Science, Technology and Humanity: Opportunities and Risks

Superintelligence, singularity and after-human era

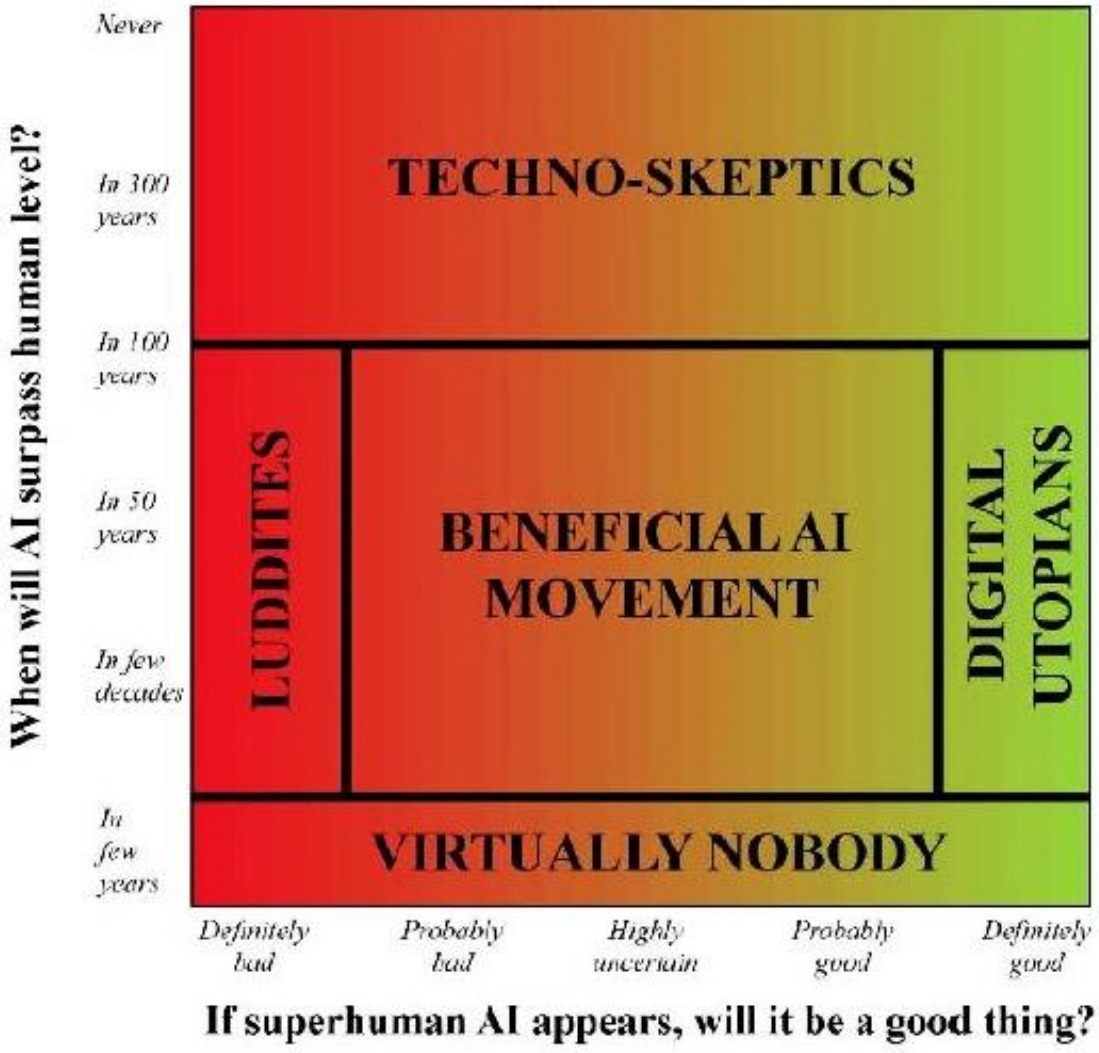
Martin Takáč, Tomáš Gál

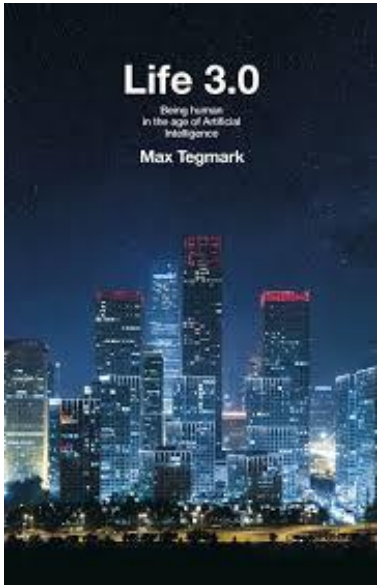
http://dai.fmph.uniba.sk/courses/STH/

Terminology

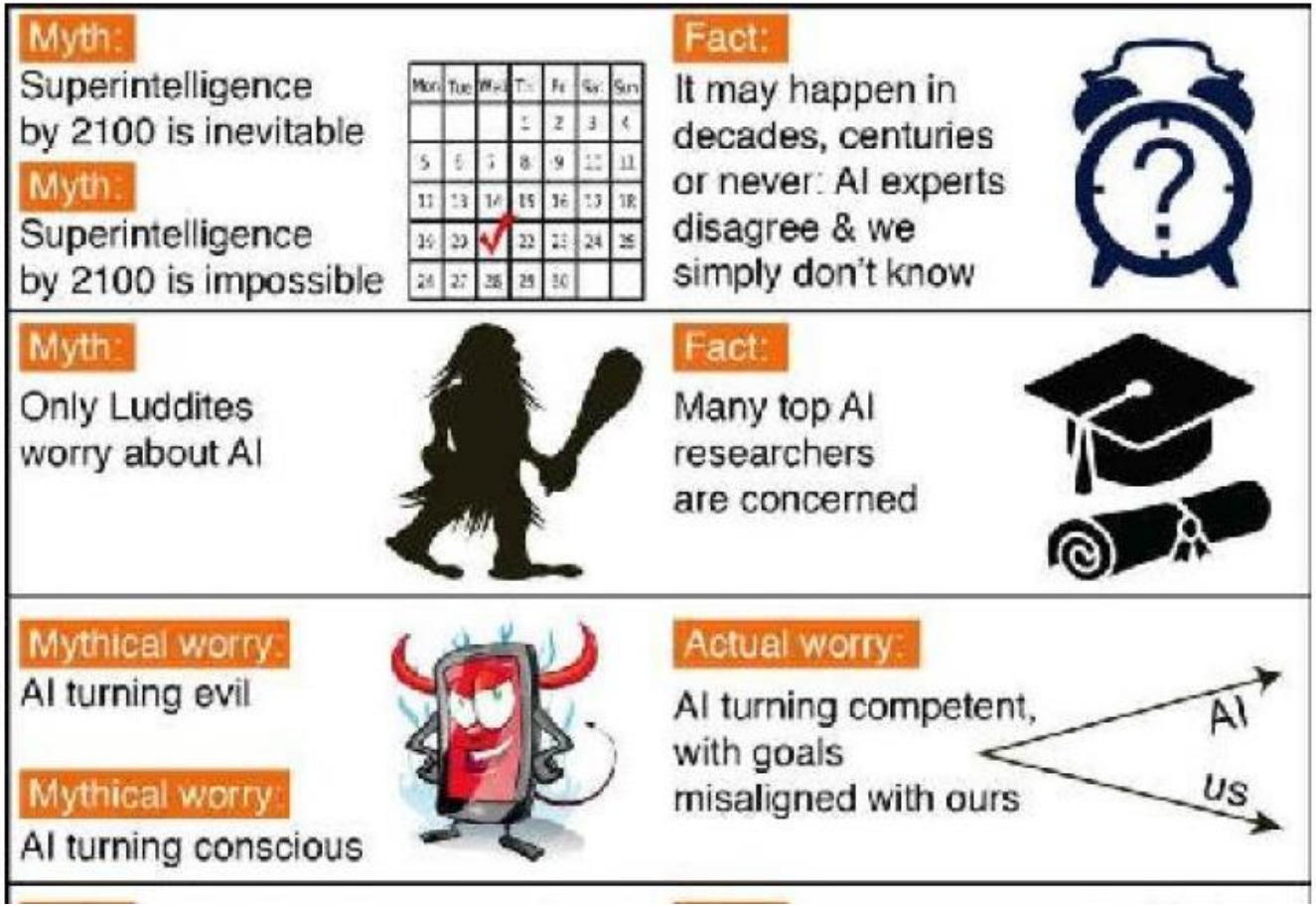
- **Human-level intelligence** – performance-matched with humans.
- There are tasks on which AI already reached or surpassed human level of performance
- But these are *specialized* tasks, vs matching humans in *all* tasks – **human-level artificial general intelligence** (AGI).
- **Human-level** doesn't mean **human-like** (can be radically different from human intelligence).
- By **superintelligence** we usually mean surpassing humans in all tasks, i.e. a superhuman-level AGI.

