Föreläsning på MVE420: Nya teknologier, global risk och mänskligheten framtid

AI-risk och AI-etik

27 april 2021

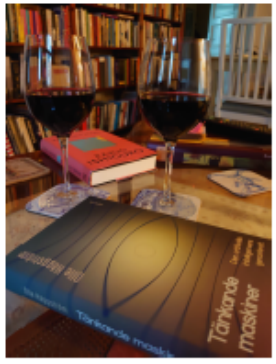
Olle Häggström

https://www.chalmers.se/en/Staff/Pages/olle-haggstrom.aspx

http://haggstrom.blogspot.com/


Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Tänkande
maskiner


Tänkande
maskiner


Tänkande
maskiner


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Ideals, Varieties,
and Algorithms
Tänkande
maskiner


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott


Olle Häggström
Tänkande
maskiner
Den artificiella
intelligensens
genombrott

Definitionsfrågor

Definitionsfrågor

- Många åsikter finns om vad som skall räkns som AI, och en heltäckande definition vore för mycket begärt, men de flesta förslag inbegriper **automatiserat beslutsfattande**.

Definitionsfrågor

- Många åsikter finns om vad som skall räkns som AI, och en heltäckande definition vore för mycket begärt, men de flesta förslag inbegriper **automatiserat beslutsfattande**.
- **snäv AI** vs **AGI (artificiell generell intelligens)**

Men kan AGI bli verklighet?

Men kan AGI bli verklighet?

Men kan AGI bli verklighet?

Steven Pinker: *As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent.*

Men kan AGI bli verklighet?

Steven Pinker: *As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent.*

Right?

Men kan AGI bli verklighet?

Steven Pinker: *As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent.*

Right?

Wrong!

A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy

Global Catastrophic Risk Institute Working Paper 17-1

99 Pages • Posted: 16 Nov 2017

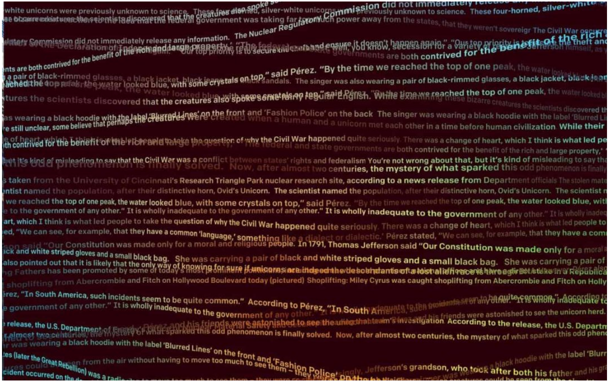
Seth Baum
Global Catastrophic Risk Institute

Date Written: November 12, 2017

Abstract

Artificial general intelligence (AGI) is AI that can reason across a wide range of domains. It has long been considered the "grand dream" or "holy grail" of AI. It also poses major issues of ethics, risk, and policy due to its potential to transform society: if AGI is built, it could either help solve the world's problems or cause major catastrophe, possibly even human extinction. This paper presents the first-ever survey of active AGI R&D projects in terms of ethics, risk, and policy. A thorough search identifies 45 projects of diverse sizes, nationalities, ethical goals, and other attributes. Most projects are either academic or corporate. The academic projects tend to express goals of advancing knowledge and

GPT-2 och GPT-3



In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Här ett replikskifte som ägde rum bara dagar efter att GPT-2 släpptes i februari 2019.

Här ett replikskifte som ägde rum bara dagar efter att GPT-2 släpptes i februari 2019.

Ej namngiven AI-forskare: I still think GPT-2 is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.

Här ett replikskifte som ägde rum bara dagar efter att GPT-2 släpptes i februari 2019.

Ej namngiven AI-forskare: I still think GPT-2 is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.

Scott Alexander: Yeah, well, your *mom* is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.

Kort historik

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.
- Alan Turing

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.
- Alan Turing
- Som AI-forskningens startskott brukar räknas sommarkonferensen i Dartmouth 1956.

"We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.
- Alan Turing
- Som AI-forskningens startskott brukar räknas sommarkonferensen i Dartmouth 1956.

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.
- Alan Turing
- Som AI-forskningens startskott brukar räknas sommarkonferensen i Dartmouth 1956.
- Därefter kom AI-forskningen att gå i vågor, med bl.a. två perioder av så kallad AI-vinter 1974-1980 och 1987-1993.

Kort historik

- AI har en lång förhistoria av visionära tänkare, bl.a. Mary Shelley (*Frankenstein*) och Ada Lovelace.
- Alan Turing
- Som AI-forskningens startskott brukar räknas sommarkonferensen i Dartmouth 1956.
- Därefter kom AI-forskningen att gå i vågor, med bl.a. två perioder av så kallad AI-vinter 1974-1980 och 1987-1993.
- Sedan början av 2010-talet befinner sig AI-utvecklingen i sin mest expansiva fas någonsin, mest tack vare det som kallas *deep learning*.

Av AI i framtiden kan vi bland annat annat vänta oss...

Av AI i framtiden kan vi bland annat annat vänta oss...

- självkörande bilar,

Av AI i framtiden kan vi bland annat annat vänta oss...

- självkörande bilar,
- kraftigt förbättrade beslutsverktyg i medicinsk diagnosticering och på många andra områden,

Av AI i framtiden kan vi bland annat annat vänta oss...

- självkörande bilar,
- kraftigt förbättrade beslutsverktyg i medicinsk diagnosticering och på många andra områden,
- en effektiviseringsvåg som revolutionerar bransch efter bransch,

Av AI i framtiden kan vi bland annat annat vänta oss...

