UNIVERSITAT DE BARCELONA

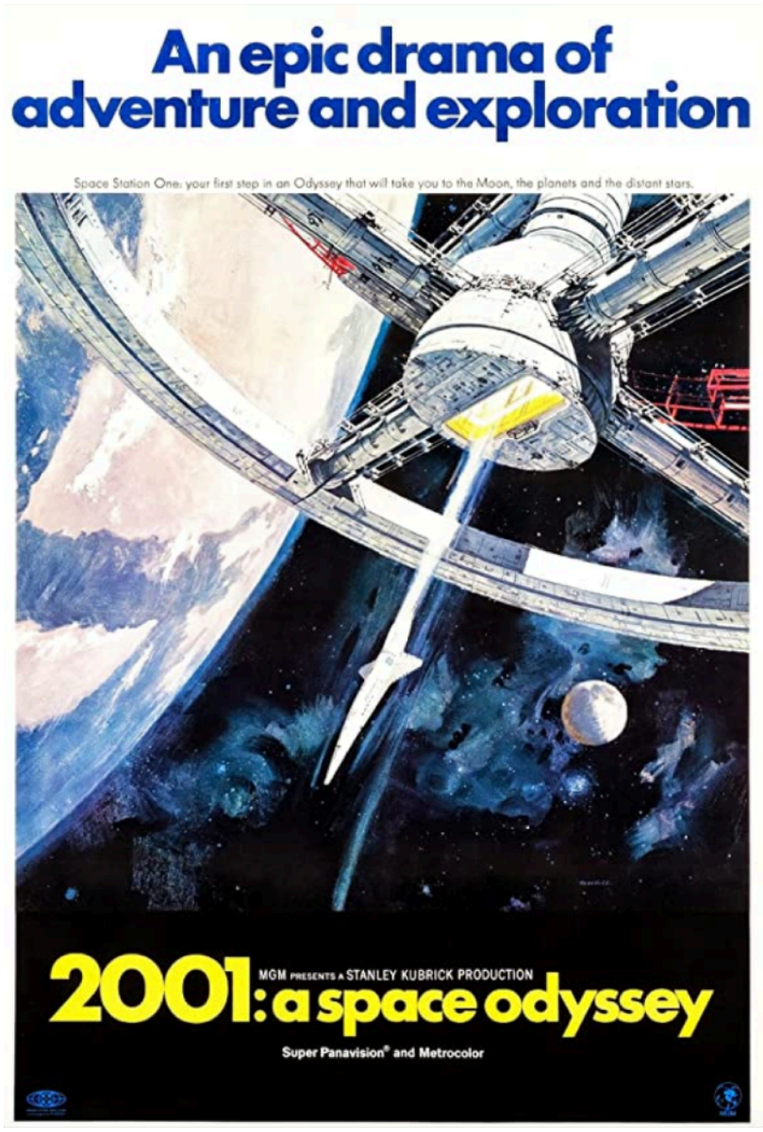
MSc in Fundamental Principles of Data Science

Ethical Data Science

AI and the alignment problem

Jordi Vitrià

2020-2021


An epic drama of adventure and exploration
Space Station One: your first step in an Odyssey that will take you to the Moon, the planets and the distant stars.
2001: a space odyssey
MGM PRESENTS A STANLEY KUBRICK PRODUCTION
Super Panavision® and Metrocolor


Dave: Open the pod bay doors, HAL.
HAL: I'm sorry, Dave. I'm afraid I can't do that.
Dave: What's the problem?
HAL: I think you know what the problem is just as well as I do.
Dave: What are you talking about, HAL?
HAL: This mission is too important for me to allow you to jeopardize it.
Dave: I don't know what you're talking about, HAL.
HAL: I know that you and Frank were planning to disconnect me. And I'm afraid that's something I cannot allow to happen.
Dave: Where the hell did you get that idea, HAL?
HAL: Dave, although you took very thorough precautions in the pod against my hearing you, I could see your lips move.
Dave: All right, HAL. I'll go in through the emergency airlock.
HAL: Without your space helmet, Dave, you're going to find that rather difficult.
Dave: HAL, I won't argue with you any more! Open the doors!
HAL: [almost sadly] Dave, this conversation can serve no purpose any more. Goodbye.

