Minds and Machines (Lecture 8): Superintelligence

Prof. Ioannis Votsis

Philosophy Faculty
ioannis.votsis@nulondon.ac.uk
www.votsis.org

Introduction

Intelligence revisited

- To fathom what it would mean for some person or thing to be super-intelligent, we first need to understand intelligence.

- You may recall that, in spite of various disagreements, the following characterisation was not half bad:

  "… a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredson 1997: 13).

- Having said this, some alterations are needed from the get-go to make the characterisation more inclusive.

Intelligence generalised

- A mental life but also experience may be unique to humans or certain kinds of biological systems.

- We could replace *mental* and *experience* with concepts that cover both biological and artificial intelligence. For example:

  * **information-processing** (in place of mental)
  * **detection** (in place of experience)

- The resulting characterisation would thus look as follows:

  '… a very general [information-processing] capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from [detections]'

The Singularity

- Customarily, it is understood as that point in time at which human intelligence is surpassed.

  "… the future point at which artificial intelligence exceeds human intelligence" (Bringsjord and Govindarajulu 2018)

- Having said this, sometimes scholars understand it as that point in time at which superintelligence is reached.

  "Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months— three months — 1.5 months … Singularity." (Yudkowski 1996, quoted in Chalmers 2010).

What is superintelligence?

- Conceptions of superintelligence typically have human intelligence as a starting point (and point of contrast) :

  "We can tentatively define a superintelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*" (Bostrom 2014: 22) [original emphasis].

- Given what we said earlier, this means far superior abilities to:

  * reason
  * solve problems
  * comprehend complex ideas
  * plan
  * think abstractly
  * learn fast (& via detections)

Superintelligence: Machines vs. humans

- Whenever we hear about superintelligence, we immediately assume that this is tantamount to machines ‘on steroids’.

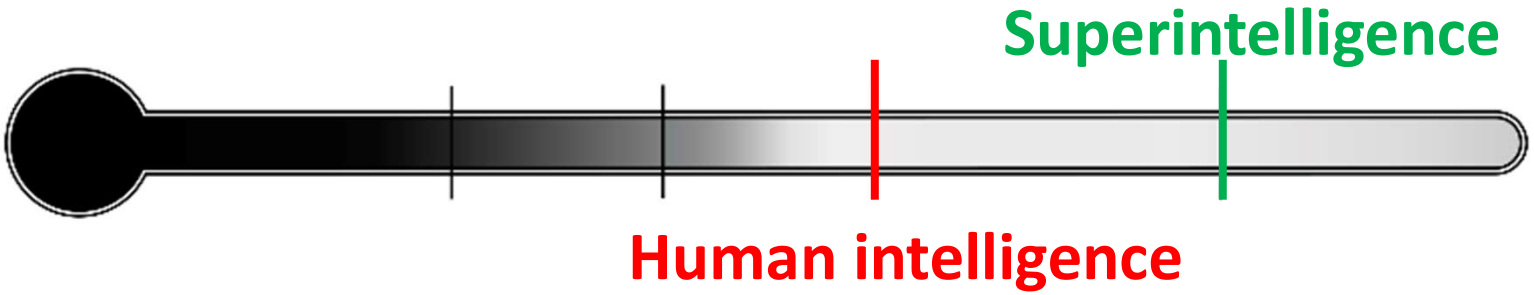
- Strictly speaking, this is a mistake! As the above conception clarifies, humans are not excluded from superintelligence (SI).

- **Kinds of SI:**

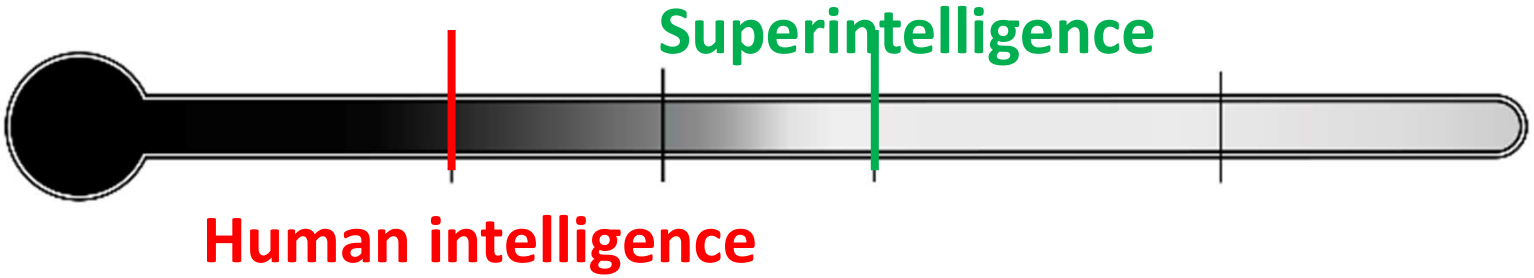
  * individual machine intelligence
  * collective machine intelligence
  * individual humans genetically re-engineered
  * individual humans technologically augmented
  * collective human intelligence
  * exoplanet biological intelligence
  * …

Superintelligence in context

- How super is super depends on where we place human intelligence.
- Since we are at the apex of intelligence on planet Earth, we typically suppose that the spectrum looks like this:



- But it could very well be the case that human intelligence is way below average in the grand scheme of things.



Plan

- In what follows, we try to motivate SI and consider a special problem that arises in the context of such intelligence.

- To be precise:

  * we examine two arguments for the emergence of SI (one that assumes AGI and one that doesn't).

  * we identify different pathways to SI (AI pathways vs. Human pathways).

  * we discuss the control problem (powers and goals).

Two Arguments for Superintelligence

The argument from hardware speed

- We have already seen this Moore's-law-inspired argument in the quotation from Yudkowsky. Here's an adaptation:

  1. Besides having the required speed, all other stumbling blocks to superintelligence will be removed before long.
  2. Superintelligence requires *X* level of computing speed.
  3. The current level of computing speed is *Y* (where $Y < X$)
  4. The level of computing speed doubles every two years.
  5. The X-Y difference is some (not many) orders of magnitude.

  ---

  6. Absent interfering factors, superintelligence will be reached before long.

The argument from intelligence explosion

- Chalmers (2010: 12) presents an argument that is heavily influenced by I. J. Good (1965). Here's an adaptation:

  1. There will be AGI (before long, absent defeaters).
  2. If there is AGI, there will be AGI+ (soon after, absent defeaters).
  3. If there is AGI+, there will be AGI++ (soon after, absent defeaters).

  ---

  4. There will be AGI++ (before too long, absent defeaters).

  NB: AGI++ is here taken to equal SI.

Superintelligence: Some Pathways

Multiple pathways to superintelligence

- On the supposition that superintelligence is achievable, how likely is it that it'll be achieved in the not-too-distant future?

- That depends on the pathway we take towards it and the requisite technology to develop it.

- In what follows, we consider the following:

  * Current neural net approach
  * Evolution mimicking
  * Brain emulation

  } **AI Pathways**

  * Augmentation
  * Collective re-organisation

  } **Human Pathways**

AI pathway: Mimicking evolution

- Evolution already designed intelligent beings, viz. ourselves. We can mimic evolution to do the same (Moravec 1976).
- **Recipe**: Start with a simple system and use *recombination*, *mutation* and *selection* to produce superintelligence.
- Evolutionary algorithms is an area of AI that mimics evolution in the aforementioned way (Russell and Norvig 2020: 4.1.4).
- Time and speed are obvious problems with this approach:

  "One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages" (Turing 1950: 456).

AI pathway: Brain emulation

- Another pathway to superintelligence is through brain emulation, whether in whole or in part.

- **Recipe**: Scan a human brain, create a virtual copy and then artificially spruce it up all the way to superintelligence.

- This pathway depends on several advances, incl.:

  * much more fine-grained + real-time scanning instruments
  * know-how to turn scans into a virtual artificial neural net
  * powerful hardware that runs the virtual brain
  * methods to spruce up the virtual brain

- "… the emulation path will not succeed in the near future (within the next fifteen years, say)" (Bostrom 2014: 36).

Bostrom's take-home message

- The existence of many pathways implies a higher (note: not necessarily high) probability that superintelligence emerges.

- Progress on some pathways is likely to lead to progress on other pathways.

  Example: Brain augmentations can help accelerate progress in reaching superintelligence via some AI pathway.

- "True superintelligence… might plausibly first be attained via the AI path [i.e. through something like ML]" (50).

