EU Parliament, Committee on Legal Affairs:

Hearing on
Legal and Ethical Aspects of Robotics and Artificial Intelligence

Robotics, artificial intelligence and risk management

Olle Häggström

http://www.math.chalmers.se/~olleh/

http://haggstrom.blogspot.se/

Developments in artificial intelligence (AI) and robotics in coming years are expected to have vast impact on the economy. In the long run, the possibilities for AI technology to bring prosperity are virtually unboundend (other than by the laws of physics).

But there are also (at least) the following three reasons to be concerned:

1. Can autonomous drones and other military AI technology become dangerous, for instance by falling into the hands of terrorists?
2. Robosourcing – will robots outcompete us on the labor market?
3. When AI has reached the level where we humans are no longer the most intelligent beings on Earth, can we expect to remain in control?

All three are important topics, but in this talk I will focus entirely on the most radical issue 3.

Question: When AI has reached the level where we humans are no longer the most intelligent beings on Earth, can we expect to remain in control?

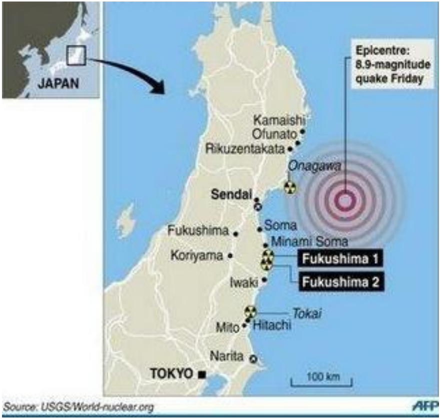
Short answer: No.

Consequence: Our fate at that point will depend on the intentions of the machines. So we'd better make sure (in advance) that those intentions are well-aligned with what we value, such as human welfare.

Three levels of disaster

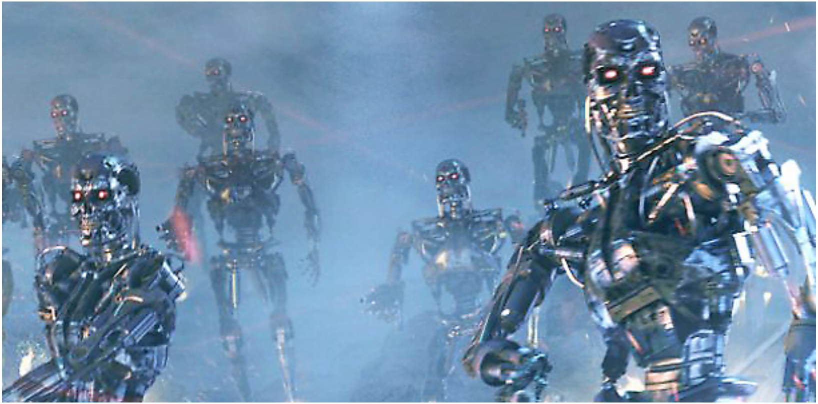
Level 1

Level 2



Source: USGS/World-nuclear.org

Level 3



"Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all. [...] It's tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake in history."

Stephen Hawking
Stuart Russell
Max Tegmark
Frank Wilczek
in *The Independent*, May 2014

Does humanity really risk getting wiped out by highly intelligent machines? To answer this, there are two main issues that we can try to separate out from each other:

(1) Will human level AI be exceeded?

(2) What will a superintelligent AI be inclined to do?

The first of these – question (1) concerning if and when AI capability will exceed human-level intelligence – can be further broken down into:

(1a) Will human level AI be reached? When?

(1b) From there, superintelligent AI? How fast?

A possibly illuminating abstract way to think about **(1)** is in terms of three magnificent libraries:

- **Library of Babel** (Borges 1941)
- **Library of Mendel** (Dennett 1995)
- **Library of Turing** (Husfeldt 2015)

Concerning **(1b)**, rapid developments towards superintelligence levels popularly go under the name **the Singularity**. The crucial theoretical issue for understanding how likely such scenarios are is (as emphasized by Eliezer Yudkowsky in his seminal 2013 manuscript *Intelligence Explosion Microecomonics*) that of whether returns on cognitive reinvestment are increasing or decreasing.

Why would an AI *want* to self-improve or (more or less equivalently) create even better AI's? This brings us to issue **(2)** – what will a superintelligent AI be inclined to do?

The best theoretical framework currently available for addressing this issues is the Omohundro–Bostrom theory of **ultimate** versus **instrumental** goals.

The orthogonality thesis: Virtually any ultimate goal is compatible with arbitrarily high levels of intelligence.

The instrumental convergence thesis: There are a number of instrumental goals that a sufficiently intelligent AI is likely to set up to help promote their final goal, pretty much no matter what the final goal is.

Some basic instrumental goals for which the instrumental convergence thesis seems to apply:

- Self-preservation (don't let them pull the plug on you).
- Acquisition of hardware (and other resources).
- Improving one's own software and hardware.
- Preservation of final goal.

How can we avoid **Paperclip Armageddon**?

How to avoid Paperclip Armageddon (or something similarly bad)?

We are very far from being able to answer that question today.
But the following seems like a good idea:

Let us try to redirect at least some of the (enormous) effort that is today put into making AI **as capable and powerful as possible** towards instead making it **safe** and ensuring that its power has **consequences beneficial to humanity**.

It seems to be crucially important that a superintelligent AI has goals and values that are well-aligned with our own.

However, because a superintelligent AI is likely to employ **perservation of final goal** as an instrumental goal, it is unlikely to permit us to tamper with its ultimate goal.

But perhaps that instrumental goal can be made to work to our advantage. Could we install suitable values into the AI before it reaches superintelligence levels? This is a key idea behind so-called **Friendly AI** proposals, advocated by Yudkowsky, Bostrom and others.

Friendly AI seems very difficult, partly because human values seem very fragile. Getting them only 99% right is likely to lead to so-called **perverse instantiation**. Most of Isaac Asimov's robot stories deal with such scenarios.