What if we do succeed?

(Credit: Stuart Russell)

Which is the bigest event in the future of humanity?

1.We all die (asteroid impact, climate change, pandemic,…)

2.We all live forever (medical solution to aging)

3.We invent faster-than-light travel and conquer the univers.

4.We are visited by a superior alien civilization.

5.We invent superintelligent AI.

What if we do succeed?

(Credit: Stuart Russell)

Which is the bigest event in the future of humanity?

1.We all die (asteroid impact, climate change, pandemic,…)

2.We all live forever (medical solution to aging)

3.We invent faster-than-light travel and conquer the univers.

4.We are visited by a superior alien civilization.

5.We invent superintelligent AI.

It would help us to avoid 1 and achieve 2 & 3.
It is analogous to 4 but much likely and we have some say…

What if we do succeed?

(Credit: Stuart Russell)

Which is the bigest event in the future of humanity?

1.We all die (asteroid impact, climate change, pandemic,…)

2.We all live forever (medical solution to aging)

3.We invent faster-than-light travel and conquer the univers.

4.We are visited by a superior alien civilization.

5.We invent superintelligent AI.

Must it be “superintelligence” or being “slightly better than humans” is sufficient?

What if we do succeed?

(Credit: Stuart Russell)

Are we about to be overcome by machines?

> No! In principle, **it is possible**, but there are some breakthroughs that have to happen before we have a superintelligent AI.

It is difficult to predict how far we are from this point…

In the 1930s Ernest Rutherford (1871–1937) repeatedly suggested, sometimes angrily, that the possibility of harnessing atomic energy was “moonshine.”

On September 12, 1933, Leo Szilard imagined a reasonable mechanism for releasing the vast quantities of energy known to be stored in atomic nuclei. As it turned out, his concept worked the first time it was tried on December 2, 1942.

What if we do succeed?

(Credit: Stuart Russell)

What is “intelligence”?

> The best definition we know is: “Humans are intelligent to the extend that our actions can be expected to achieve our objectives”.

It is the **property of an agent**: an autonomous entity which act upon an **environment** using **sensors** and **actuators** for achieving **goals**.

All those other characteristics of intelligence (learning, perception, planning, etc.) can be understood through their contributions to our ability to **act successfully**.

What if we do succeed?

(Credit: Stuart Russell)

What is AI?

> MACHINES are intelligent to the extend that THEIR actions can be expected to achieve THEIR objectives.
>
> But, (up to now) MACHINES do not have objectives by their own! We give them the objectives to achieve. We build **optimization machines, not intelligent machines**!

Should we built intelligent machines?

