- 10% participation grade
- 20% short writing assignment (500 words)
- 20% short writing assignment (500 words)
- 50% end-of-semester essay (2,000 words)

You earn your participation grade by coming to class with questions, participating in the work at your table, and participating in class discussion. This assessed work is formative not summative: the aim is to help you understand the reading of the week and guide and focus the class discussion, not to assess your pre-class understanding of the material. The participation grade measures your willingness to contribute to this work, and contribute thoughtfully, not on the correctness or otherwise of your contributions. If you miss the class for good reason (illness, personal emergency, etc.), please let me know so I can this into account in working out your participation grade.

The writing assignments are assessed differently. Like with other Philosophy courses, these assignments are assessed using the Philosophy-specific marking guidelines. You can find a copy of these guidelines in the ‘Course Info’ folder. The guidelines emphasise correctness in your writing and good understanding of the course material (as well as rigour, precision, and clarity).

To choose a question for your writing assignments:

- For the short writing assignment, choose any of the questions we discuss in class (see in the online workbooks). For your writing assignment, you will likely need to narrow the focus of your answer to cover it in 500 words. For example, you may may say the start of your answer, 'I will focus on X here,' where X is just one issue raised by the question.
- For the end-of-semester essay, you may choose either one of the questions we discuss in class or you may come up with your own question. If you choose your own question, you must email it to me so I can confirm its suitability before you start to write.

In class, we will discuss strategies for writing excellent essays and short writing assignments (see 'Strategy for writing a 1st class essay' in the 'Course Info' folder). One important piece of advice is that depth is much more important than breadth in a philosophy writing assignment: better to have a narrow focus and cover one issue in depth rather than many topics superficially.

Regarding the reading for your essay and writing assignments, 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, visit the

https://www.cm.edu/teaching/assessment/howto/basics/formative-summative.html
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.

I’m happy to give advice on essays, and suggest appropriate readings after you have explored yourself.

Word limits for the writing assignments include footnotes, but excludes bibliography. These are hard word limits – do not go over them.

4 Contact details

You can talk to me talk about possible essay titles, plans for your essay, questions about the course, or anything else related. I am usually quick to respond over email. However, if you would rather see me face-to-face, please book to see me in my office hours. To do this, click the following link: https://calendly.com/sprevak/office-hours-meeting. If you cannot come in the listed hours, send me an email (mark.sprevak@ed.ac.uk) and we can arrange another time. My office is 5.12 in the Dugald Stewart Building

5 Reading

5.1 Background reading

A good starting point is to read one of the books listed below.

- Wallach, W., Allen, C. (2008), Moral Machines, Oxford University Press

5.2 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 with you come up with a focused question for class.

All the essential readings and as many as possible of the secondary readings are posted as PDFs in the 'Readings' folder.

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 of a passage, 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 tips for reading philosophy (as he says, ‘you should expect to read a philosophy article more than once’).

5.2.1 Week 1 – What is ethics of AI?

Essential reading:
5.2.2 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/ijcs/2012/00000019/F0020001
  - http://www.ingentaconnect.com/content/ijcs/2012/00000019/F0020007
- E. Yudkowsky. ‘Three Major Singularity Schools’, blogpost on Machine Intelligence Research Institute, September 2007
- The Singularity Film has some nice interviews with experts.

5.2.3 Week 3 – The 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. '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

5.2.4 Week 4 – Racist AI

Essential video:


Essential reading:


Secondary:

- 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

5.2.5 Week 5 – Autonomous weapons

Essential video:

- N. Sharkey, 'Killer Robots in War and Civil Society', video talk, 10 August 2015.
Essential reading:


Secondary:

- H. M. Roff & D. Danks (in press). "Trust but Verify": The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics*

### 5.2.6 Week 6 – Falling in love with AI

Essential video:


Essential reading:


Secondary:

- Devlin, K. 'In defence of sex machines: why trying to ban sex robots is wrong', *The Conversation*, 17 September, 2015
- Eskens, R. (2017) 'Is sex with robots rape?' *Journal of Practical Ethics*
- Essays in this edited collection:
- *Her* is a rather good film that explores some of these ideas

### 5.2.7 Week 7 – Humans need not apply

Essential video:

- CGP Grey, 'Humans Need Not Apply', video talk, 13 August 2014
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
- Sharkey, A. (2016) Should we welcome robot teachers?, *Ethics and Information Technology* 18, pp. 283–297

### 5.2.8 Week 8 – Teaching ethics to robots

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:


Secondary:

- Goodall, N. J. (2014) 'Ethical Decision Making During Automated Vehicle Crashes', *Transportation Research Record: Journal of the Transportation Research Board* 2424, 58–65
- Lin, P. (2016) 'Why Ethics Matters for Autonomous Cars' in M. Maurer et al. (eds.), *Autonomous Driving*
5.2.9  Week 9 – Robot rights

Essential video:


Essential reading:


Secondary:

- Singer, P. (1974) 'All Animals are Equal', Philosophical Exchange, 5, Article 6

5.2.10  Week 10 – Understanding alien minds

Essential video:

- Tufekci, Z. 'Machine intelligence makes human morals more important', TED talk, 11 November 2016
- Coldewey, D., 'Laying a trap for self-driving cars', TechCrunch, 17 March 2017

Essential reading:


Secondary:
5.2.11 Week 11 – Final report

No new readings – see the description of what we do in Week 11 above.

5.3 Useful websites

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.

- Center for Human-Compatible AI has a superb bibliography.
- Ethics + Emerging Sciences Group often links to interesting news stories.
- Ethics of AI conference at NYU streamed video discussions from October 2016.
- Future of Humanity Institute is the other big UK centre worth checking out.
- Future of Life Institute has lots of useful resources and links.
- LessWrong has interesting posts on the alignment problem and related issues in decision making and a useful wiki
- Leverhulme Centre for the Future of Intelligence is a major UK centre on this topic and it's worth checking out its events and news.
- The Machine Intelligence Research Institute has many useful publications and blog posts

5.4 Useful journals

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