- självkörande bilar,
- kraftigt förbättrade beslutsverktyg i medicinsk diagnosticering och på många andra områden,
- en effektiviseringsvåg som revolutionerar bransch efter bransch, och
- en betydande andel av den ekonomiska tillväxten det närmaste årtiondet eller så.

Men det finns också risker:

Men det finns också risker:

- automatiserad diskriminering

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text,

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...

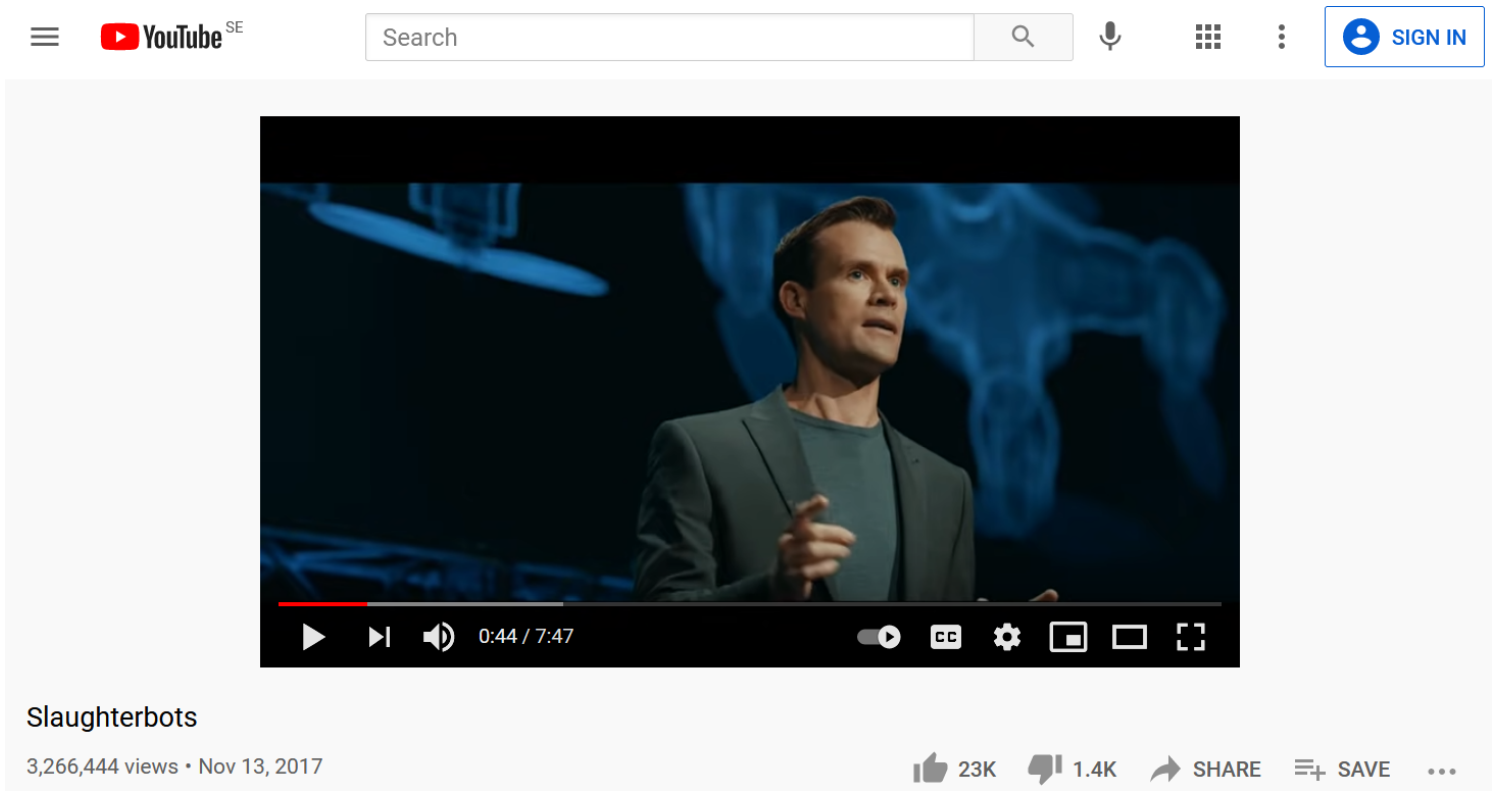
Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning
- autonoma vapen

https://www.youtube.com/watch?v=9CO6M2HsoIA




#Futureofeu

Vid mötet i Bryssel framhöll Pinker att det skulle krävas en galning för att konstruera något så hemskt som *“a swarm of robots designed to attack individual people based on facial recognition”*,

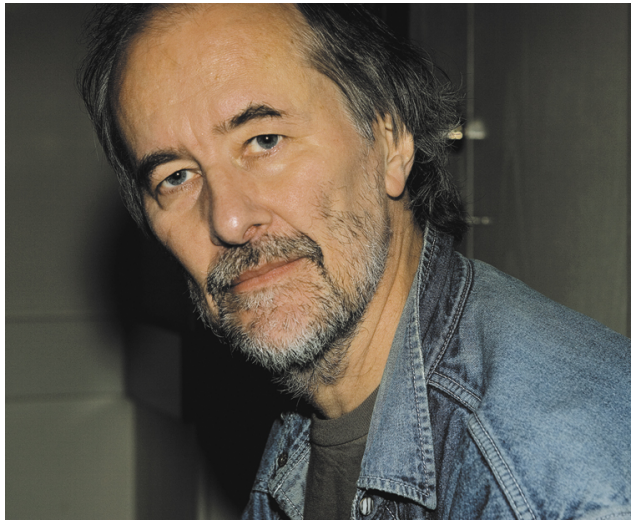
Vid mötet i Bryssel framhöll Pinker att det skulle krävas en galning för att konstruera något så hemskt som *“a swarm of robots designed to attack individual people based on facial recognition”*, och att det i dag inte längre finns utrymme för sådana enstaka galningar eftersom alla betydande ingenjörsprojekt numera genomförs i stora samarbeten.