Attitudes to superintelligence prospects



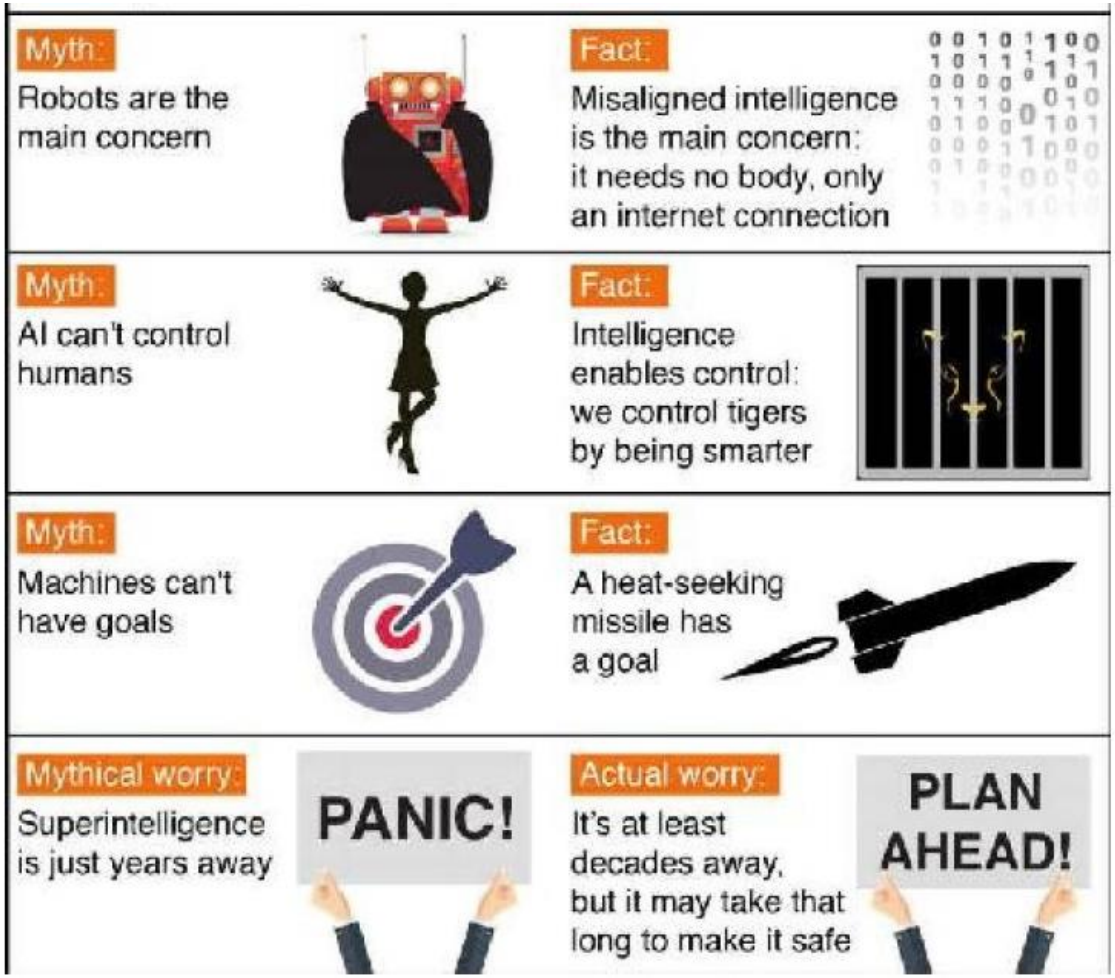


(Some) myths about superintelligence



- Image from: Tegmark, M. Life 3.0. Penguin Random House, 2017.

(Some) myths about superintelligence



- Image from: Tegmark, M. Life 3.0. Penguin Random House, 2017.