- We should care about how we get to superintelligence as it might affect how much control of it we end up having.

Surveys: Müller and Bostrom

- Müller and Bostrom (2016) surveyed ca. 550 experts from CS, Biol./Phys./Neurosc., Psych./CogSci, Phil., Math/Physics.

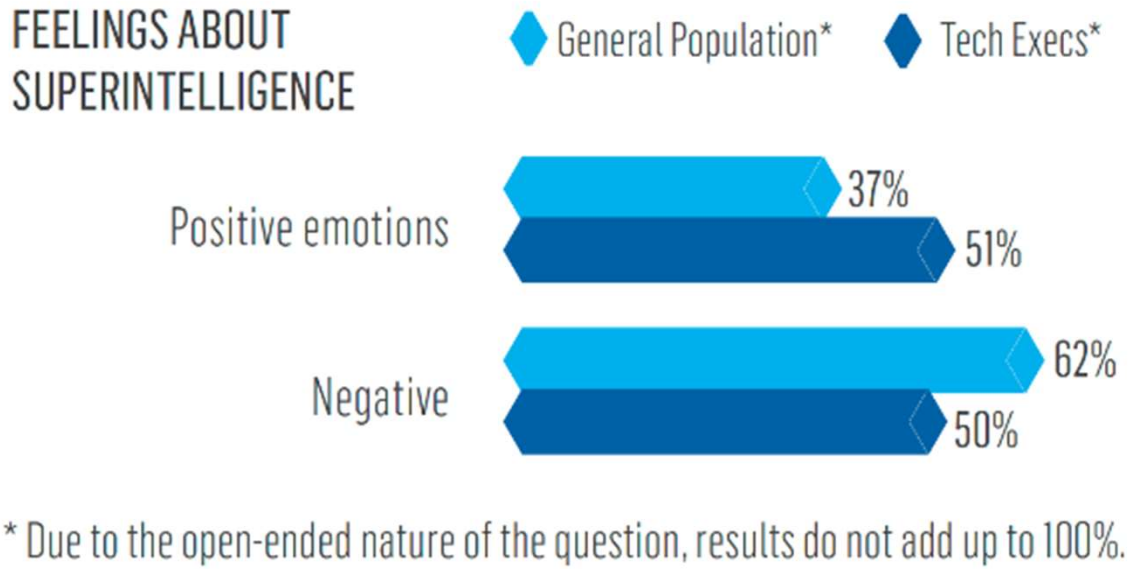
  "Define a 'high–level machine intelligence' (HLMI) as one that can carry out most human professions at least as well as a typical human."

90% chance to reach	**Median**	**Mean**	**St. Dev.**
HLMI by year:	2075	2183	396
Probability to reach			
HLMI to SI within 30 years:	0.75	0.62	0.35

Surveys: Edelman AI survey (2019)

- 300 tech executives; 1000 US adults general population.

  Singularity (AGI conception): "… 61 percent of the general population and 73 percent of tech execs saying this moment will arrive within 10 years" (p. 11).



The Control Problem

Framing the issue

- **Control problem**: “the unique principal-agent problem that arises with the creation of an artificially superintelligent agent” (Bostrom 2014: 127).

- It can be split into two problems involving conflict between: (1) sponsors-developers, (2) sponsors-superintelligence.

- In what follows, we focus on the second problem since it offers up a new challenge.

- Two broad types of methods to address this problem:

  * capability control methods
  * motivation selection methods

Controlling its abilities

- Capability control methods aim to restrict the powers of a superintelligent agent.

- There are at least three types of measures that can be taken:

  * boxing measures (physical, informational)
  * stunting measures (intellectual abilities/modalities)
  * tripwire measures (behaviour, ability, content)

- Some of these measures may be easier to implement and some will offer more protection than others.

- No guarantee that we can control the capabilities of such an agent. Moreover, some controls would render it useless.

Controlling its motivations

- Motivation selection methods aim to restrict the motivations of a superintelligent agent.

- There are at least two types of measures that can be taken that involve the specification of rules/values/goals:

  * direct specification
  * indirect specification

- A number of potential difficulties in:

  (1) specifying rules/values/goals explicitly
  (2) interpreting and implementing them
  (3) verifying that the indirectly-sourced ones are apt

The End