CBMM Retreat, April 2016, MIT

Brainstorming/Discussion Session: Ethical issues for AI and CBMM

Panel moderators: Matt Wilson, Max Tegmark

Synopsis: In this session we will discuss some of the issues raised by the development of AI and the resulting applications in society. Suggested topics include Technology and Society, Autonomy and Critical Decision-making (autonomous vehicles, medical diagnostics, lethal autonomous weapons), Relation to CBMM mission, Public outreach.

Below is some suggested reference material to look at to stimulate discussion. The listed times are just estimates for convenience and do not reflect the time that will be spent in the session.

Readings and links (1 hour total):

Video

Humans need not apply

https://www.youtube.com/watch?v=7Pq-S557XQU&feature=youtu.be

Short, entertaining video highlighting the issue of worker displacement through technology (running time 15 mins)

Commentary

Robotics: Ethics of Artificial Intelligence

http://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611

Perspectives from several science and technology researchers on societal concerns regarding intelligent machines (reading time 10 mins)

Extended discussion

The Ethics of Artificial Intelligence

https://intelligence.org/files/EthicsofAI.pdf

Wordy but interesting in-depth discussion of ethical issues raised by AI (reading time 30 mins)

Research Projects

Future of Life 2015 Project Grants

http://futureoflife.org/AI/2015awardees

Example of funded research in the area of ethics and AI (reading time 3 mins)

Outreach

Stanford AI100 project

http://news.stanford.edu/news/2014/december/ai-century-study-121614.html

Announcement of the long-term Stanford project assessing societal impact of AI (reading time 2 mins)

Issues for discussion:

Transparency – predictability

It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also transparent to inspection—to name one of many socially important properties. Unfortunately, the better the AI system, the harder it often is to explain. The features that contribute to probability-based assessments such as Bayesian analyses are straightforward to present; deep-learning networks, less so. (Altman, Nature, 2015)

Morality in critical decision making

How should robots ethically decide, as they develop autonomy and free will, what to do?

Technology and job displacement

Disparity in access to AI technology

AI and liability

When an AI system fails at its assigned task, who takes the blame?

Public outreach

Experts need to become the messengers. Through social media, researchers have a public platform that they should use to drive a balanced discussion. (Hauert, Nature, 2015)

Development of Artificial General Intelligence (AGI): It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance.

Thus the discipline of AI ethics, especially as applied to AGI, is likely to differ fundamentally from the ethical discipline of noncognitive technologies, in that:

- The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmers do everything right;
- Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system's safe behavior in all operating contexts;
- Ethical cognition itself must be taken as a subject matter of engineering.

(modified from Bostrom, The ethics of artificial intelligence, 2011)

A *superintelligence* is any intellect that is vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills.

A prerequisite for having a meaningful discussion of superintelligence is the realization that superintelligence is not just another technology, another tool that will add incrementally to human capabilities. Superintelligence is radically different.

Superintelligence may be the last invention humans ever need to make.

Technological progress in all other fields will be accelerated by the arrival of advanced artificial intelligence.

Superintelligence will lead to more advanced superintelligence.

Artificial minds can be easily copied.

Emergence of superintelligence may be sudden.

Artificial intellects need not have humanlike motives.

Artificial intellects may not have humanlike psyches.

(modified from Bostrom, **Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence**, 2003)