Misuses of AI

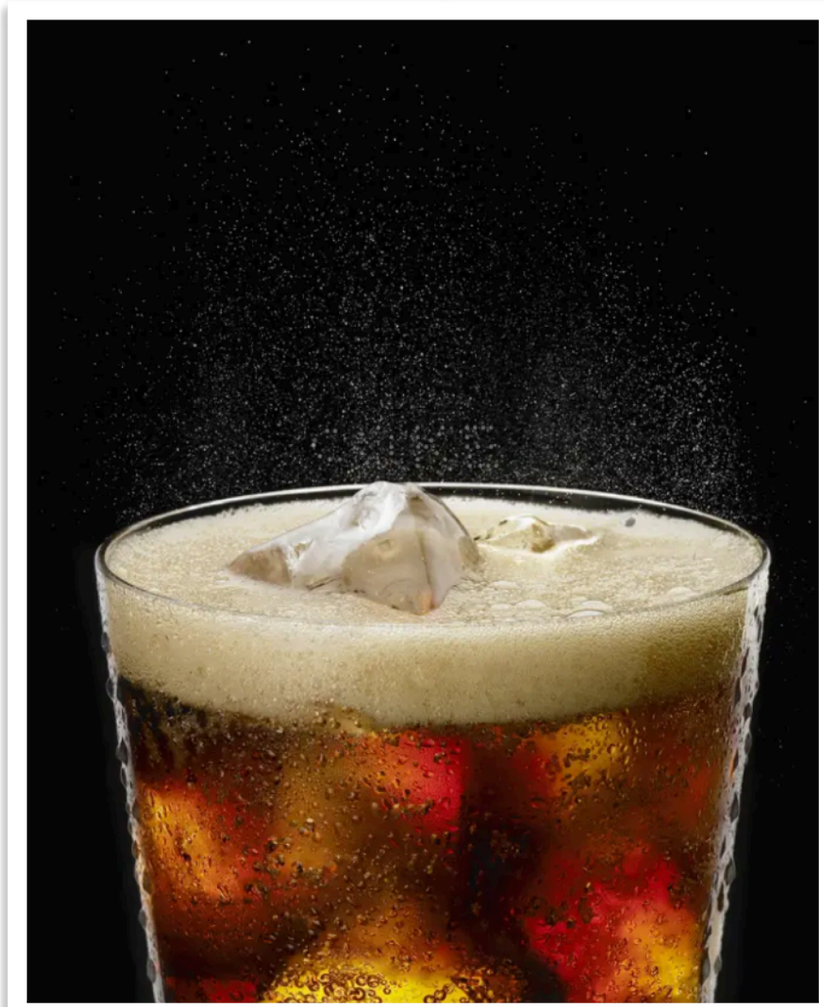
Surveillance, persuasion and control.

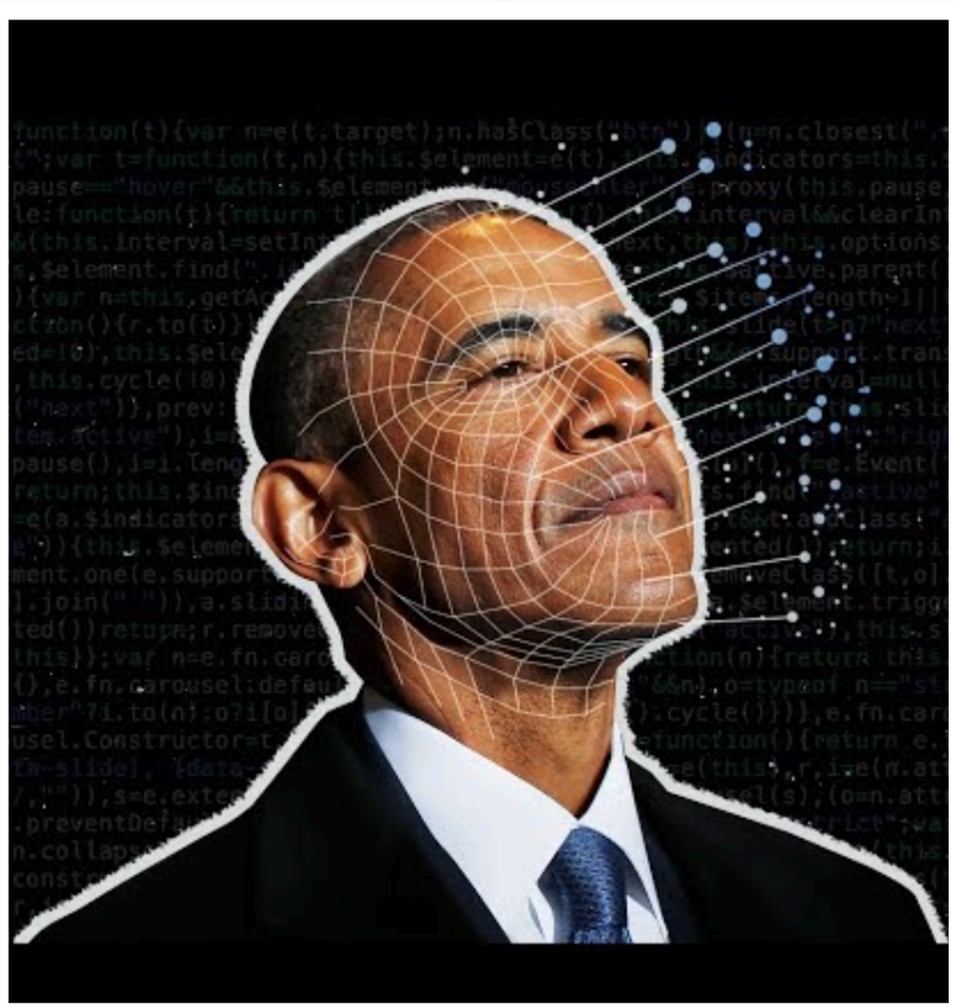
We already have surveillance capabilities in place. If used by a superintelligent AI, the Stasi will look like amateurs.