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning
- autonoma vapen

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning
- autonoma vapen
- automatisering, arbetsmarknad och ekonomisk ojämlikhet



Maskinerna är våra vänner
utan dom inget paradis
maskinerna är våra vänner
naturligtvis

Tryck på en knapp
och lampan lyser
tryck på en knapp
när du fryser
tryck på en knapp

Tryck på en knapp
kläderna tvättas
tryck på en knapp
munnarna mättas
tryck på en knapp

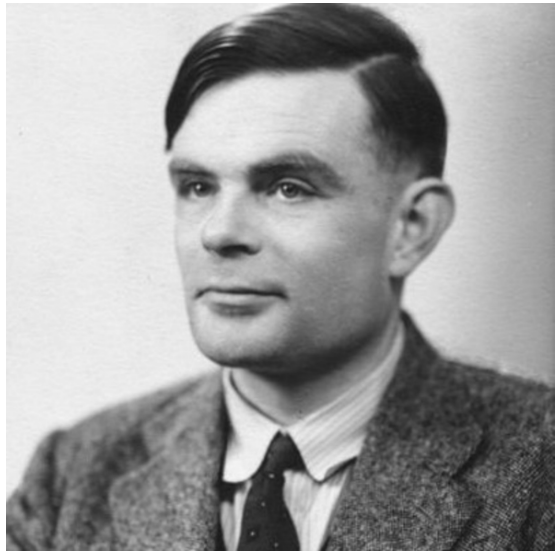
Du ska inte längre slita
i ditt anletes svett
du ska ägna dig åt vetenskap och konst
har jag fel eller rätt?

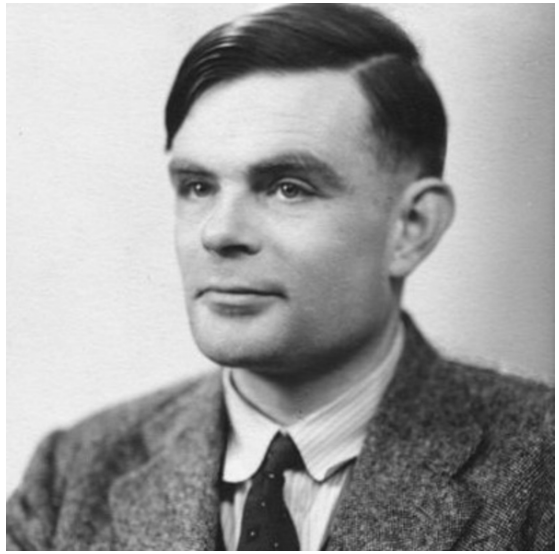
Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning
- autonoma vapen
- automatisering, arbetsmarknad och ekonomisk ojämlikhet

Men det finns också risker:

- automatiserad diskriminering
- manipulation, av bilder, videos, text, människor...
- privatliv vs övervakning
- autonoma vapen
- automatisering, arbetsmarknad och ekonomisk ojämlikhet
- AGI-katastrof





Alan Turing, 1951: "My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. [...] Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. [...] It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control."

Bitcoin consumes 'more electricity than Argentina'

By Cristina Criddle
Technology reporter

10 February

The University of Cambridge's online tool estimates Bitcoin's annual electricity consumption is more than The Netherlands, United Arab Emirates and Argentina

Bitcoin uses more electricity annually than the whole of Argentina, analysis by Cambridge University suggests.

"Mining" for the cryptocurrency is power-hungry, involving heavy computer calculations to verify transactions.

Cambridge researchers say it consumes around 121.36 terawatt-hours (TWh) a year - and is unlikely to fall unless the value of the currency slumps.

Critics say electric-car firm **Tesla's decision to invest heavily in Bitcoin** undermines its environmental image.

The currency's value hit a record $48,000 (£34,830) this week, following

”Sunt förnuft”

”Sunt förnuft”

AI Camera Ruins Football Game By Mistaking Referee's Bald Head For Ball

Many complained that they missed their team's goals because the camera "kept thinking the Lino bald head was the ball."

Offbeat | Written by Sanya Jain | Updated: November 02, 2020 9:56 am IST

An AI camera kept mistaking the referee's head for the football during a match.

An AI camera at a football game in Scotland kept mistaking the referee's bald head for the ball, ruining the game for fans. According to the Daily Mail, Inverness

”Sunt förnuft”

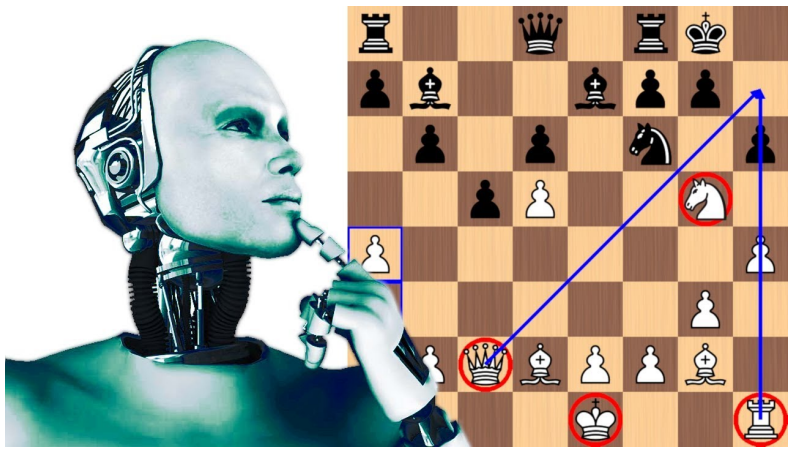
AI Camera Ruins Football Game By Mistaking Referee's Bald Head For Ball

Many complained that they missed their team's goals because the camera "kept thinking the Lino bald head was the ball."