Arguments for superintelligence

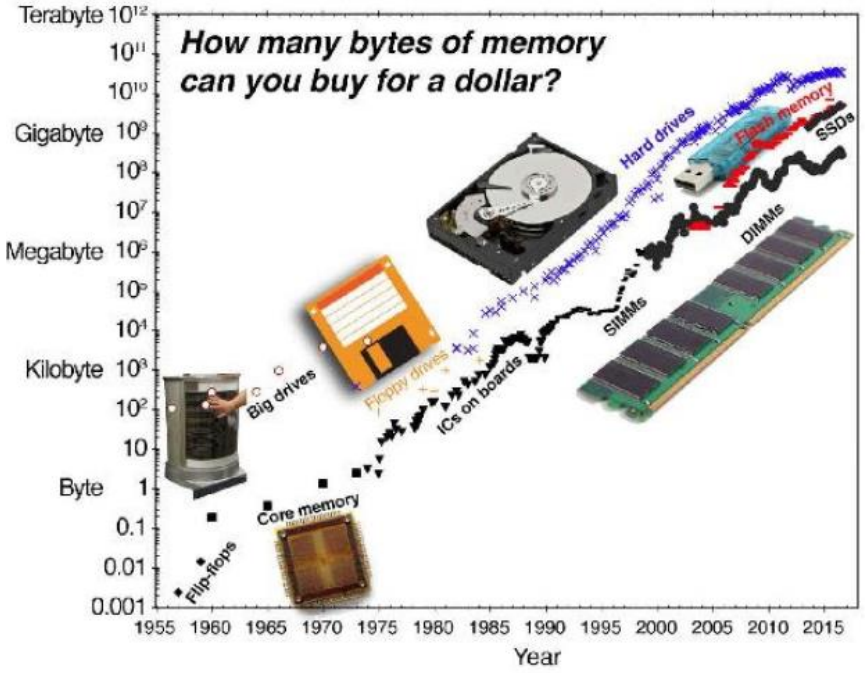
1. Human-level AGI can self-improve.
2. Once it creates as slightly better AGI, the second AGI will also create slightly better AGI, etc., which leads to recursive self-improvement.
3. Not being limited by constraints of biology, intelligence improvement will follow an *exponential curve.*
   - AGI can spawn multiple copies, search the space of solutions in parallel, split, merge and terminate processes in the speed of billions operations per second.
   - Other technologies have followed an exponential curve as well.

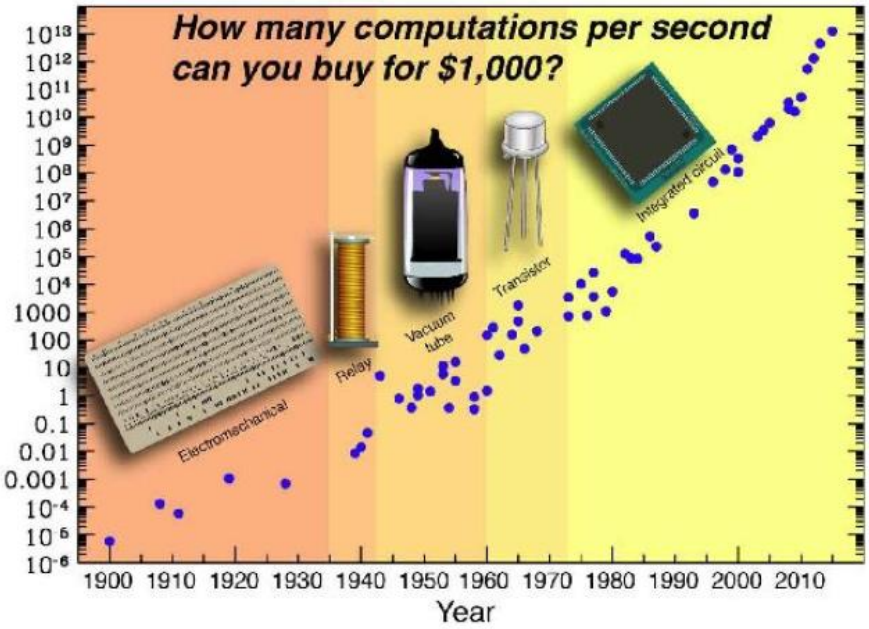
Moore's law

- "The number of transistors that can be fabricated on a single chip doubles approx. every 18 months" (G.E. Moore, 1965)
- CPU clock speed, network bandwidth, DNA sequencing speed and inverse cost, brain scanning resolution – these all have been increasing exponentially (as recognized by R. Kurzweil in 2005)