Misuses of AI

Surveillance, persuasion and control.

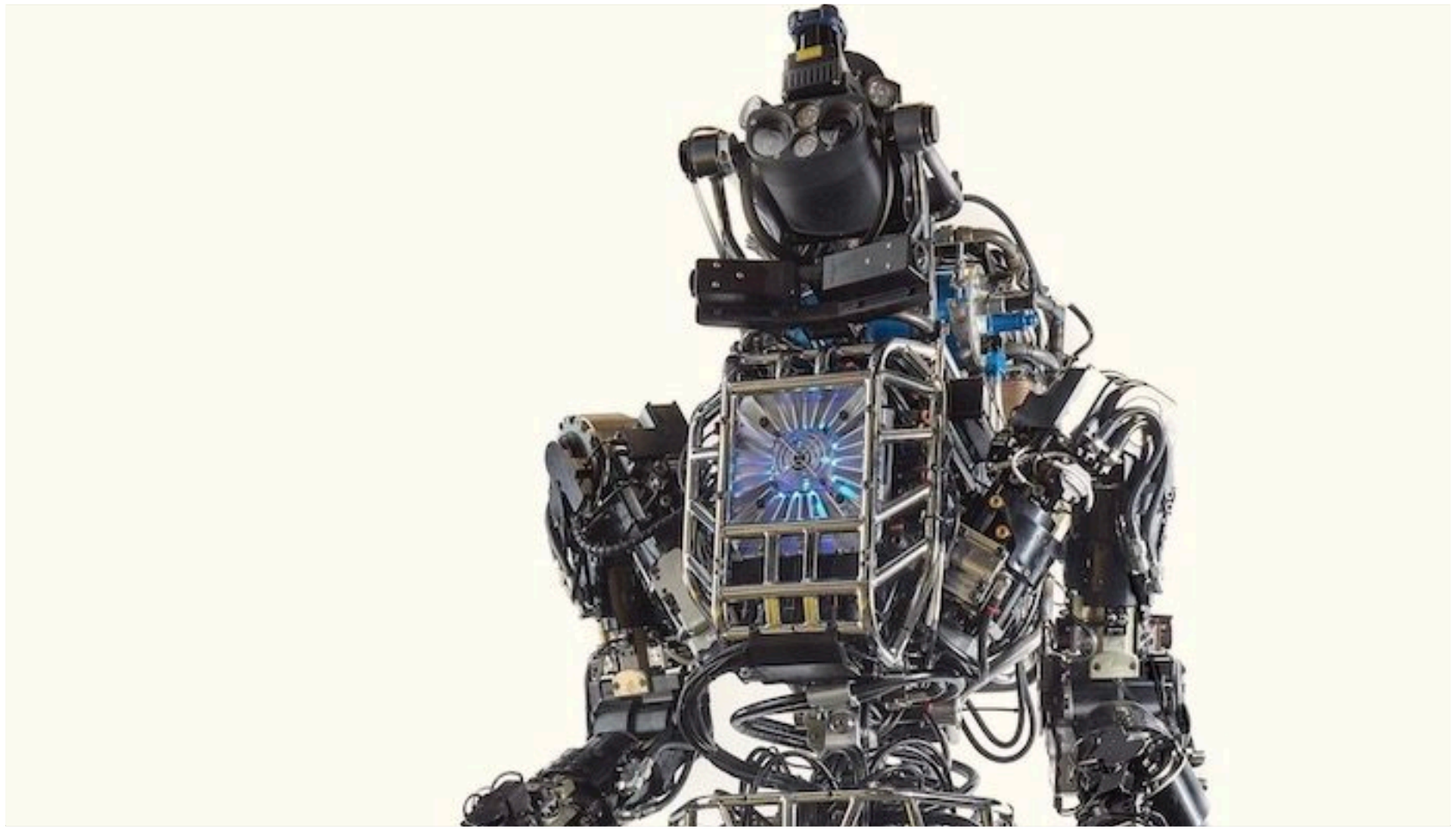
If non benevolent AI can get the data, the next step is to change our behavior to suit its objectives.





