Ethics of Artificial Intelligence

Course guide v1.17 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) have 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.

[1] For more information on flipped classrooms, and why one might use them instead of lectures, see: Gibbs, Graham. 'Lectures don't work, but we keep using them' *Times Higher Education*, November 21, 2013; Wilson, Karen, and James H. Korn. 'Attention during lectures: Beyond ten minutes.' *Teaching of Psychology* 34.2 (2007): 85–89; Bligh, Donald A. *What's the Use of Lectures?* Intellect books, 1998; Bishop, Jacob Lowell, and Matthew A. Verleger. 'The flipped classroom: A survey of the research.' *ASEE National Conference Proceedings* 30.9 (2013).

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. 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 summary of the main points from the reading. Note that this is not a lecture – students will be expected to contribute to the summary. During the remainder of the class, we will work through some of the questions you bring, breaking them up into smaller questions where necessary.

As part of the class, different tables will be given different tasks. A table might be asked to summarise a part of the reading, argue for a particular case (*pro* or *contra*), assess the merits of a given view, think of counterexamples to a generalisation or fallacies with an argument, or find out some information that is relevant to answering the question. Work together with others at your table to solve the task. We will share the work of each table and discuss how it helps answer our questions. Some tasks 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 good 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*
3. You should familiarise yourself with Google Docs if you are not already: https://goo.gl/vkApYk.

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:[2] 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).

I have posted some examples of questions you could use for writing assignments or essays in a 'Example questions' document in the Google drive. However, you are not restricted to these questions. You are encouraged to use the questions that have come up in class, which are listed in the workbooks. You are also welcome to think up a new question from scratch. If you create a new question, **email the question 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 breath 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 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.

[2] https://www.cmu.edu/teaching/assessment/howto/basics/formative-summative.html

3.1 Advice for short writing assignments

For your short writing assignments, you will likely need to narrow the focus of your answer significantly to cover it in 500 words. You might want to start your answer with, 'I will focus on X here', where X is just one issue raised by the question.

You should think of a short writing assignment as a miniature essay. It is assessed in the same way. Here are some dos and don'ts:

Don't:

- Attempt to summarise or engage with an entire paper – too much for a short assignment!
- Merely express likes/dislikes – justify your view with rational argument
- Make *ad hominen* attacks on the author – engage with the substance of what he/she says

Do:

- Explain the literature in your own words
- Use simple worked examples to illustrate your points and demonstrate understanding
- Focus exclusively on developing 1 (or at most 2) points – it's ok to ignore everything else
- Draw on the further readings and your own research where appropriate
- Explore a problem/question/counterexample
- Consider possible responses on behalf of the author
- Be honest if you don't understand something – but provide some careful candidate hypotheses about what you *think* it might be

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.

- Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press
- 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:

- N. Bostrom and E. Yudkowsky. 'The ethics of artificial intelligence'. In W. M. Ramsey and K. Frankish, editors, *The Cambridge Handbook of Artificial Intelligence*, pages 316–334. Cambridge University Press, Cambridge, 2014.

Secondary:

- 'Benefits & risks of artificial intelligence', Future of Life Institute
- 'Top 9 ethical issues in artificial intelligence', World Economic Forum, 21 Oct 2016
- E. Yudkowsky. 'Artificial intelligence as a positive and negative factor in global risk' In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, 308—345. New York: Oxford University Press, 2008.
- K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans. 'When Will AI Exceed Human Performance? Evidence from AI Experts', arXiv:1705.08807, May 2017 and associated blogpost on AI Impacts ('Some survey results!')
- S. J. Russell, D. Dewey, and M. Tegmark, 'Research priorities for robust and beneficial artificial intelligence', *AI Magazine*, 2015

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:

- Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-1), 7-65.

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`

- Armstrong, S., Sandberg, A. & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds & Machines* (2012) 22: 299–324.
- Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Chapters 2–6
- Chalmers, D. (2012). The Singularity: A reply to commentators. *Journal of Consciousness Studies*, 19(7-8), 141–167.
- E. Yudkowsky. 'Three Major Singularity Schools', blogpost on Machine Intelligence Research Institute, September 2007
- Good, I.J. (1965) Speculations concerning the first ultraintelligent machine, in Alt, F. & Rubinoff, M. (eds.) *Advances in Computers*, vol 6, New York: Academic
- Shanahan, M. (2015) *The Technological Singularity*, Cambridge, MA: MIT Press
- The Singularity Film has some nice interviews with experts.

5.2.3 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:

- Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds & Machines* 22: 71–85.

