Omar Fakhri
Ethics of A.I.

Course Description

One of the first digital computers, the ENIAC, was developed just about seven decades ago. This computer wasn't programmable, weighed about 50 tons, and took up about 1,8000 square feet. Just two years ago, Facebook had to shut down an AI because it had developed its own language that humans couldn't understand. We can fit a terabyte of information in a device that is as small as my pinky finger. We've come a long way in such a short time. How far will technology advance in the next few decades? Is it possible that AIs will eventually become so advanced that they become smarter than us?

In this upper division course on the ethics of artificial intelligence, we will begin the course by talking about preliminary issues, such as: can machines think, and if they can how can we find out? Can computational machines emulate human cognitive activity? Then, we will examine the so-called "singularity," and different arguments for it. This is the idea that given the rapid advancements in technology, we will see unexpected consequences. One argument for the singularity goes as follows: Humans might eventually build machines that are smarter than us, but these machines will be able to build machines that are even smarter than themselves. And so on. There might be a limitation to this process due to the laws of nature and the limited resources in our universe. However, it is safe to say that these advance machines will bring about unexpected consequences that we need to seriously consider.

From here, we will focus our discussion on the ethical issues that might arise if there was a singularity. In particular, we will look at issues such as human enhancement and impairment, eliminating death, living with A.I., the ethical status of A.I., and uploading our minds to machines. Unlike other questions in philosophy, philosophers haven't spent two thousand or so years trying to answer ethical questions about A.I., although some of the issues we will discuss do overlap with traditional issues in philosophy. This should be encouraging to you because the possibility of producing an original contribution is quite high.

Course Requirements:

- Biweekly Assignments 40% – These are short reaction papers between 400-500 words. You are expected to summarize an important part of the reading and then critically evaluate it.

Pick only **one** of the following requirements (note: for those interested in applying to graduate programs in philosophy, I highly recommend doing the latter option):

- Three Papers 20% each – These are shorter papers, about 5-7 double-spaced pages. Prompts will be handed out a week before the paper is due.

Or

- Long Paper 60% - This is a substantial paper, about 15 double-spaced pages. You will be required to get your paper topic approved by me first. Ideally, you should aim to do this at least a month before the paper is due. This paper will engage with a big bulk of the assigned readings and perhaps some outside sources as well. If you decide to write this paper instead

of the three short papers, please let me know as soon as you make this decision. I will provide extensive comments on this paper, and I would be happy to read future drafts of it, even after the class is done.

Required Text:

There will not be a required text. The reading will be distributed via the course website.

Course Schedule

Week 1: Turing test; schools of singularity
Turing: Computing Machinery and Intelligence
Yudkowsky: Three Major Singularity Schools

Week 2: Singularity: intelligence explosion
Vinge: The Coming Technological Singularity
Good: Speculations Concerning the First Ultraintelligent Machine
Optional: Kurzweil: *The Singularity is Near*, selections

Week 3: Singularity: speed explosion
Solomonoff: The Time Scale of Artificial Intelligence - Reflections on Social Effects
Yudkowsky: Staring at the Singularity

Week 4: Minds can't be emulated by computational machines
Dreyfus: *What Computers Can't Do*, selections
Penrose: *Shadows of the Mind*, selections
Optional: Lucas: Minds, Machines, and Godel
Optional: Block: Psychologism and Behaviorism

Week 5: Minds can be emulated by computational machines
Chalmers: Minds, Machines, and Mathematics
Sandberg & Bostrom: Whole Brain Emulation: A Roadmap
Optional: Chalmers: *The Conscious Mind*, chapter 9

Week 6: Human enhancement
Savulescu: Genetic Interventions and the Ethics of Enhancement of Human Beings, pp. 1-7
Bostrom: In Defense of Posthuman Dignity, pp. 6-11

Week 7: Human impairment
Levy: Deafness, Culture, and Choice
Anstey: Are Attempts to Have Impaired Children Justifiable? (Start)

Week 8: Human impairment (continued)
Anstey: Are Attempts to Have Impaired Children Justifiable? (finish)
Cullen, Dasgupta, and Levi: A Puzzle About Impairment

Week 9: Eliminating Death
Bostrom: The Fable of the Dragon

Kagan: The Badness of Death

Week 10: Living with artificial intelligence

Chalmers: The Singularity, pp. 1-33
Bostrom and Yudkowski: The Ethics of Artificial Intelligence, pp. 1-3, 14-18 (start)

Week 11: Living with artificial intelligence (continued)

Bostrom and Yudkowski: The Ethics of Artificial Intelligence, pp. 1-3, 14-18 (finish)
Harari: Universal Basic Income is Neither Universal nor Basic

Week 12: The ethical status of artificial intelligences

Bostrom and Yudkowski: The Ethics of Artificial Intelligence, pp. 6-9.
Schwitzgebal and Garza: A Defense of the Rights of Artificial Intelligences, pp. 98-103, pp. 107-111

Week 13: Mind Uploading

Chalmers: The Singularity, pp. 33-40
Kagan: The Value of Life

Week 14: Extra time