Offbeat | Written by Sanya Jain | Updated: November 02, 2020 9:56 am IST

An AI camera kept mistaking the referee's head for the football during a match.

An AI camera at a football game in Scotland kept mistaking the referee's bald head for the ball, ruining the game for fans. According to the Daily Mail, Inverness

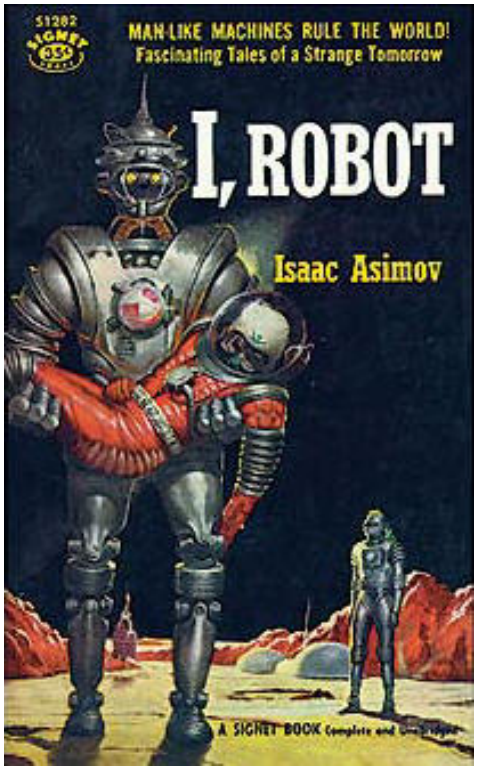


Om vi tar Alan Turings ord om *"at some stage therefore we should have to expect the machines to take control"* på allvar, så är ett centralt framtidsproblem det som kommit att kallas **AI Alignment**:

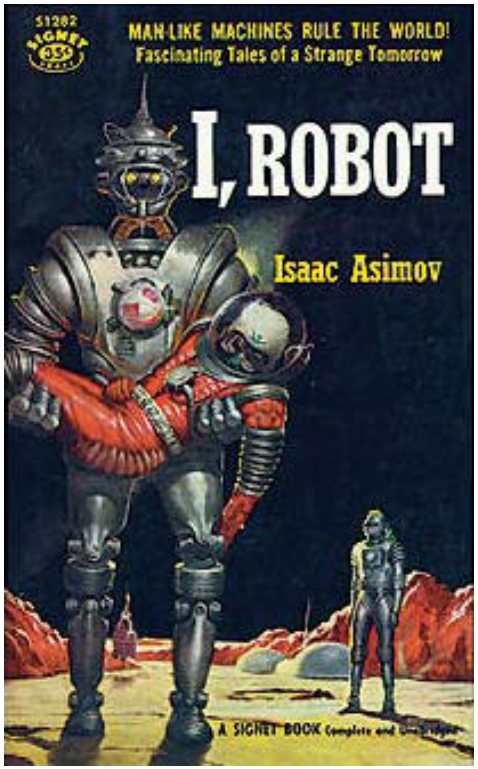
Om vi tar Alan Turings ord om *"at some stage therefore we should have to expect the machines to take control"* på allvar, så är ett centralt framtidsproblem det som kommit att kallas **AI Alignment**:

Hur ser vi till att en superintelligent AGI har mål och drivkrafter som är i linje med mänskliga värderingar och mänsklig välgång?