Secondary:

- Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Chapters 7–8, 12
- 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
- E. Yudkowsky. 'The AI Alignment Problem: Why It's Hard, and Where to Start', recorded lecture at Stanford University on May 5, 2016 for the Symbolic Systems Distinguished Speaker series.
- N. Soares, 'The Value Learning Problem'. Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016) New York, NY, USA 9–15 July 2016
- 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

5.2.4 Week 4 – Racist AI

Essential video:

- O'Neil, C. (2016). 'The era of blind faith in big data must end', TED talk, April 2017

Essential reading:

- Binns, R. (2017). 'Algorithmic Accountability and Public Reason', *Philosophy & Technology*

Secondary:

- A. Chander, The Racist Algorithm, 115 *Michigan Law Review* 1023, 1046 (2017)
- B. Goodman and S. Flaxman (2016) 'European Union regulations on algorithmic decision-making and a "right to explanation" ', arXiv:1606.08813
- H. Nissenbaum (2001). How computer systems embody values. *Computer*, 34(3), 120–119.
- J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, H. Yu, Accountable Algorithms, *University of Pennsylvania Law Review* 165 (2017)
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 1–21.
- N. Diakopoulos (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2). New York, NY, 56–62.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group.
- Pasquale, F. (2015). *The Black Box Society*, Harvard University Press
- 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:

- Sparrow, R. (2007). 'Killer robots', *Journal of Applied Philosophy*, 24, 62–77.

Secondary:

- B. J. Strawser (2010) Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles, *Journal of Military Ethics*, 9:4, 342–368
- H. M. Roff & D. Danks (in press). "Trust but Verify": The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics*
- Muller, V. C. and Simpson, T. W. 'Autonomous Killer Robots Are Probably Good News' in Ezio Di Nucci & Filippo Santoni de Sio (eds.): *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*. London: Ashgate.
- N. Sharkey (2010) Saying 'No!' to Lethal Autonomous Targeting, *Journal of Military Ethics*, 9:4, 369–383
- N. Sharkey (2012) The Evitability of Autonomous Robot Warfare, *International Review of the Red Cross*, 94/886: 787–99.
- R. C. Arkin (2010) The Case for Ethical Autonomy in Unmanned Systems, *Journal of Military Ethics*, 9:4, 332–341
- Simpson, T. W. and Muller, V. C. (2016) Just war and robot's killings, *Philosophical Quarterly*, 66 (263), 302–22

5.2.6 Week 6 – Falling in love with AI

Essential video:

- Devlin, K. (2017). 'Sex robots', TED talk, April 2017

Essential reading:

- Turkle, S. (2011), *Alone Together*, Basic Books – Introduction ('Alone Together') & Chapter 3 ('True Companions')

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:
  - Danaher, J., McArthur, N. (Eds.), (2017). *Robot Sex: Social and Ethical Implications*, MIT Press.
- Levy, D. (2009) *Love and Sex with Robots*, Duckworth
- Richardson, K. (2015). 'The Asymmetrical "Relationship": Parallels Between Prostitution and the Development of Sex Robots', *SIGCAS Computers & Society*, 45, 290–293
- Sharkey, A. (2014) Robots and human dignity: a consideration of the effects of robot care on the dignity of older people, *Ethics and Information Technology* 16, pp. 63–75
- Sharkey, A. (2016) Should we welcome robot teachers?, *Ethics and Information Technology* 18, pp. 283–297
- Sharkey, A. and Sharkey, N. (2012) Granny and the robots: Ethical issues in robot care for the elderly, *Ethics and Information Technology* 14, pp. 27–40
- Sparrow, R. (2016) Robots in aged care: a dystopian future?, *AI and Society* 31, pp. 445–454
- Sparrow, R. (2017) 'Robots, rape, and representation', *International Journal of Social Robotics* 4, 465–477
- The Verge, (2018), 'Sony's Aibo is a very good robot dog', news report, 9 January 2018
- 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
- Autor, D. H., 'Will automation take away our jobs?', TED talk, September 2016

Essential reading:

- Autor, D. H. (2015), 'Why Are There Still So Many Jobs? The History and Future of Workplace Automation', *The Journal of Economic Perspectives*, 29, pp. 3–30

Secondary:

- Arntz, M., T. Gregory and U. Zierahn (2016), 'The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis', *OECD Social, Employment and Migration Working Papers*, No. 189, OECD Publishing, Paris.
- Brynjolfsson, E. and McAfee, A. (2014). *The Second Machine Age*, WW Norton and Co.
- Collins, K. 'A programmer automated their data-entry job. Now the question is whether to tell their employer', *Quartz*, 30 June 2017
- Danaher, J. (2017) Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life, *Science and Engineering Ethics* 23, pp. 41–64
- Standage, T. 'Automation and anxiety', special report in *The Economist*, 25 June 2016

5.2.8 Week 8 – Good and bad 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:

- Allen, C., Varner, G., Zinser, J. (2000) 'Prolegomena to any future artificial moral agent' *Journal of Experimental & Theoretical Artificial Intelligence* 12, 251–261

Secondary:

- Allen, C., Smit, I., Wallach, W. (2005) 'Artificial morality: Top-down, bottom-up, and hybrid approaches', *Ethics and Information Technology* 7, 149–155
- Anderson, M., Anderson, S. L. (2010) 'Robot be good' *Scientific American* 303, 72–77
- Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352, 1573–1576.
- Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Chapter 13
- Gogoll, J., Muller, J. F. (2017) 'Autonomous Cars: In Favor of a Mandatory Ethics Setting' *Science and Engineering Ethics* 23, 681–700
- Goodall, N. J. (2014) 'Ethical Decision Making During Automated Vehicle Crashes', *Transportation Research Record: Journal of the Transportation Research Board* 2424, 58–65
- Howard, D., Muntean, I. (2017) 'Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency' in T.M. Powers (ed.), *Philosophy and Computing*, Philosophical Studies Series 128
- Lin, P. (2016) 'Why Ethics Matters for Autonomous Cars' in M. Maurer et al. (eds.), *Autonomous Driving*
- Nyholm, S., Smids, J. (2016) 'The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?' *Ethical Theory and Moral Practice* 19, 1275–1289
- Savulescu, J., Giubilini, A. (2017) 'The artificial moral advisor' *Philosophy & Technology*
- Savulescu, J., Maslen, H. (2015) 'Moral Enhancement and Artificial Intelligence: Moral AI?' in J. Romportl et al. (eds.), *Beyond Artificial Intelligence*
- Wallach, W., Allen, C. (2008), *Moral Machines*, Oxford University Press
- Yudkowsky, E. (2004), 'Coherent extrapolated volition', Machine Intelligence Research Institute
- The Moral Machine is a website with an interesting collection of moral dilemmas

5.2.9 Week 9 – Robot rights

Essential video:

- Cohen, G. A.I. Ethics: Should We Grant Them Moral and Legal Personhood?, video posted to YouTube, 23 September 2016
- Darling, K. (2015) 'Children Beating Up Robot Inspires New Escape Maneuver System', *IEEE Spectrum*, 6 August 2015

Essential reading:

- Korsgaard, K. M. (2004). 'Fellow Creatures: Kantian Ethics and Our Duties to Animals', in *The Tanner Lectures on Human Values*, Grethe B. Peterson (ed.), Volume 25/26, Salt Lake City: University of Utah Press.

Secondary:

- Bryson, J. J. (2010) 'Robots Should Be Slaves' in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issue*, Yorick Wilks (ed.), pp 63-74
- Dennett, D. C. (1978) 'Why you can't make a computer that feels pain'. *Synthese* 38, 415–449.
- Gruen, L. (2017) 'The Moral Status of Animals', *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.)
- Kagan, S. (2016), 'What's Wrong with Speciesism?' *Journal of Applied Philosophy* 33, 1–21 and responses in the same journal issue:
  - http://onlinelibrary.wiley.com/doi/10.1111/japp.2016.33.issue-1/issuetoc
- Singer, P. (1974) 'All Animals are Equal', *Philosophic Exchange*, 5, Article 6
- Singer, P. (1993), *Practical Ethics*, second edition, Cambridge: Cambridge University Press; first edition, 1979.
- Singer, P. (2009), 'Speciesism and Moral Status', *Metaphilosophy*, 40, 567–581
- Solum, L. B., (1992) 'Legal Personhood for Artificial Intelligences', *North Carolina Law Review* 70, 1231–1287
- Wegner, D. M., Gray, K. (2016), *The Mind Club*, Penguin Books

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:

- Lipton, Z. C. (2017) 'The mythos of model interpretability', arXiv:1606.03490v3

Secondary:

- Amodei D., Olah, C., Steinhardt, J., Christiano, C., Schulman, J., Mané, D. (2016) 'Concrete Problems in AI Safety', arXiv:1606.06565
- Athalye, A. 'Robust Adversarial Examples', blogpost on OpenAI, 17 July 2017
- Bostrom, N. (2009) 'Pascal's mugging' *Analysis* 69, 443–445.
- Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015) 'Explaining and harnessing adversarial examples', arXiv:1707.07397v2
- Goodfellow, I., Papernot, N., Huang, S., Duan, V., Abbeel, P. & Clark, J. 'Attacking Machine Learning with Adversarial Examples', blogpost on OpenAI, 24 February 2017
- Karpathy, A. 'Visualizing what ConvNets learn', post on Stanford CS class CS231n: Convolutional Neural Networks for Visual Recognition
- Kurakin, A., 'Adversarial Examples In The Physical World - Demo', YouTube video to accompany paper arXiv:1607.02533, 1 November 2016
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J. (2017) 'Building machines that learn and think like people', *Behavioral and Brain Sciences*
- Leike, J. et al, (2017) 'AI Safety Gridworlds', arXiv:1711.09883v2

- MITCSAIL, 'Fooling Image Recognition with Adversarial Examples', YouTube video to accompany paper arXiv:1707.07397v2, 2 November 2017
- Mordvintsev, A., Olah, C., Tyka, M. 'Inceptionism: Going Deeper into Neural Networks', blogpost on Google Research Blog, 13 July 2015
- Ribeiro, M. T. et al (2016) ' "Why Should I Trust You?" Explaining the Predictions of Any Classifier', arXiv:1602.04938v3

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*