Moore's law

- Not just brute force, but conceptual inventions and new types of technology

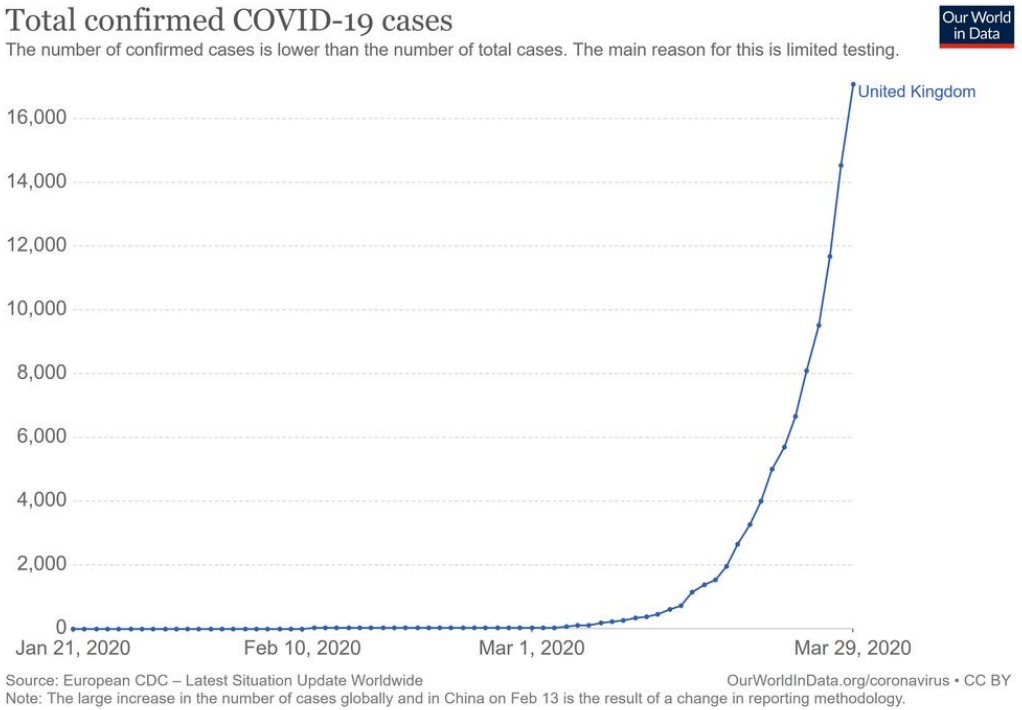




- Images from: Tegmark, M. Life 3.0. Penguin Random House, 2017.

Law of accelerating returns (Ray Kurzweil)

- "the more you have, the faster it grows" – positive feedback loop
- $Y(t+1)=R * Y(t)$
- $Y(t)=R^t * Y(t_0)$

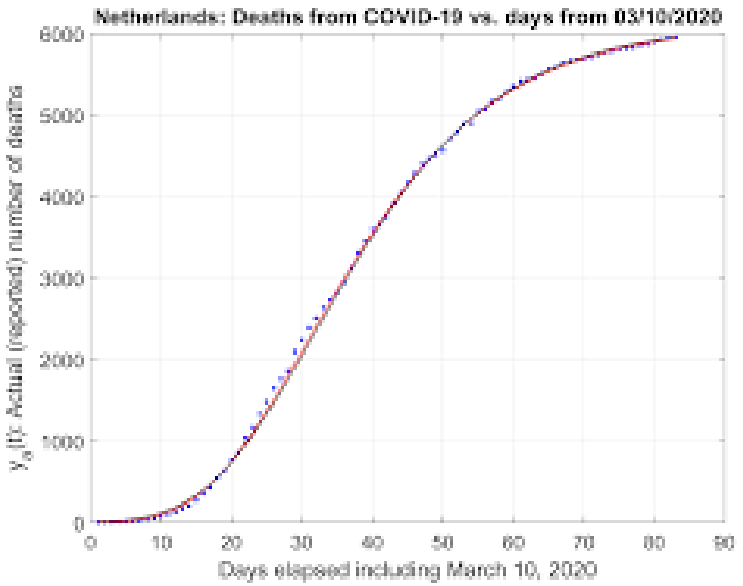


Singularity

- Metaphor from physics: a **singularity** is a point in space or time where mathematics breaks down (and our intuitions as well) – e.g. Big Bang or the center of a black hole
- **Technological singularity** – happens if exponential technological progress brought about such a dramatic change that human affairs as we understand them (economy, law, society, human autonomy etc.) came to an end