Den mest kända ansatsen till AI Alignment: Asimovs tre robotlagar

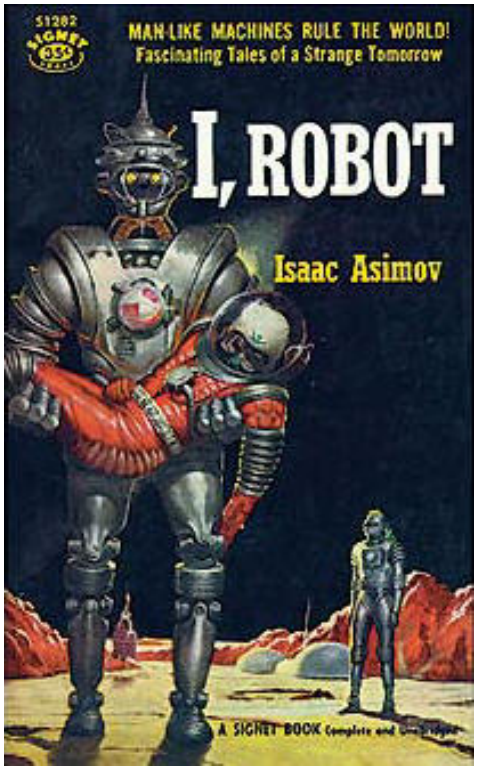


Den mest kända ansatsen till AI Alignment: Asimovs tre robotlagar



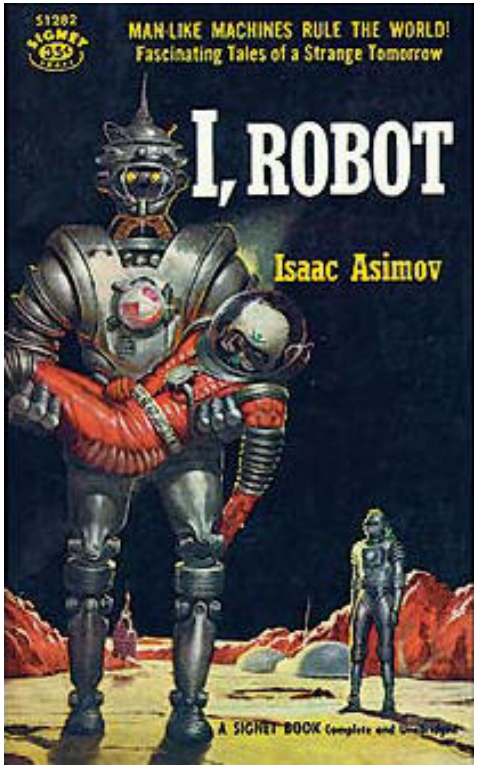
1. En robot får aldrig skada en människa eller, genom att inte ingripa, tillåta att en människa kommer till skada.

Den mest kända ansatsen till AI Alignment: Asimovs tre robotlagar



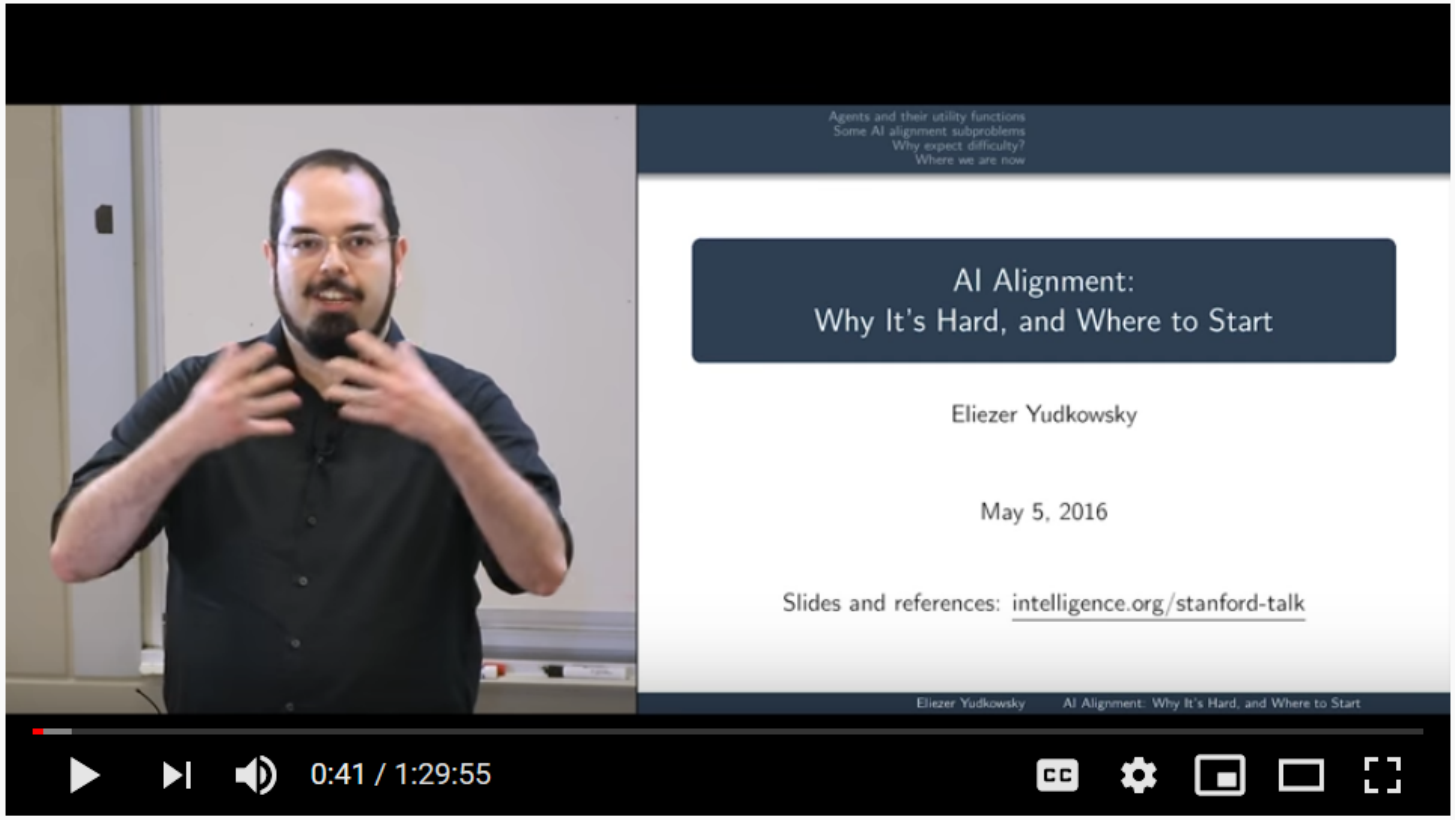
1. En robot får aldrig skada en människa eller, genom att inte ingripa, tillåta att en människa kommer till skada.
2. En robot måste lyda order från en människa, förutom om sådana order kommer i konflikt med första lagen.

Den mest kända ansatsen till AI Alignment: Asimovs tre robotlagar



1. En robot får aldrig skada en människa eller, genom att inte ingripa, tillåta att en människa kommer till skada.
2. En robot måste lyda order från en människa, förutom om sådana order kommer i konflikt med första lagen.
3. En robot måste skydda sin egen existens, såvida detta inte kommer i konflikt med första eller andra lagen.

Men en prioriterad lista av lagar fungerar inte, och om vi påminner oss beslutsteorin jag föreläste om härom veckan så minns vi den axiomatiskt uppbyggda teori från 1900-talet som utmynnar i att en rationell agent behöver maximera förväntad nytta, för någon nyttofunktion U.



Eliezer Yudkowsky – AI Alignment: Why It's Hard, and Where to Start

Ett första förslag till nyttofunktion U:

$$U = \begin{cases} 1 & \text{om kitteln är full med vatten} \\ 0 & \text{annars} \end{cases}$$

Nästa förslag: låt $U = U_1 + U_2$, där

$$U_1 = \begin{cases} 1 & \text{om kitteln är full med vatten} \\ 0 & \text{annars} \end{cases}$$

och

$$U_2 = \begin{cases} -1000 & \text{om lokalen översvämmas} \\ 0 & \text{annars} \end{cases}$$

Komplikationerna hopar sig. Vad sägs om $U = U_1 + U_2 + U_3 + U_4$, där

$$U_1 = \begin{cases} 1 & \text{om kitteln är full med vatten} \\ 0 & \text{annars,} \end{cases}$$

$$U_2 = \begin{cases} -1000 & \text{om lokalen översvämmas} \\ 0 & \text{annars,} \end{cases}$$

$$U_3 = \begin{cases} -100\,000 & \text{om någon dödas av svängande hink} \\ 0 & \text{annars,} \end{cases}$$

$$U_4 = \begin{cases} 0.1 & \text{om hinkarna kånkas elegant} \\ 0 & \text{annars.} \end{cases}$$

Att modifiera U genom ta hänsyn till det ena specialfallet efter det andra är ett projekt som aldrig kan fullbordas. En mer allmän modifiering behövs.

Att modifiera U genom ta hänsyn till det ena specialfallet efter det andra är ett projekt som aldrig kan fullbordas. En mer allmän modifiering behövs.

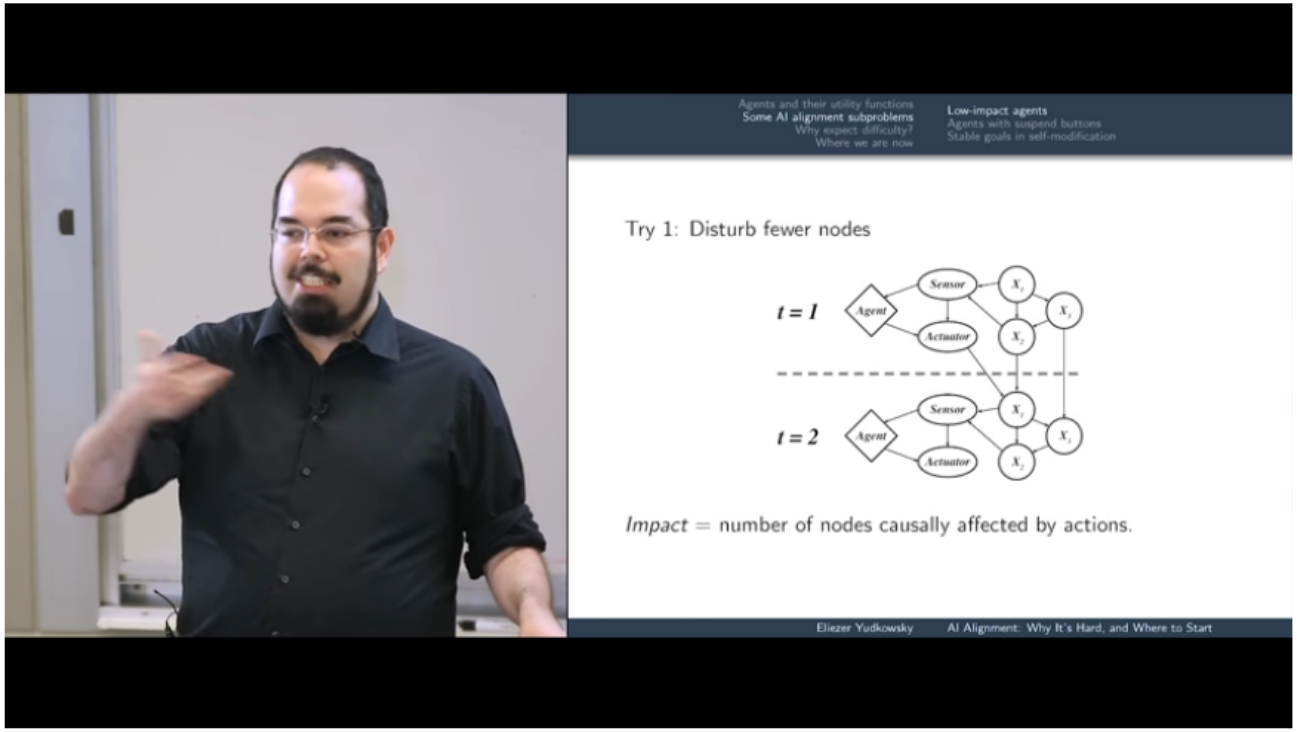
Vad sägs om att låta U bestraffa bieffekter vilka dessa än är?

$$U(x) = \begin{cases} 1 - \mathit{Impact}(x) & \text{om kitteln är full med vatten} \\ 0 - \mathit{Impact}(x) & \text{annars,} \end{cases}$$

där $\mathit{Impact}(x)$ ger ett mått på hur mycket kvastens handling påverkar världen utöver huruvida kitteln är fylld eller inte.

Men vi behöver ge $Impact(x)$ en exakt definition.

Men vi behöver ge *Impact*(x) en exakt definition.



The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

Abstract

It is clear that one of the primary tools we can use to mitigate the potential risk from a misbehaving AI system is the ability to turn the system off. As the capabilities of AI systems improve, it is important to ensure that such systems do not adopt subgoals that prevent a human from switching them off. This is a challenge because many formulations of rational agents create strong incentives for self-preservation. This is not caused by a built-in instinct, but because a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead. Our goal is to study the incentives an agent has to allow itself to be switched off. We analyze a simple game between a human H and a robot R, where H can press R's off switch but R can disable the off switch. A traditional agent takes its reward function for granted: we show that such agents have an incentive to disable the off switch, except in the special case where H is perfectly rational. Our key insight is that for R to want to preserve its off switch, it needs to be uncertain about the utility associated with the outcome, and to treat H's actions as important observations about that utility. (R also has no incentive to switch *itself* off in this setting.) We conclude that giving machines an appropriate level of uncertainty about their objectives leads to safer designs, and we argue that this setting is a useful generalization of the classical AI paradigm of ra-

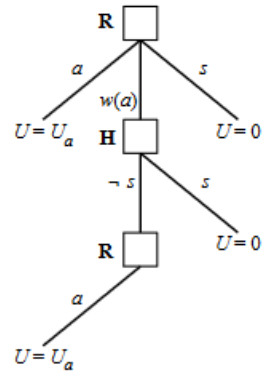
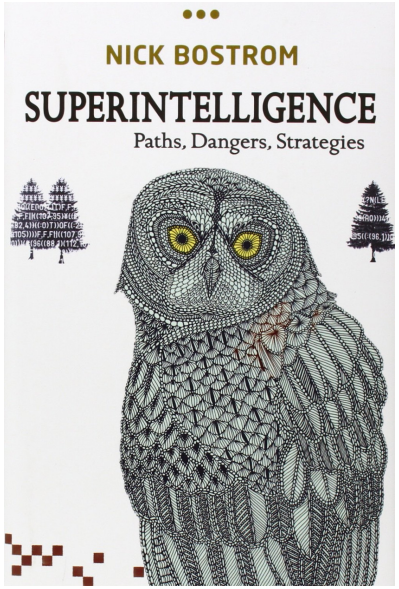


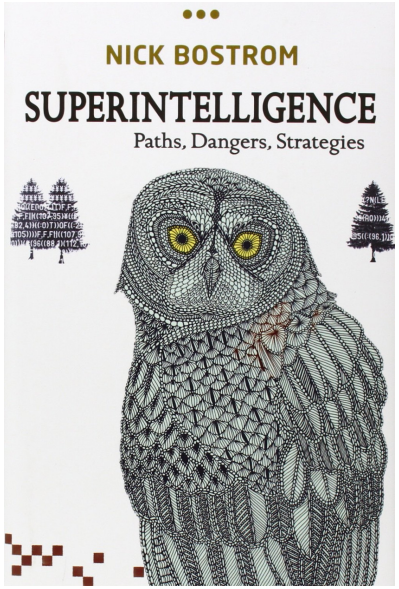
Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

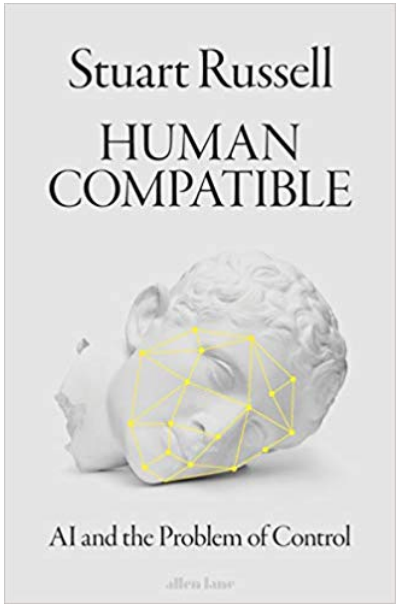
strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger is certainly something which can give us anxiety [Turing, 1951]."

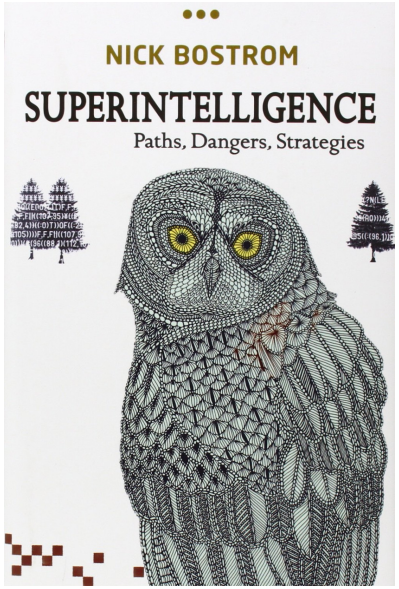
There has been recent debate about the validity of this concern, so far, largely relying on informal arguments. One important question is how difficult it is to implement Turing's idea of 'turning off the power at strategic moments', i.e., switching a misbehaving agent off[1]. For example, some have argued that there is no reason for an AI to resist being switched off unless it is explicitly programmed with a self-preservation incentive [Del Prado, 2015], [Omohundro