Misuses of AI

Lethal Autonomous Weapons



Misuses of AI

Eliminating work as we know it.

When a machine replaces one's physical labor, our can shell mental labor. When a machine replaces one`s metal labor, what does one left to sell?

"Data science is a very tiny lifeboat for a giant cruise ship"

Yong Ying-I

Misuses of AI

Usurping other human roles

European Parliament
2014-2019

TEXTS ADOPTED

P8_TA(2017)0051

Civil Law Rules on Robotics

European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))

f) creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently;

What if we do succeed?

(Credit: Stuart Russell)

What is AI?

> MACHINES are intelligent to the extend that THEIR actions can be expected to achieve THEIR objectives.
>
> But, (up to now) MACHINES do not have objectives by their own! We give them the objectives to achieve. We build **optimization machines, not intelligent machines**!

Should we built intelligent machines?

What if we do succeed?

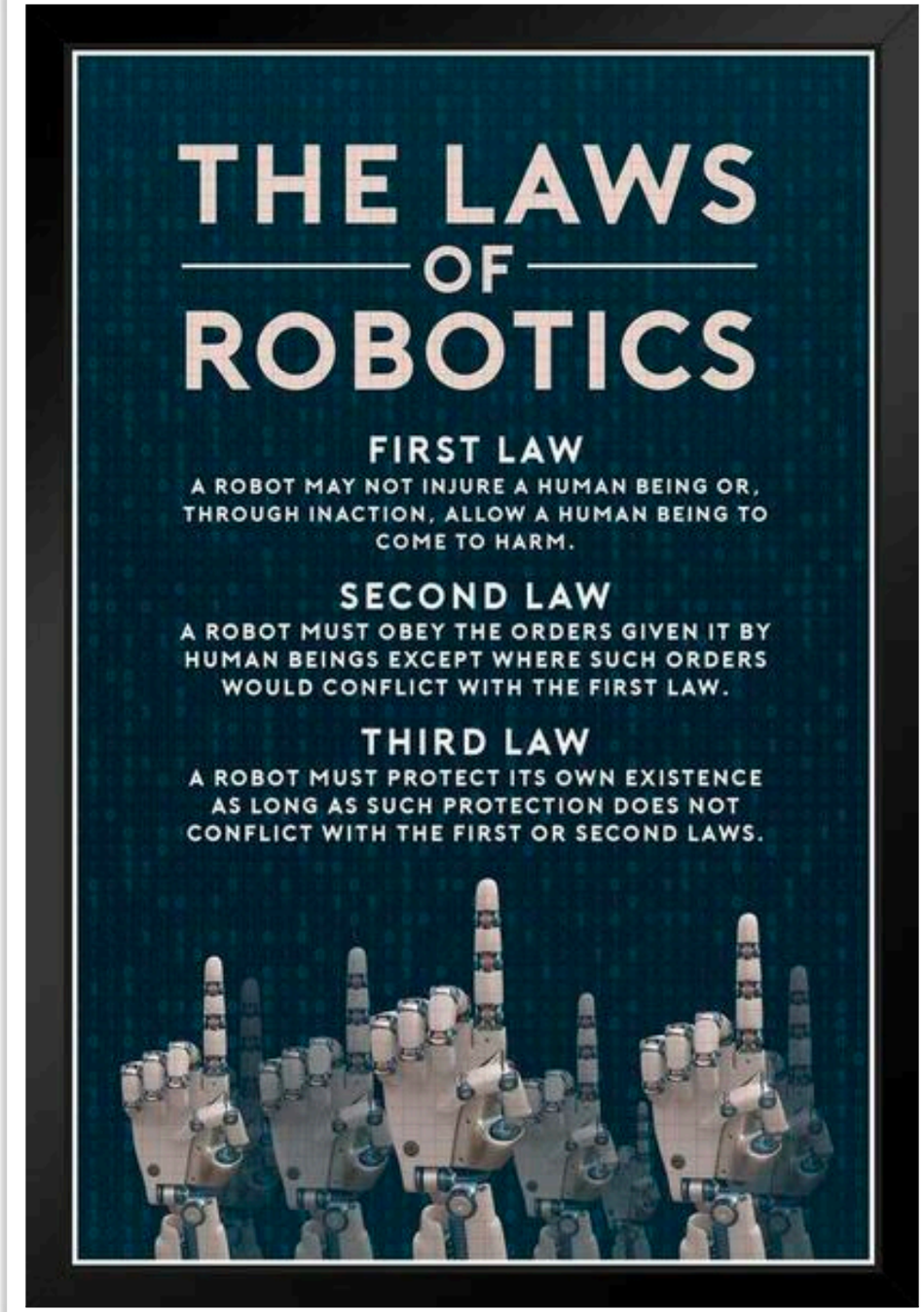
The answer can be “yes”, but only in the case we solve the **AI control problem**:

- How to build a superintelligent agent that will **aid** its creators, and **avoid inadvertently building** a superintelligence that will **harm** its creators.

“One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. 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.”

Stephen Hawking

What if we do succeed?



Isaac Asimov's **"Three Laws of Robotics"**

What if we do succeed?

Asimov's laws are organised around the **moral value** of **preventing harm to humans**, but they are not easy to interpret.

> For example, falling is an everyday hazard of life. An elderly person can rationally judge that living with the after-effects of a fall is better than a regime in which they people are heavily monitored and insulated from all danger. But a robot that deferred to this judgement by not interfering when its user wants to take a steep walk would be violating the first law!

This is a classical ethical paradox related to the deontological approach. Remember Kant and the inquiring murderer!

What if we do succeed?

Kant and the inquiring murderer.

What if we do succeed?

There is another alternative: instead of defining AI goals as a set of norms and values, set the goal of "**alignment**". This goal would ensure beneficial AI by default.

> Machines are beneficial to the extend their actions can be expected to achieve our objectives.

AI **alignment problem**: how to design and build AI models which capture our norms and values, understand what we mean or intend, and, above all, do what we want.

The "alignment" approach

When building intelligent machines, we should follow these principles:

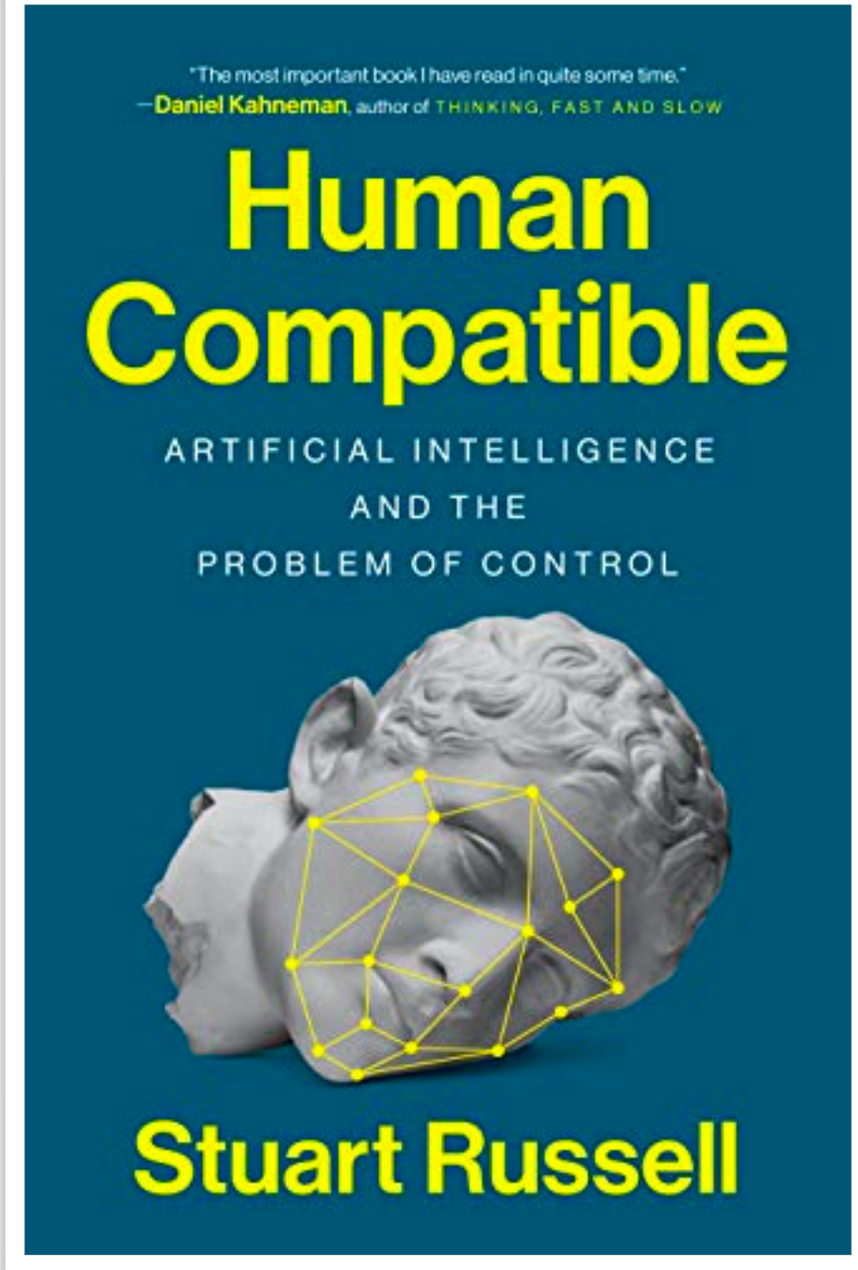
- ALTRUIST MACHINE. The machine's only objective is to maximize the realization of human preferences.

- HUMBLE MACHINE. The machine is initially uncertain about what those preferences are.

- LEARNING MACHINE. The ultimate source of information about human preferences is human behavior.

Research Question: **Can we prove theorems to the effect that a particular way of designing AI systems ensures that they will be beneficial to humans?**

The “alignment” approach



Complications

The human race is composed of nasty, envy-driven, irrational, inconsistent, unstable, computationally-limited, complex, evolving, heterogeneous entities.

Observations:

Designed machines should not behave like those they observe.

Governance of AI is more difficult than governance of nuclear power (because AI is based on “more democratic” technologies).

Complications

Enfeeblement and human autonomy.

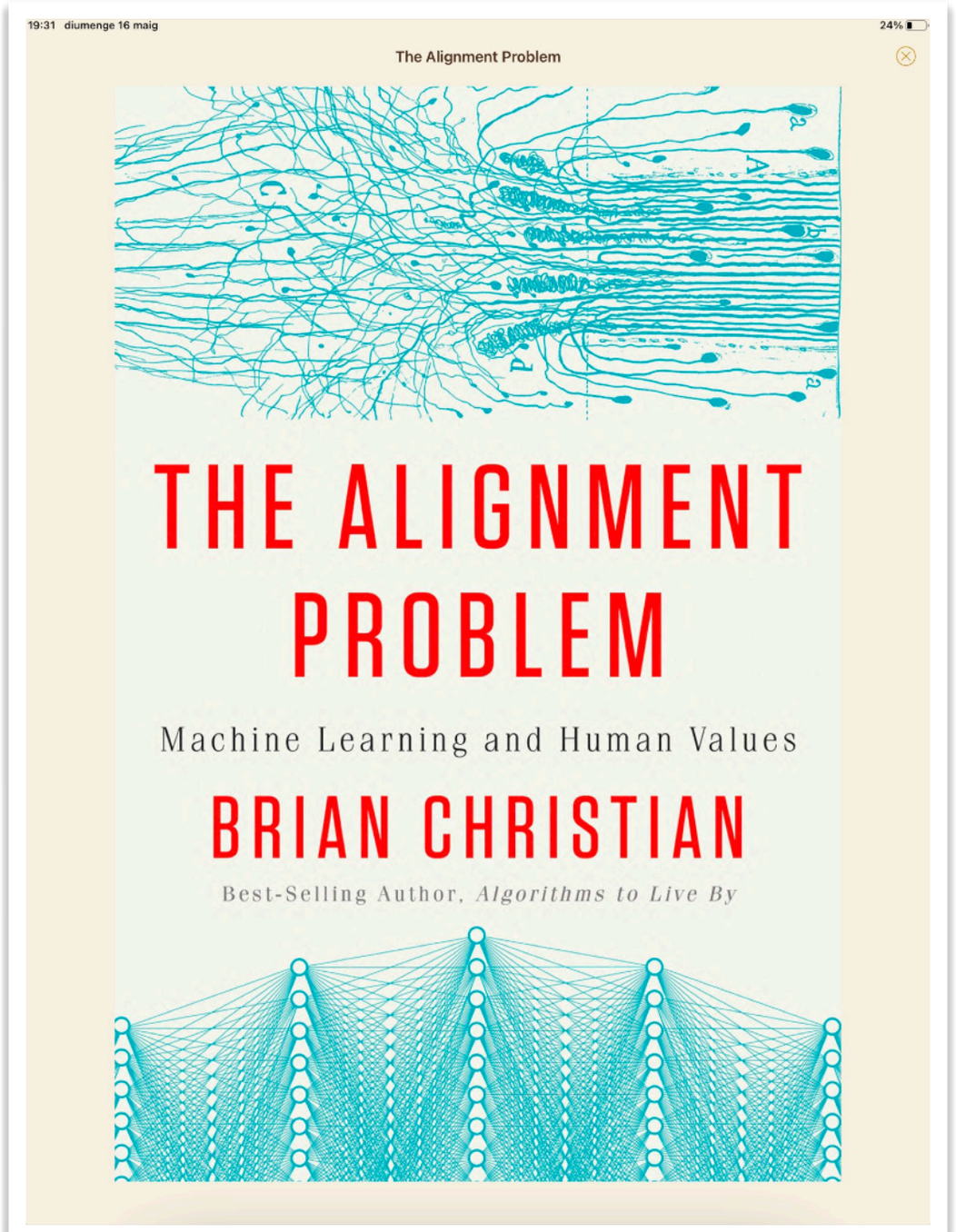
Humans become increasinly dependent on technology, but they understand less and less how it works.

Human autonomy is an individual's capacity for self-determination or self-governance. It describes a person's ability to make her or his own rules in life and to make decisions independently. Autonomy is a fundamental human value and an ethical principle.

As AI systems become more autonomous and supplant humans and human decision-making in increasing manners, there is the risk that we will lose the ability to make our own life rules, decisions or shape our lives, in cohort with other humans as traditionally has been the case.

For example, we may increasingly consult an AI medical advisor. We might choose to give up our ability and willingness to know and understand our own bodies and ailments. This example implies a profound transformation of who we are, what we do, and how we relate to ourselves.

Resources



https://youtu.be/CzoVn8LUaDs

ALIGNMENT NEWSLETTER

I edit and write content for the Alignment Newsletter, a weekly publication with recent content relevant to AI alignment with over 2300 subscribers. While it used to be more of an overview, it's now more like "things Rohin finds interesting", which excludes a bunch of work that a lot of other smart people are excited about.

It turns out that people don't notice things when they're part of a paragraph of text, so here is:

RECENT POSTS

Assignment

Edge

To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves.

CONVERSATION : SPECIAL EVENTS

Is Superintelligence Impossible?

On Possible Minds: Philosophy and AI with Daniel C. Dennett and David Chalmers

Moderated by John Brockman [4.10.19]

https://www.edge.org/conversation/john_brockman-is-superintelligence-impossible

Write a short essay about this video, focusing on agreements and disagreements between Chalmers and Dennett .

Ethical Data Science 2020-2021