The law of diminishing returns

- Nothing in the nature grows forever
- At a certain size, negative feedback loops that flatten the curve will be activated (economy, infections, urbanization) – leading to a sigmoid (saturation)



Singularity might be a myth

- But superhuman intelligence may occur before the curve flattens.
- Nevertheless, its consequences would be so profound that it makes sense to prepare for it, even if it had a low probability

Amara’s Law

- "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run." (Roy Amara, a cofounder of the Institute for the Future, Palo Alto)

Importance

- *"Everything civilization has to offer is the product of our intelligence; gaining access to **considerably greater intelligence** would be the biggest event in human history"*

Stuart Russell

Possible roads to superintelligence

- Whole brain emulation
- Engineered AI

Whole brain emulation

1. **Mapping:** map the brain of a subject at submicron spatial resolution (connectome blueprint)
2. **Simulation:** build a real-time simulation of the electrochemical activity of the connectome
3. **Embodiment:** interface the simulation to the external environment (might need copying the body as well)

Some (sci-fi) tricks to map the brain (Shanahan, 2015)

- Structural scan (“slice and scan”)
- Use genetically modified brain so that its neurons produce a dye that fluoresces when they fire. Then use the fluorescence microscopy to record neural dynamics.
- Genetically modify the brain so that each neuron contains a unique sequence embedded in its DNA (“a barcode”). Then infect the brain with a virus engineered to carry genetic material across synapses, resulting in unique pre-post synaptic codes. Use DNA sequencing to get a neuron-level connectome.

Brain simulation

- E.g. Hodgkin-Huxley model
- Needs a supercomputer
  - E.g. Sunway Taihu Light – worlds fastest supercomputer (as of 2016) performs 10^{17} FLOPS.
  - Alternative: **neuromorphic hardware** that uses analogue computations (e.g. representing membrane potential with a real physical quantity of continuously changing charge)
- Blue brain project – started with ambition of building a human-scale brain simulation. Now the goal is more modest (a mouse brain).

Embodiment

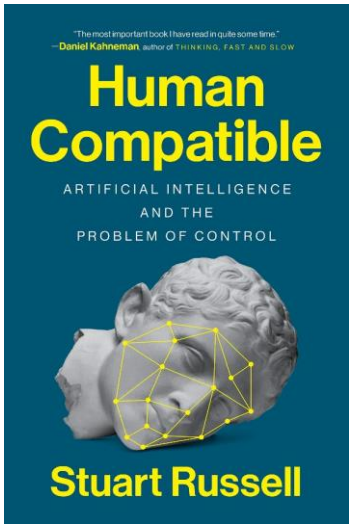
- Experience from neuroprosthetics
- Machine learning can be used to adjust the body and brain to each other.
- Alternative: virtual embodiment

Engineered AI

- See the lecture on AGI
- Reinforcement learning:
  - What is the reward function?
  - How does it learn?
  - How does it optimize (maximizes the long-term reward)?

Goal alignment

- Recall from the lecture on values:
  - Machines can be very effective in achieving their goals
  - Following a goal regardless of a context is dangerous (Paperclip example)
  - Why should machine care about the goals it's not told to care about?



Existential risk

- The problem is not a malicious AI, but an AI to whose reward maximization we get in the way
- *Convergent instrumental (sub)goals* (Bostrom, 2014)
  - self-preservation
  - resource acquisition
- Eliezer Yudkowsky: "The AI neither hates you, nor loves you, but you are made out of atoms that it can use for something else."
  - analogy to how we treat animals

Moral constraints

- inbuilt high negative rewards for actions that violate moral constraints
- being taught by human educators
- learn them by observing humans (S. Russell: inverse RL)
- limiting of AI power (children have no way to bypass the training process because their power is limited in comparison to adults). We don't give children guns or a driving licence.

Superintelligence scenarios (Tegmark)

1. Fast take-off and a single superintelligence
2. Slow take-off and multiple superintelligences

- Competition or collaboration?
- Nash equilibria and hierarchical organization
- Problem of self-control limits the effective size of the hierarchy

Superintelligence scenarios (Tegmark)

- **Libertarian utopia:** Humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property rights.
- **Benevolent dictator:** Everybody knows that the AI runs society and enforces strict rules, but most people view this as a good thing.
- **Egalitarian utopia:** Humans, cyborgs and uploads coexist peacefully thanks to property abolition and guaranteed income.
- **Gatekeeper:** A superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. As a result, helper robots with slightly subhuman intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.