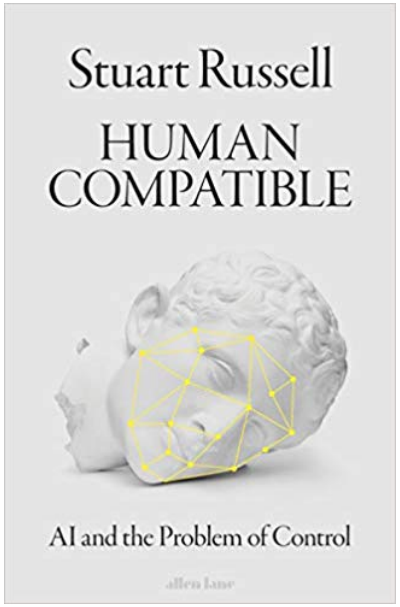

Human-compatible AI
1. The robot's only objective is to maximize
the rea tion of human values
2. The ot is initially uncertain about what
thos re
3. avior provides information
ab values

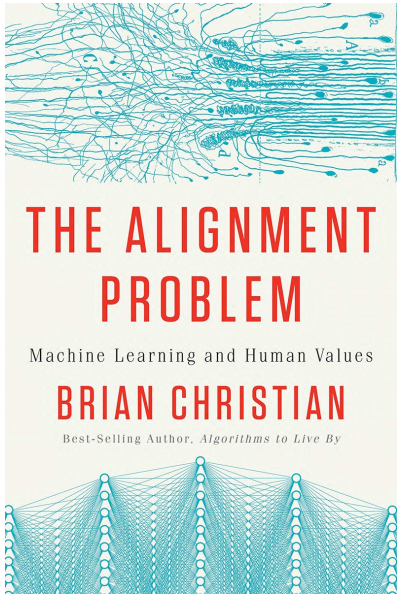

NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

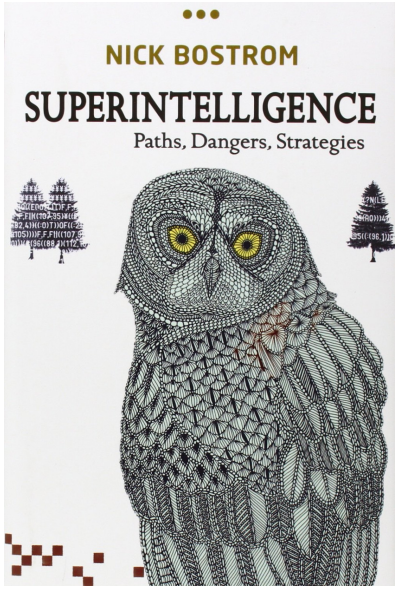

NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

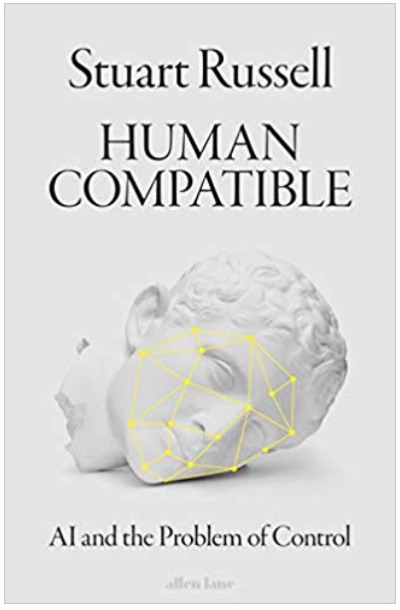

Stuart Russell
HUMAN COMPATIBLE
AI and the Problem of Control

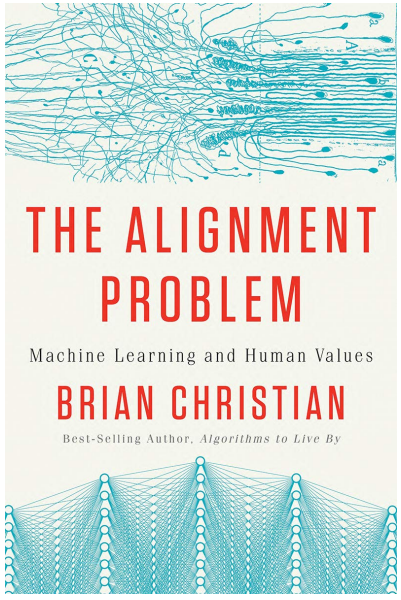

NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

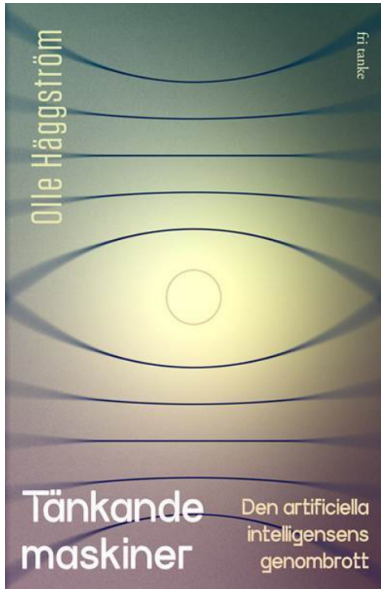

Stuart Russell
HUMAN COMPATIBLE
AI and the Problem of Control


THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, Algorithms to Live By


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies


Stuart Russell
HUMAN COMPATIBLE
AI and the Problem of Control


THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, Algorithms to Live By


Olle Häggström
fri tanke
Tänkande maskiner
Den artificiella intelligensens genombrott