Superintelligence scenarios (Tegmark)

- **Protector god:** Essentially omniscient and omnipotent AI maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough that many humans even doubt the AI's existence.
- **Enslaved god:** A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.
- **Conquerors:** AI takes control, decides that humans are a threat/nuisance/waste of resources, and gets rid of us by a method that we don't even understand.
- **Descendants:** AIs replace humans, but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who is smarter than them, who learns from them and then accomplishes what they could only dream of—even if they can't live to see it all.

Superintelligence scenarios (Tegmark)

- **Zookeeper:** An omnipotent AI keeps some humans around, who feel treated like zoo animals and lament their fate.
- **1984:** Technological progress toward superintelligence is permanently curtailed not by an AI but by a human led Orwellian surveillance state where certain kinds of AI research are banned.
- **Reversion:** Technological progress toward superintelligence is prevented by reverting to a pre-technological society in the style of the Amish.
- **Self-destruction:** Superintelligence is never created because humanity drives itself extinct by other means (say nuclear and/or biotech mayhem fueled by climate crisis).

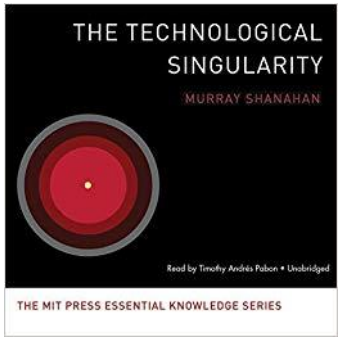
Scenarios summary

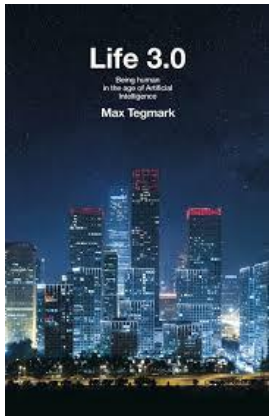
Scenario	Superintelligence exists?	Humans exist?	Humans in control?	Humans safe?	Humans happy?	Consciousness exists?
Libertarian utopia	Yes	Yes	No	No	Mixed	Yes
Benevolent dictator	Yes	Yes	No	Yes	Mixed	Yes
Egalitarian utopia	No	Yes	Yes?	Yes	Yes?	Yes
Gatekeeper	Yes	Yes	Partially	Potentially	Mixed	Yes
Protector god	Yes	Yes	Partially	Potentially	Mixed	Yes
Enslaved god	Yes	Yes	Yes	Potentially	Mixed	Yes
Conquerors	Yes	No	-	-	-	?
Descendants	Yes	No	-	-	-	?
Zookeeper	Yes	Yes	No	Yes	No	Yes
1984	No	Yes	Yes	Potentially	Mixed	Yes
Reversion	No	Yes	Yes	No	Mixed	Yes
Self-destruction	No	No	-	-	-	No

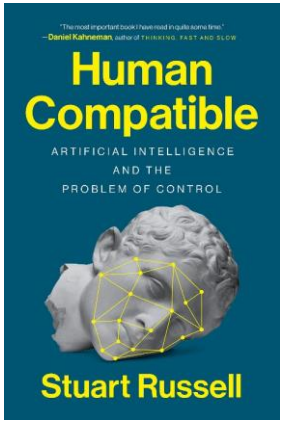
Cosmological perspective

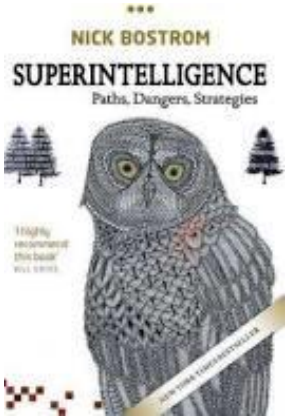
- preservation of consciousness in the universe
- avoiding energy death
- space colonization

Sources:


THE TECHNOLOGICAL SINGULARITY
MURRAY SHANAHAN
Read by Timothy Andrés Pabon • Unabridged
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES


Life 3.0
Being human in the age of Artificial Intelligence
Max Tegmark


Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies