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

Kursintro och input till era val av projektämnen

23 mars 2021

Olle Häggström

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

http://haggstrom.blogspot.com/

Ämnet för denna kurs är nya och framtida teknologier som...

Ämnet för denna kurs är nya och framtida teknologier som...

- kan innebära revolutionerande framsteg, ekonomiskt och i termer av välbefinnande, hälsa, längre liv, etc,

Ämnet för denna kurs är nya och framtida teknologier som...

- kan innebära revolutionerande framsteg, ekonomiskt och i termer av välbefinnande, hälsa, längre liv, etc,
- men som också för med sig risker av liknande dignitet, och i vissa fall till och med mänsklighetens undergång.

Ämnet för denna kurs är nya och framtida teknologier som...

- kan innebära revolutionerande framsteg, ekonomiskt och i termer av välbefinnande, hälsa, längre liv, etc,
- men som också för med sig risker av liknande dignitet, och i vissa fall till och med mänsklighetens undergång.

Hur kan vi styra framtiden i sådan riktning att vi drar nytta av teknologierna samtidigt som de största riskerna undviks?

Sådana frågor är **starkt tvärvetenskapliga**.

Sådana frågor är **starkt tvärvetenskapliga**. De för också långt bortom de korta perspektiv där vi vet med bestämdhet vad vi kan vänta oss och vad som går att göra.

Sådana frågor är **starkt tvärvetenskapliga**. De för också långt bortom de korta perspektiv där vi vet med bestämdhet vad vi kan vänta oss och vad som går att göra. Det ställer stora krav på oss att **spekulera med gott omdöme**,

Sådana frågor är **starkt tvärvetenskapliga**. De för också långt bortom de korta perspektiv där vi vet med bestämdhet vad vi kan vänta oss och vad som går att göra. Det ställer stora krav på oss att **spekulera med gott omdöme**, vilket bl.a. innebär...

- en väl avvägd grad av ödmjukhet, och en förmåga att skilja mellan spekulation och prediktion,

Sådana frågor är **starkt tvärvetenskapliga**. De för också långt bortom de korta perspektiv där vi vet med bestämdhet vad vi kan vänta oss och vad som går att göra. Det ställer stora krav på oss att **spekulera med gott omdöme**, vilket bl.a. innebär...

- en väl avvägd grad av ödmjukhet, och en förmåga att skilja mellan spekulation och prediktion, och
- förmågan att genom studium av argument för och emot olika spekulativa scenarier kunna nyansera, och skilja de välgrundade från de ogrundade, samt de rimliga från de orimliga.

Till de teknikområden vi har anledning att studera hör...

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,
- syntetisk biologi,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,
- syntetisk biologi,
- massövervakning,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,
- syntetisk biologi,
- massövervakning,
- geoengineeringmetoder för att manipulera jordens klimat,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,
- syntetisk biologi,
- massövervakning,
- geoengineeringmetoder för att manipulera jordens klimat,
- rymdkolonisering,

Till de teknikområden vi har anledning att studera hör...

- artificiell intelligens,
- "förbättrandet" av människokroppen, människans kognitiva kapacitet och den mänskliga naturen, på t.ex. farmakologisk eller genetisk väg, eller via elektronik inbyggd innanför skallbenet,
- radikal nanoteknologi,
- syntetisk biologi,
- massövervakning,
- geoengineeringmetoder för att manipulera jordens klimat,
- rymdkolonisering,
- ...

Kursens två huvudkomponenter är...

Kursens två huvudkomponenter är...

- **föreläsningarna**, till stor del inriktade på att ge teoretiska redskap och ramverk för att bedriva detta slags futurologi,

Kursens två huvudkomponenter är...

- **föreläsningarna**, till stor del inriktade på att ge teoretiska redskap och ramverk för att bedriva detta slags futurologi,
- **projektet**, där studenterna två och två väljer ut ett ämne (t.ex. något visst teknikområde, någon global utmaning eller någon katastrofrisk) att behandla.

Examinationen består av

Examinationen består av

- **projektet**, som redovisas
  - **skriftligt**, och
  - **muntligt**.

Examinationen består av

- **projektet**, som redovisas
  - **skriftligt**, och
  - **muntligt**.
- en individuellt författad **essä** om innehållet i föreläsningarna.

Kursen har ett antal obligatoriska **deadlines**.

Kursen har ett antal obligatoriska **deadlines**.

Den första är **tisdagen den 30 mars kl 23.59**, då varje grupp skall ha lämnat in ett förslag om ämne för projekt, inklusive tankar om vilken litteratur som kan vara lämplig att stödja projektet på.

Kursen har ett antal obligatoriska **deadlines**.

Den första är **tisdagen den 30 mars kl 23.59**, då varje grupp skall ha lämnat in ett förslag om ämne för projekt, inklusive tankar om vilken litteratur som kan vara lämplig att stödja projektet på.

(Dessförinnan skall ni ha lyft preliminära tankar med detta muntligt, vid undervisningstillfället **på fredag den 26 mars**.)

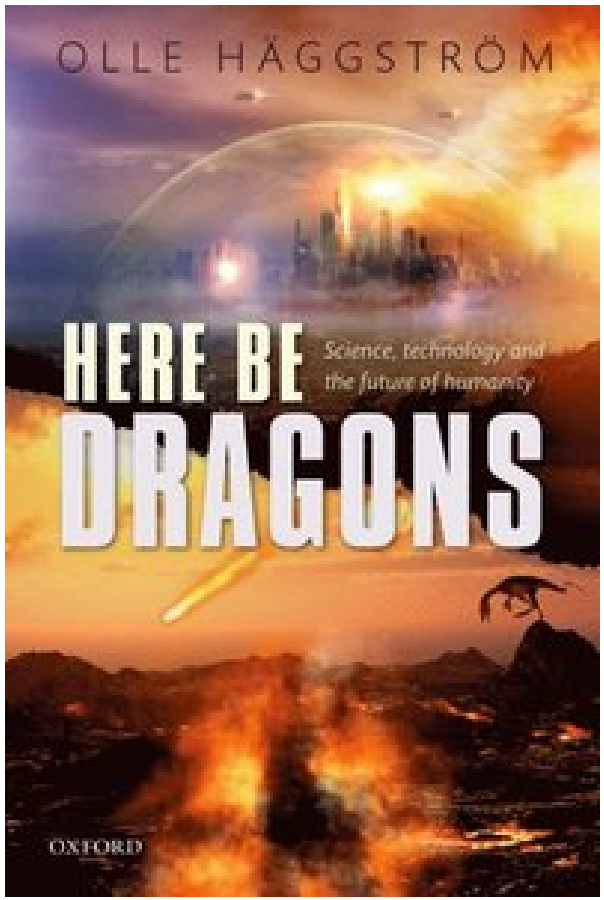
Kursen har ett antal obligatoriska **deadlines**.

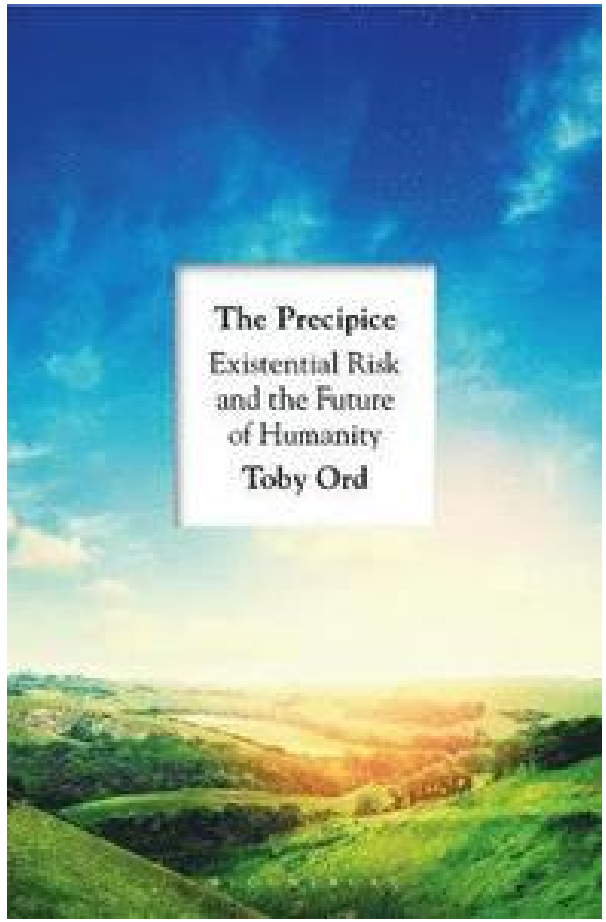
Den första är **tisdagen den 30 mars kl 23.59**, då varje grupp skall ha lämnat in ett förslag om ämne för projekt, inklusive tankar om vilken litteratur som kan vara lämplig att stödja projektet på.

(Dessförinnan skall ni ha lyft preliminära tankar med detta muntligt, vid undervisningstillfället **på fredag den 26 mars**.)

För övriga deadlines, se Canvas och Kurs-PM.

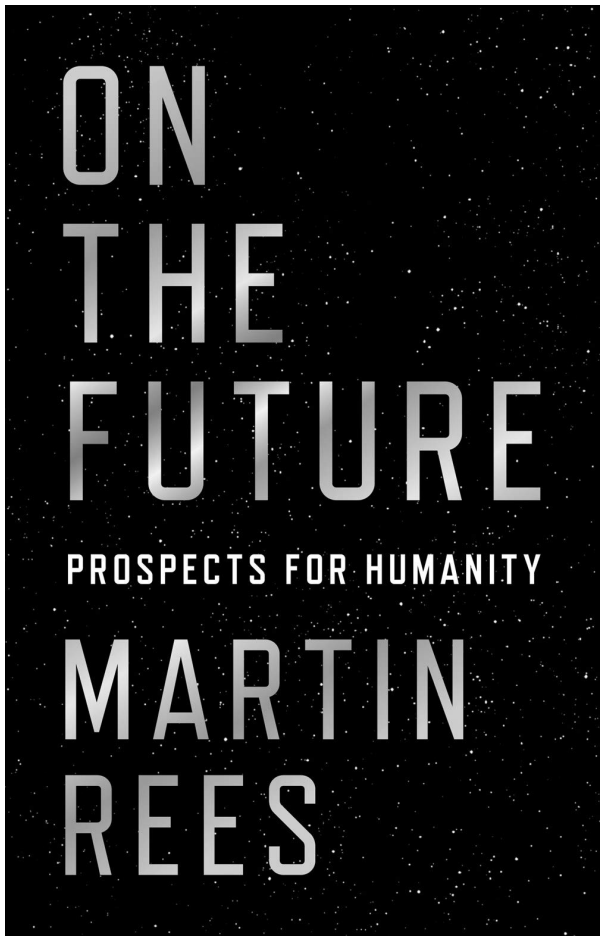
Någon obligatorisk kurslitteratur har vi inte, men en lämplig plats startpunkt för den som söker inspiration för projektämne kan vara min bok **Here Be Dragons: Science, Technology and the Future of Humanity** (2016) eller Toby Ords **The Precipice: Existential Risk and the Future of Humanity** (2020).



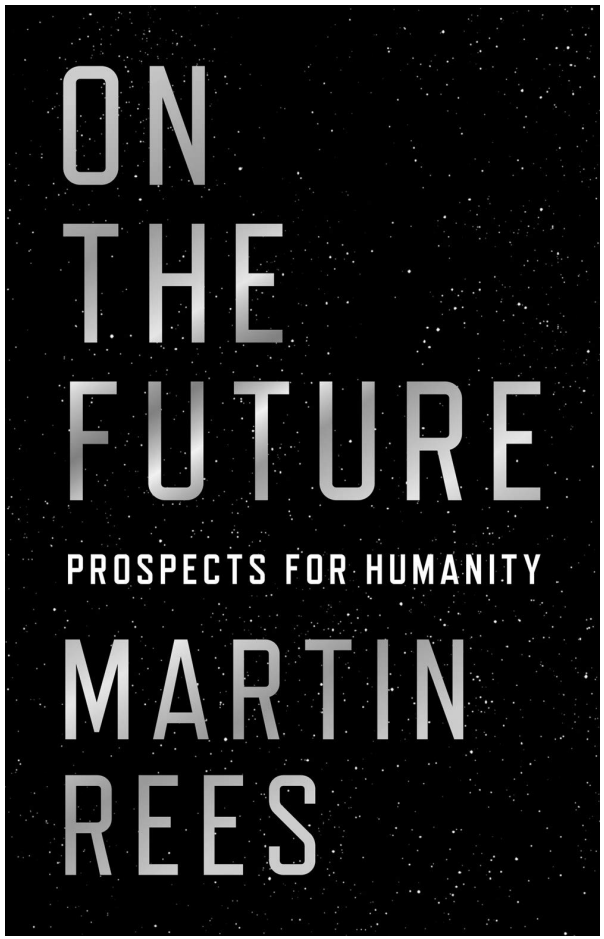


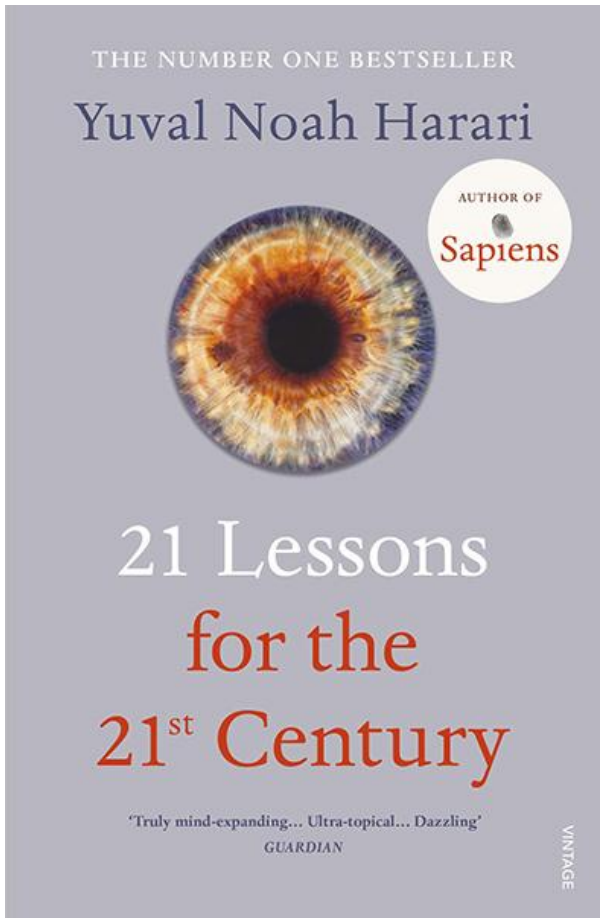
Två andra böcker som tar breda grepp om framtiden:

Två andra böcker som tar breda grepp om framtiden:



Två andra böcker som tar breda grepp om framtiden:





En av ambitionerna med kursen är att ge förmåga att välgrundat ta ställning till radikala teknikoptimistiska visioner liknande dem som uttrycks i citaten på de följande fyra bilderna.

The first ultraintelligent machine is the last invention that man need ever make. **(Jack Good, 1965)**

We have the means right now to live long enough to live forever. Existing knowledge can be aggressively applied to dramatically slow down aging processes so we can still be in vital health when the more radical life-extending therapies from biotechnology and nanotechnology become available. But most baby boomers won't make it because they are unaware of the accelerating aging process in their bodies and the opportunity to intervene. **(Ray Kurzweil, 2005)**

Imagine what the world might be like if we were really good at making things – better things – cleanly, inexpensively, and on a global scale. What if ultra-efficient solar arrays cost no more to make than cardboard and aluminum foil and laptop computers cost about the same? Now add ultra-efficient vehicles, lighting, and the entire behind-the-scenes infrastructure of an industrial civilization, all made at low cost and delivered and operated with a zero carbon footprint.

Imagine what the world might be like if we were really good at making things – better things – cleanly, inexpensively, and on a global scale. What if ultra-efficient solar arrays cost no more to make than cardboard and aluminum foil and laptop computers cost about the same? Now add ultra-efficient vehicles, lighting, and the entire behind-the-scenes infrastructure of an industrial civilization, all made at low cost and delivered and operated with a zero carbon footprint.

If we were that good at making things, the global prospect would be, not scarcity, but unprecedented abundance – radical, transformative, and sustainable abundance. We would be able to produce radically more of what people want and at a radically lower cost – in every sense of the word, both economic and environmental. **(Eric Drexler, 2013)**

I see a bright future for the biotechnology industry when it follows the path of the computer industry [...], becoming small and domesticated rather than big and centralized. [...]

I see a bright future for the biotechnology industry when it follows the path of the computer industry [...], becoming small and domesticated rather than big and centralized. [...]

Domesticated biotechnology, once it gets into the hands of housewives and children, will give us an explosion of diversity of new living creatures, rather than the monoculture crops that the big corporations prefer. New lineages will proliferate to replace those that monoculture farming and deforestation have destroyed. Designing genomes will be a personal thing, a new art form as creative as painting or sculpture. [...]

I see a bright future for the biotechnology industry when it follows the path of the computer industry [...], becoming small and domesticated rather than big and centralized. [...]

Domesticated biotechnology, once it gets into the hands of housewives and children, will give us an explosion of diversity of new living creatures, rather than the monoculture crops that the big corporations prefer. New lineages will proliferate to replace those that monoculture farming and deforestation have destroyed. Designing genomes will be a personal thing, a new art form as creative as painting or sculpture. [...]

The final step in the domestication of biotechnology will be biotech games, designed like computer games for children down to kindergarten age but played with real eggs and seeds rather than with images on a screen. Playing such games, kids will acquire an intimate feeling for the organisms that they are growing. The winner could be the kid whose seed grows the prickliest cactus, or the kid whose egg hatches the cutest dinosaur. **(Freeman Dyson, 2007)**

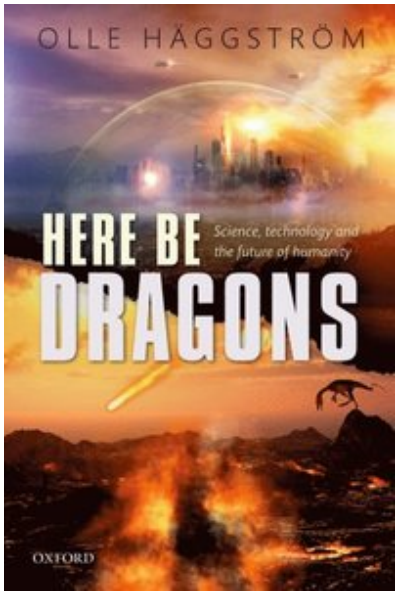
Jämte dessa entusiastiska visioner finns också diverse globala katastrof- och undergångsscenarier. Kärnvapenkrig och klimatkatastrof är de två mest kända exemplen, men det finns en rad andra som vi har anledning att ta på allvar.

Jämte dessa entusiastiska visioner finns också diverse globala katastrof- och undergångsscenarier. Kärnvapenkrig och klimatkatastrof är de två mest kända exemplen, men det finns en rad andra som vi har anledning att ta på allvar.

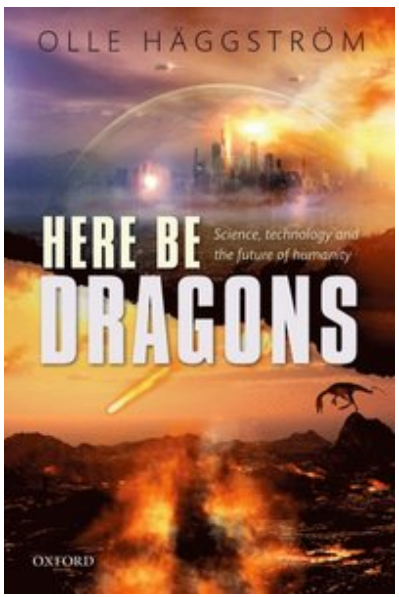

FÖRELÄSNING
Shinya Yamanaka, nobelpristagare i fysiologi eller medicin föreläser om "A New Era of Medicine with iPS Cells"
Torsdag den 7 september
09:00–10:00
Läs mer och anmäl dig på chalmers.se
PA GANG
INFORMA
3

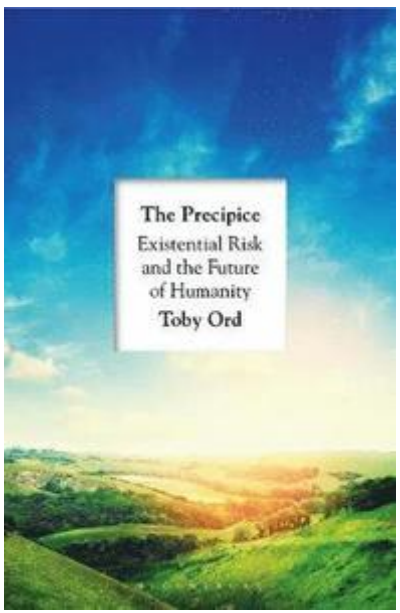
Breda svep över olika existentiella risker ges i var och en av följande böcker.

Breda svep över olika existentiella risker ges i var och en av följande böcker.

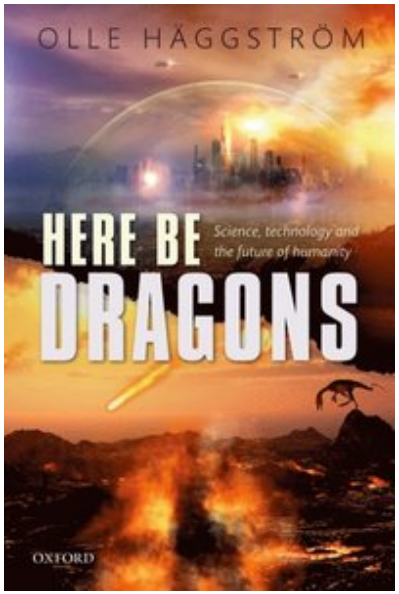


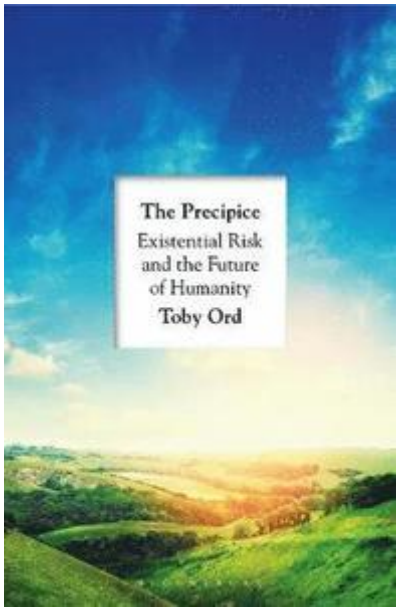
Breda svep över olika existentiella risker ges i var och en av följande böcker.

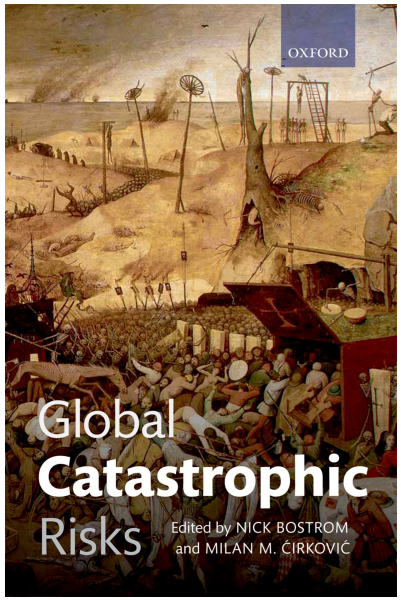




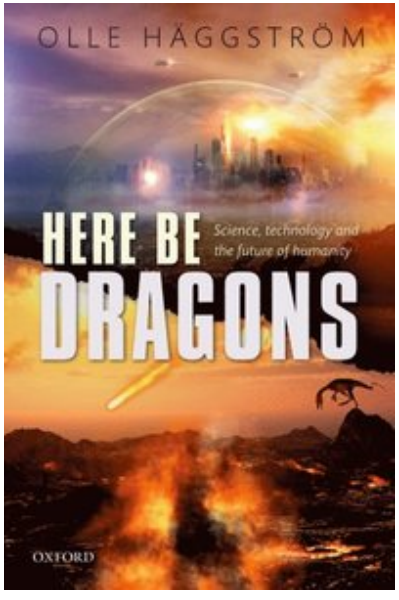
Breda svep över olika existentiella risker ges i var och en av följande böcker.

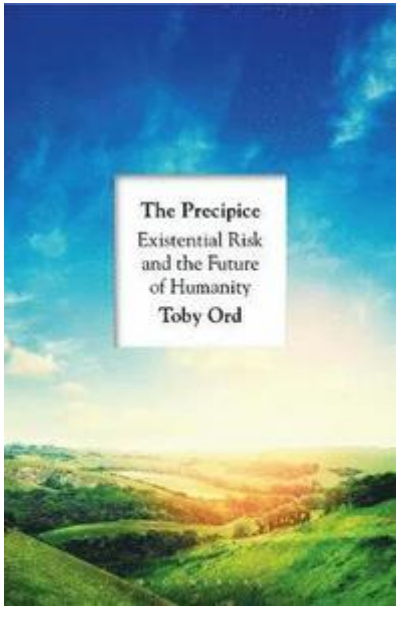


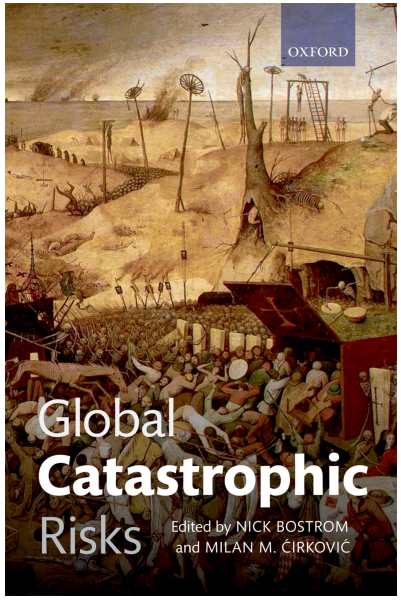


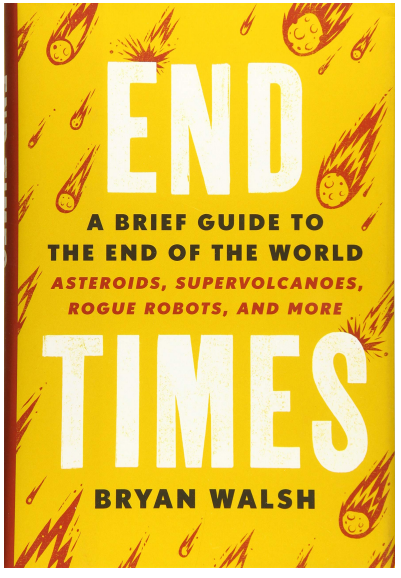


Breda svep över olika existentiella risker ges i var och en av följande böcker.









Några andra källor om existentiell risk:

Några andra källor om existentiell risk:

- Olle Häggström och Catherine Rhodes (2019) *Existential Risk to Humanity*, specialnummer av tidskriften *Foresight*.

Några andra källor om existentiell risk:

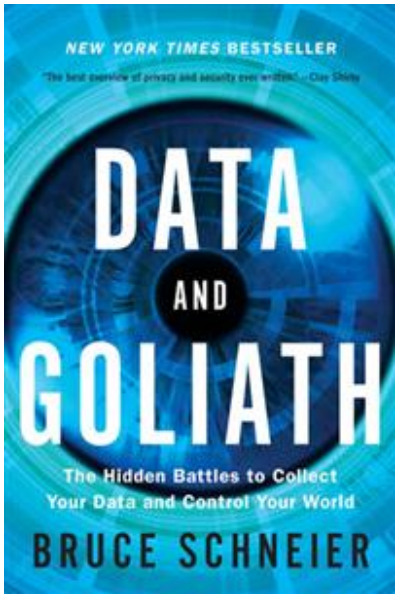
- Olle Häggström och Catherine Rhodes (2019) *Existential Risk to Humanity*, specialnummer av tidskriften *Foresight*.
- Dennis Pamlin och Stuart Armstrong (2015) *Global Challenges – 12 Risks that Threaten Civilization*, Global Challenges Foundation.

Några andra källor om existentiell risk:

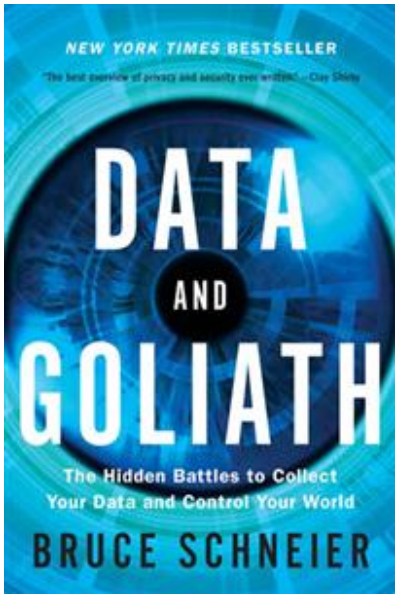
- Olle Häggström och Catherine Rhodes (2019) *Existential Risk to Humanity*, specialnummer av tidskriften *Foresight*.
- Dennis Pamlin och Stuart Armstrong (2015) *Global Challenges – 12 Risks that Threaten Civilization*, Global Challenges Foundation.
- Nick Bostrom (2019) The vulnerable world hypothesis, *Global Policy* **10**, 455-476.

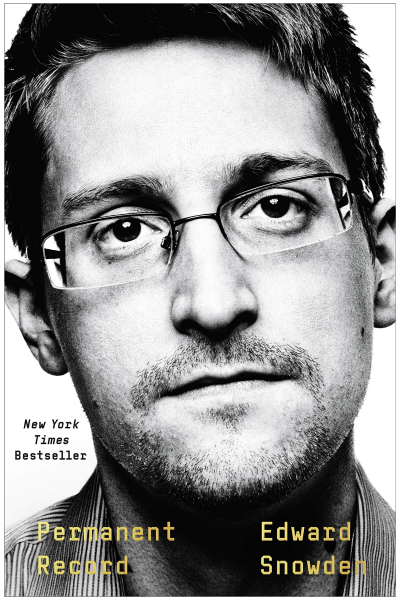
På tal om massövervakning finns annat att ta del av:

På tal om massövervakning finns annat att ta del av:

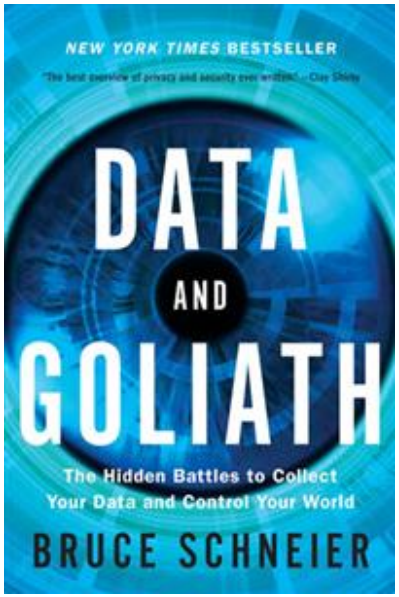


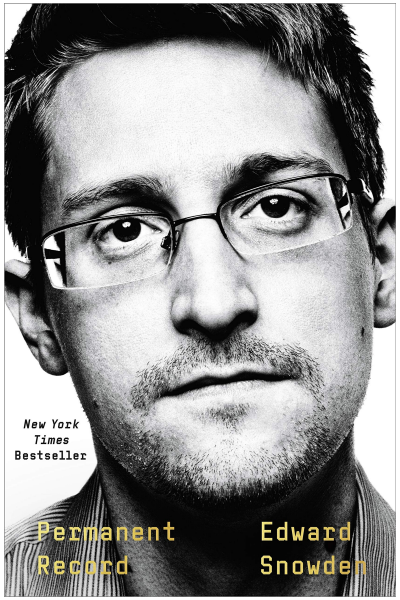
På tal om massövervakning finns annat att ta del av:

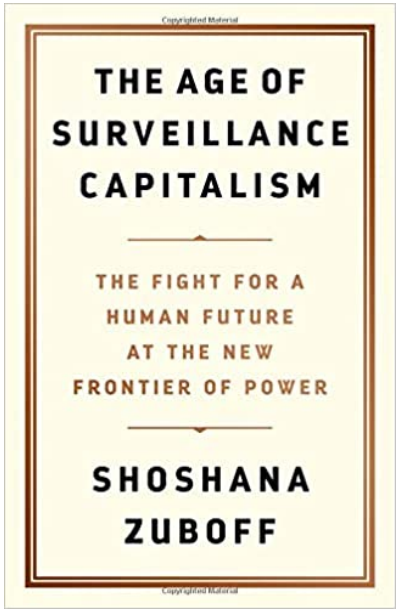




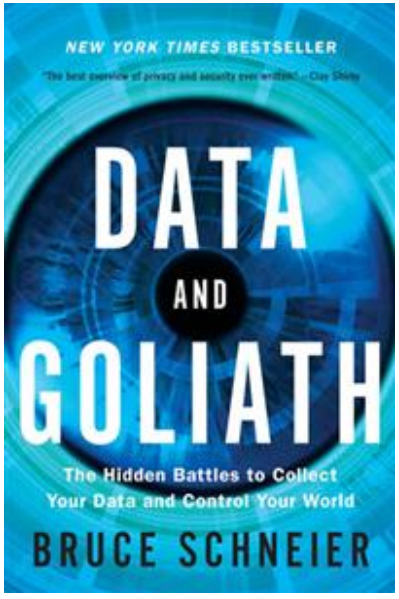
På tal om massövervakning finns annat att ta del av:

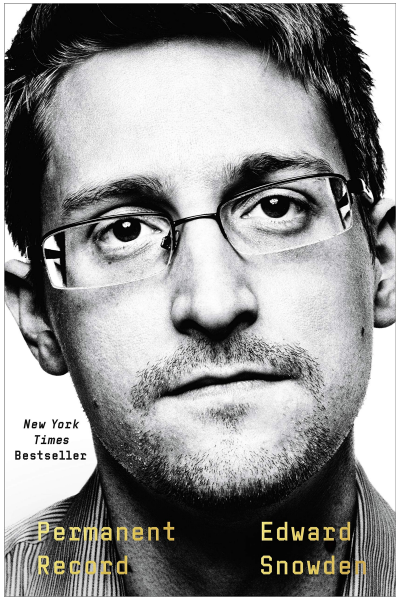


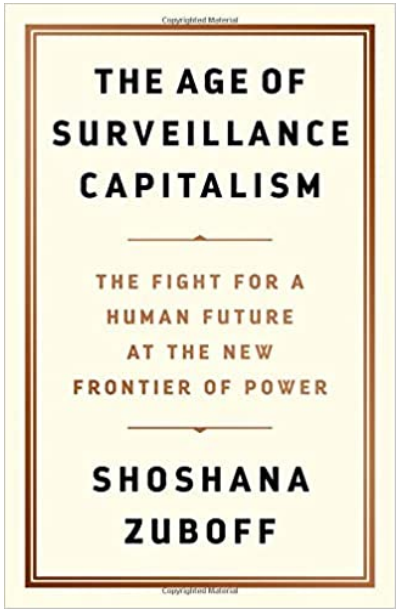




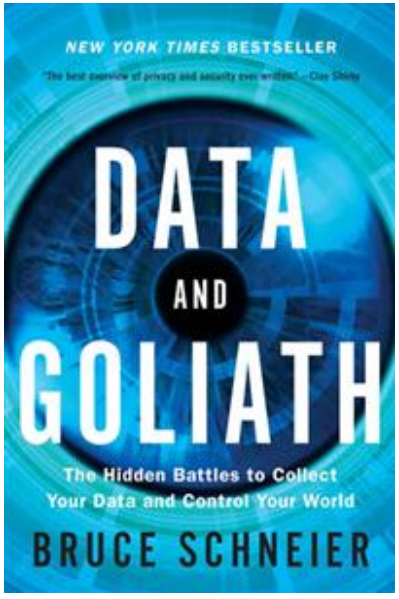
På tal om massövervakning finns annat att ta del av:

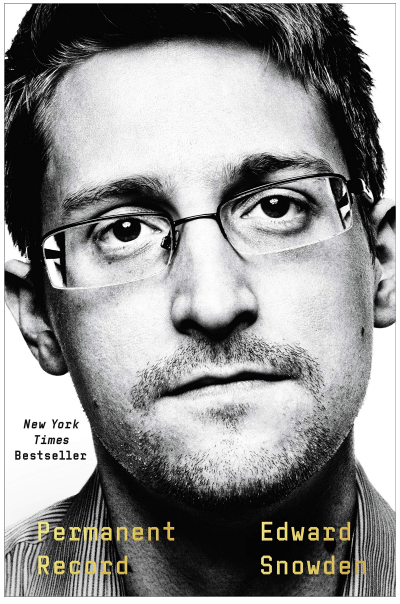


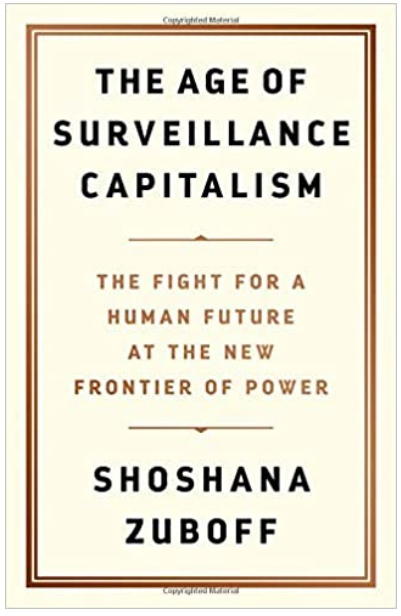




På tal om massövervakning finns annat att ta del av:







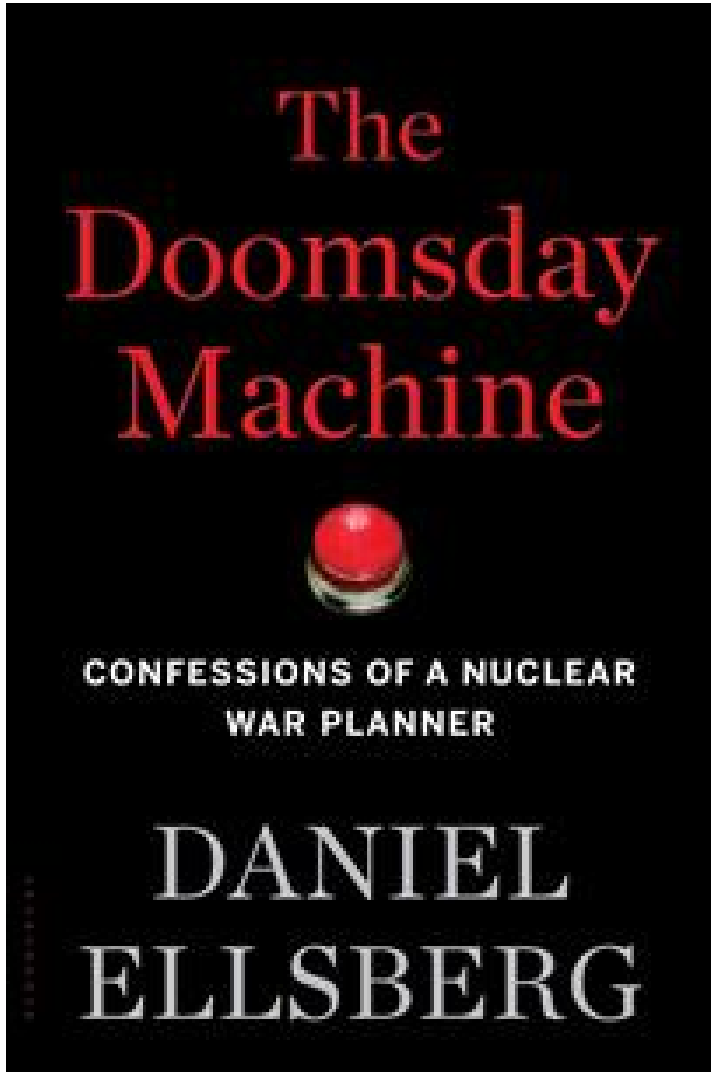
En av huvudpersonerna i *The Social Dilemma* är Tristan Harris, som nyligen blev djupintervjuad i *80,000 Hours Podcast*: `https://80000hours.org/podcast/episodes/tristan-harris-changing-incentives-social-media/`

Internet of Things

Internet of Things

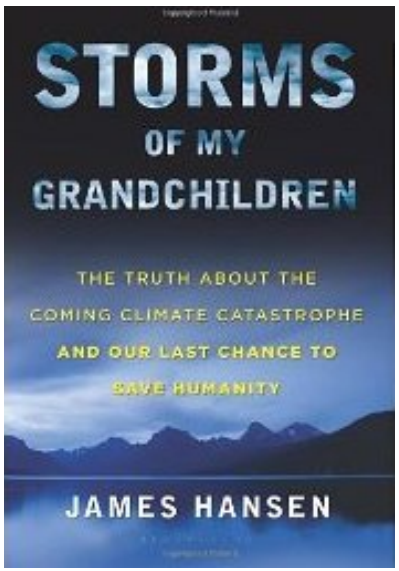
Kärnvapen

Kärnvapen

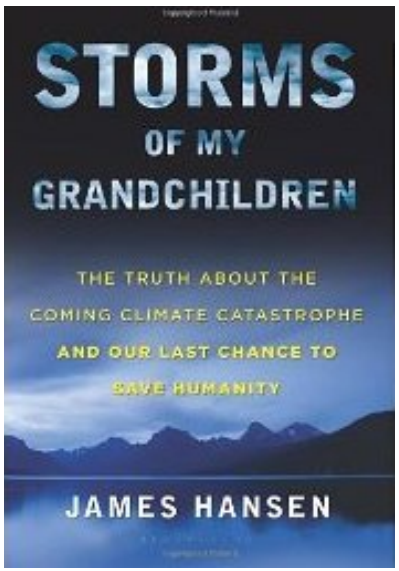

The
Doomsday
Machine
CONFESSIONS OF A NUCLEAR
WAR PLANNER
DANIEL
ELLSBERG

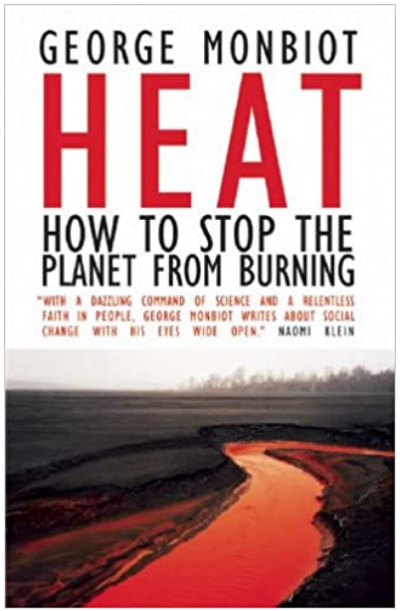
Klimatförändringar

Klimatförändringar

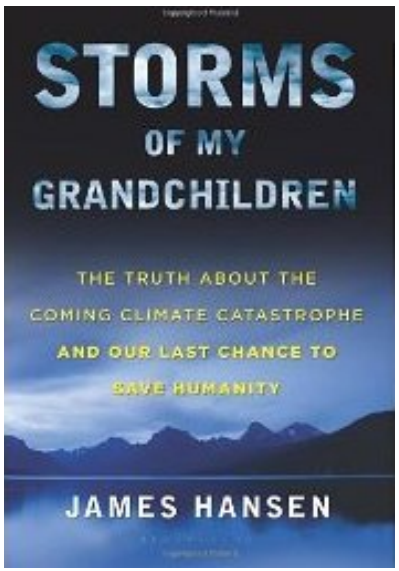


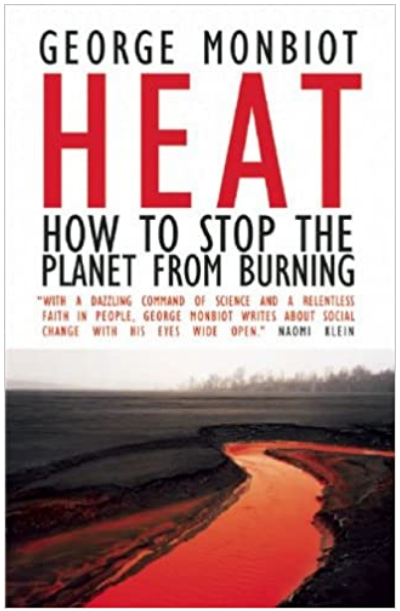
Klimatförändringar



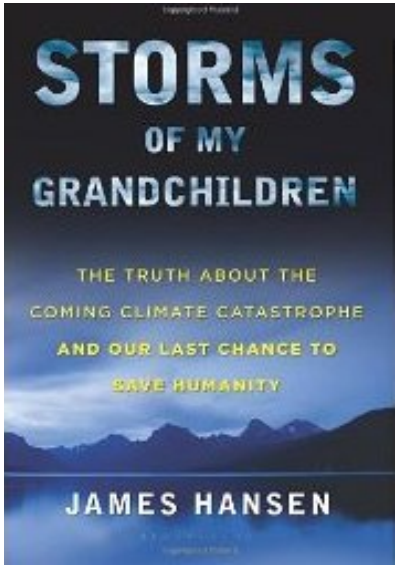


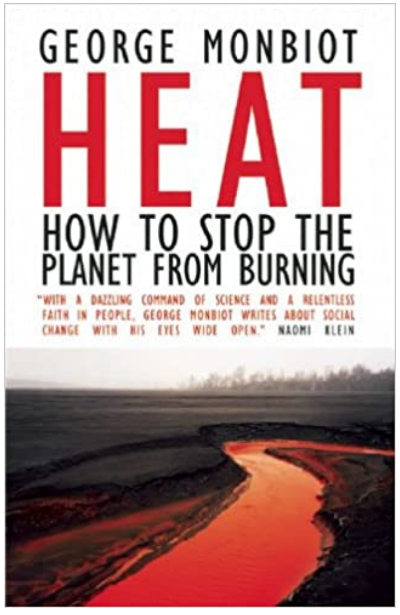
Klimatförändringar

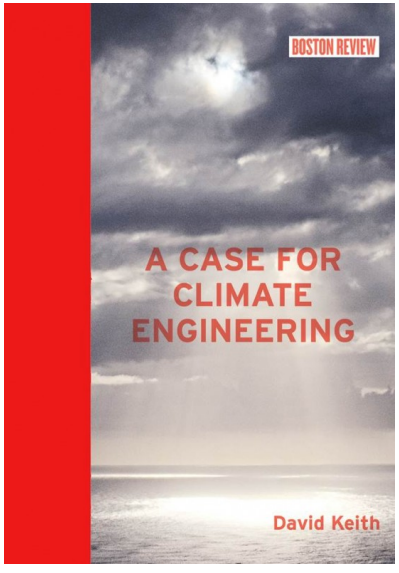




Klimatförändringar

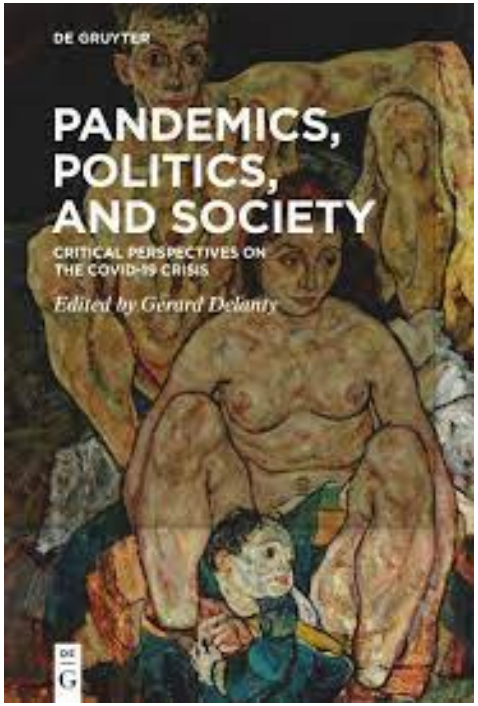




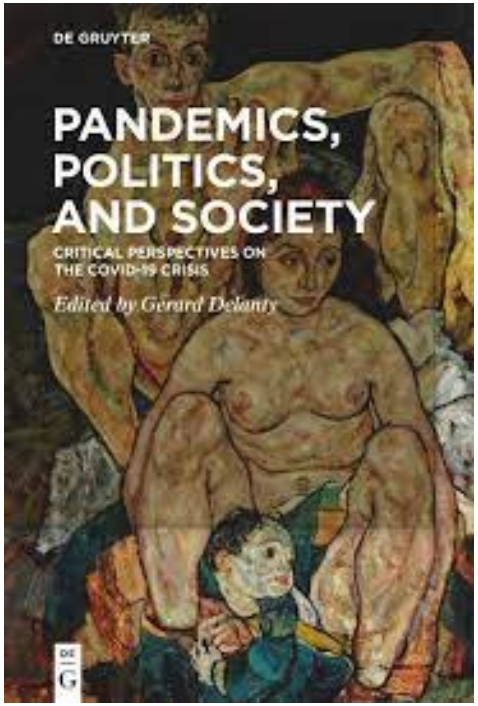


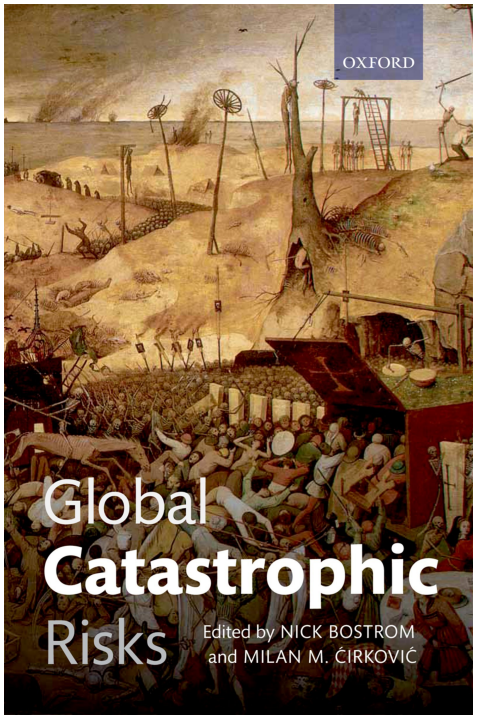
Covid-19 och andra pandemier

Covid-19 och andra pandemier

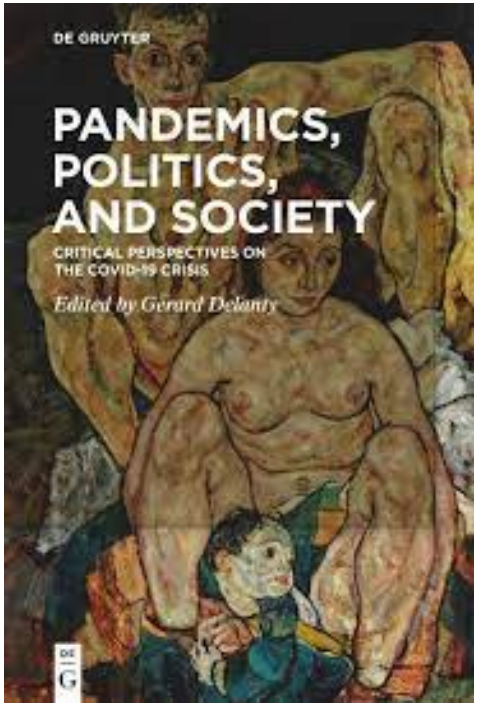


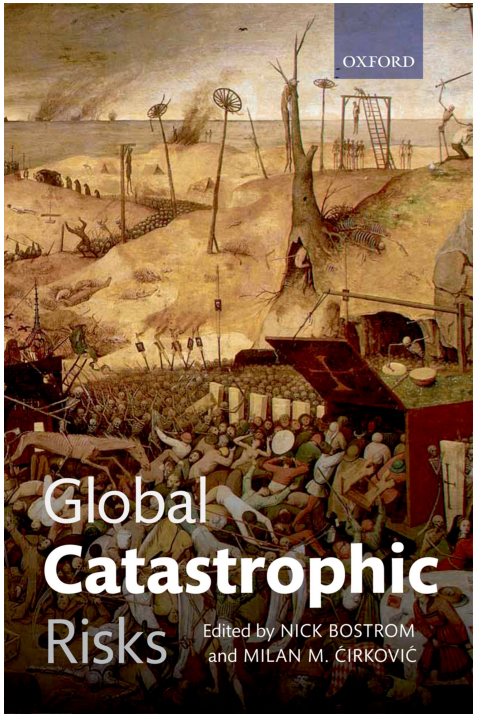
Covid-19 och andra pandemier

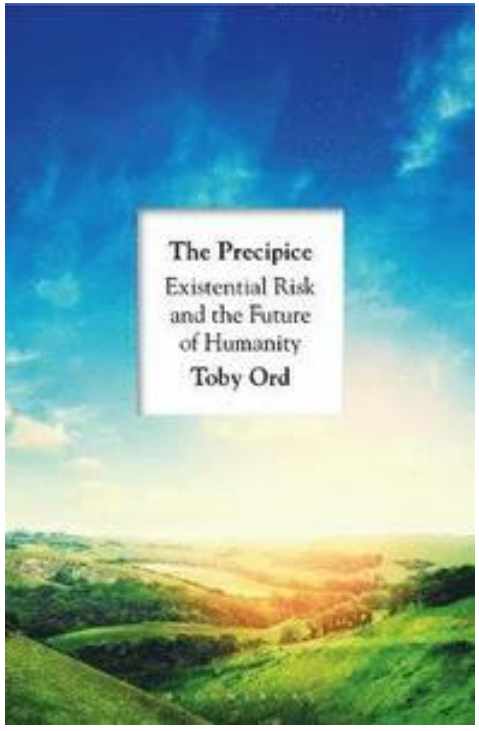




Covid-19 och andra pandemier

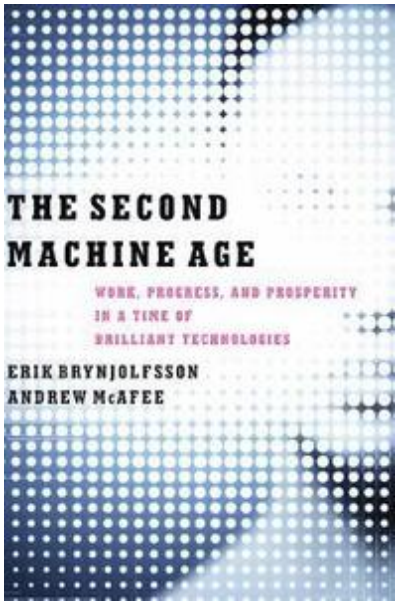




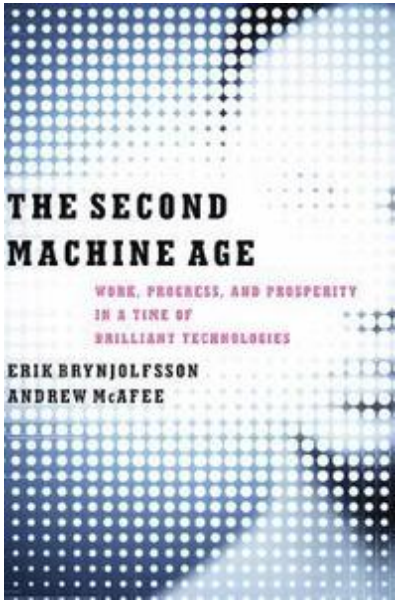


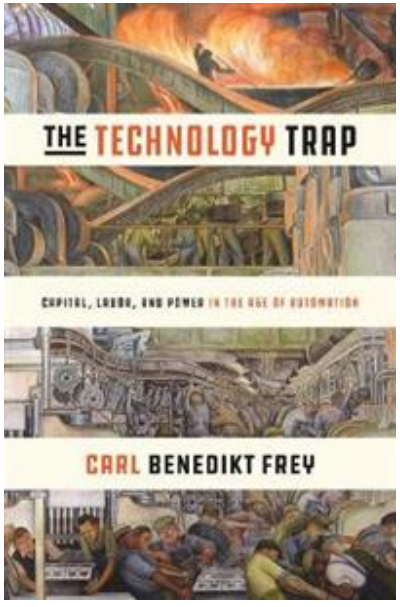
Automatisering och arbetsmarknad

Automatisering och arbetsmarknad

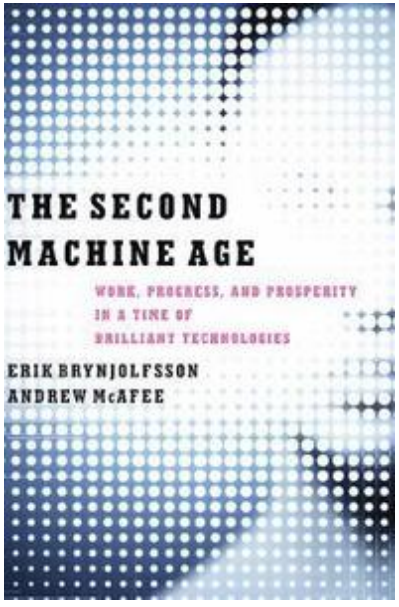


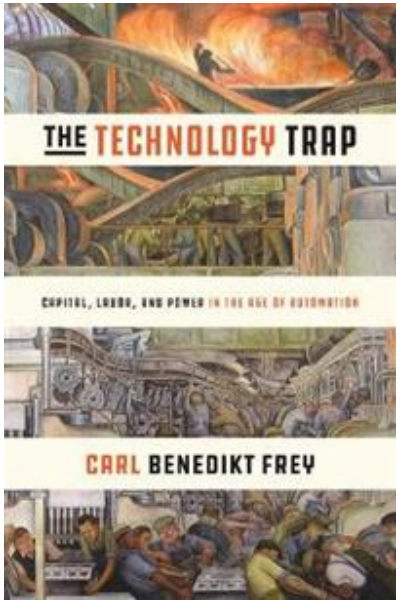
Automatisering och arbetsmarknad

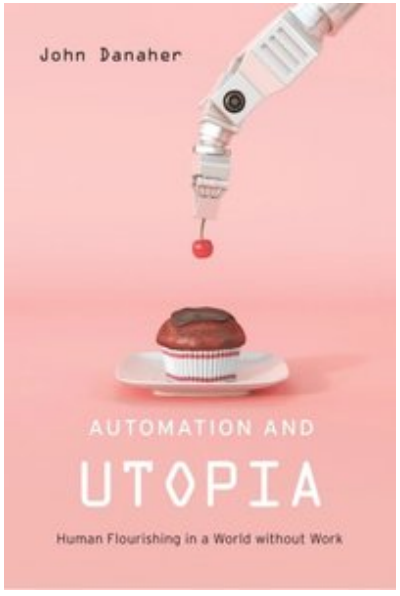




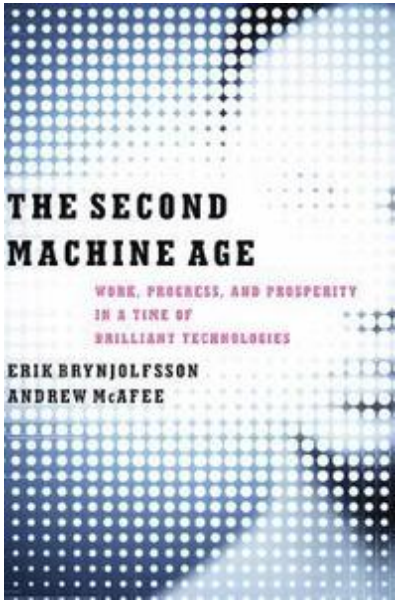
Automatisering och arbetsmarknad

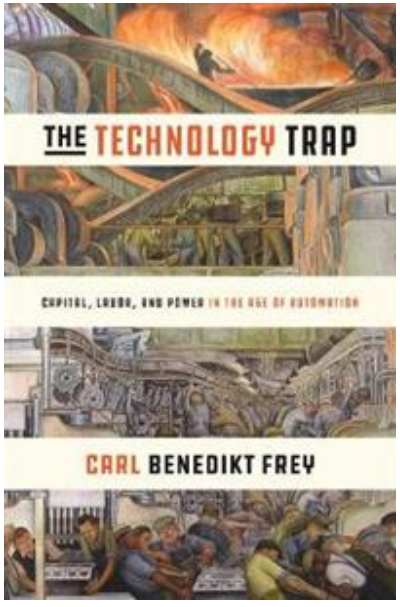


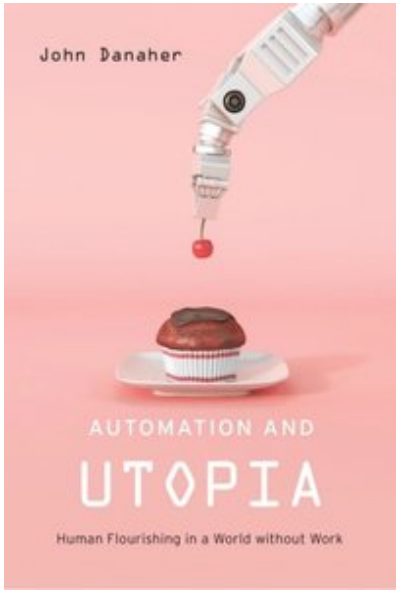


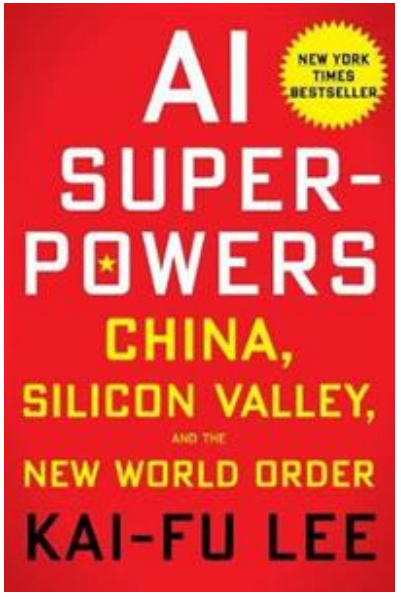


Automatisering och arbetsmarknad



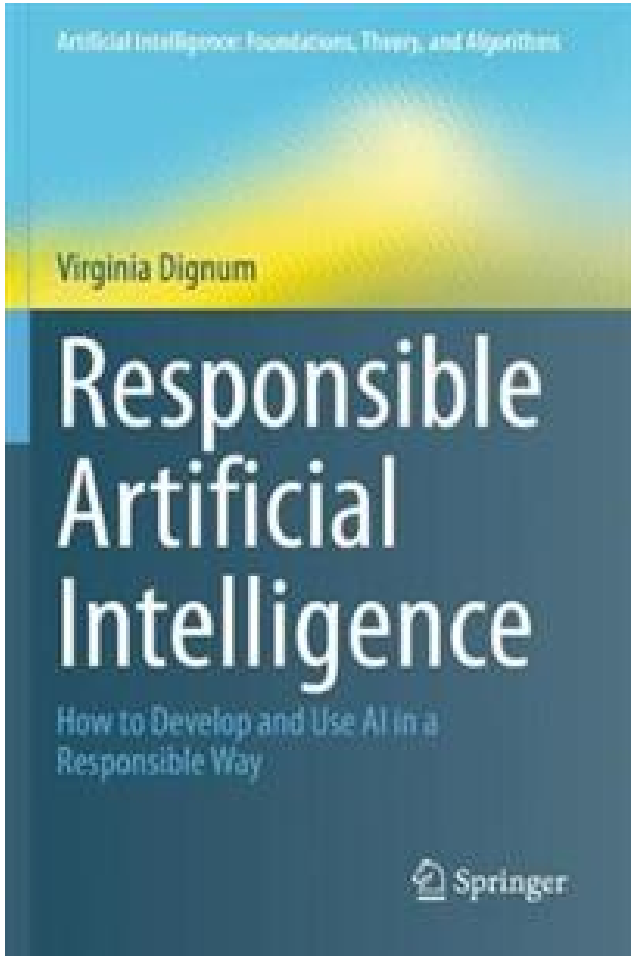




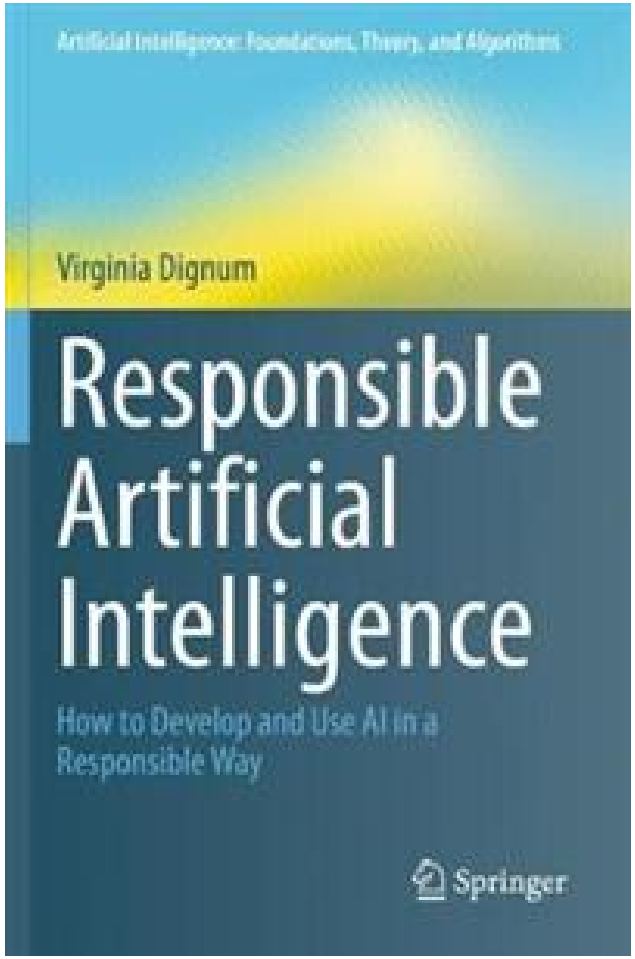


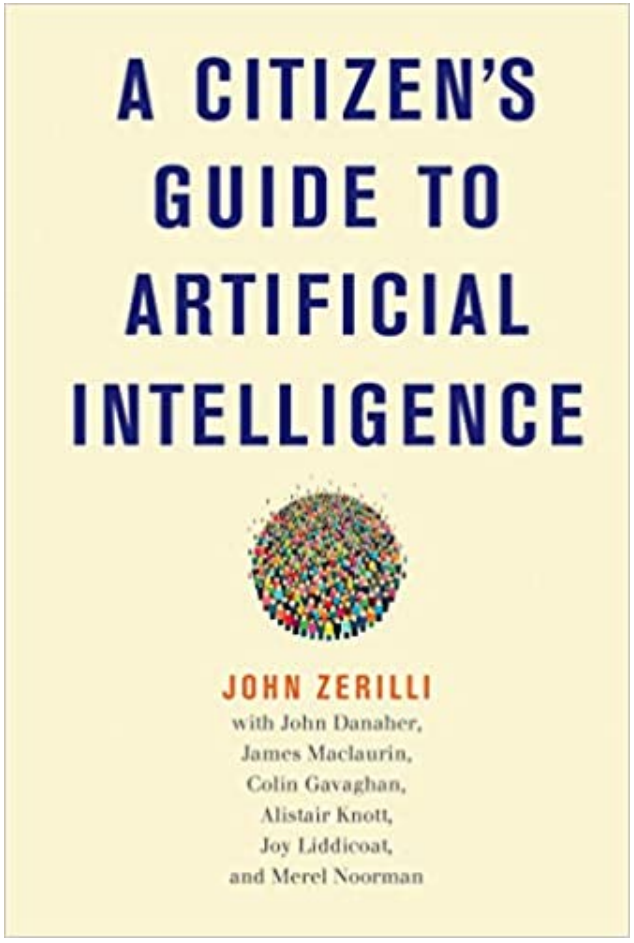
Andra samhällsaspekter på AI

Andra samhällsaspekter på AI



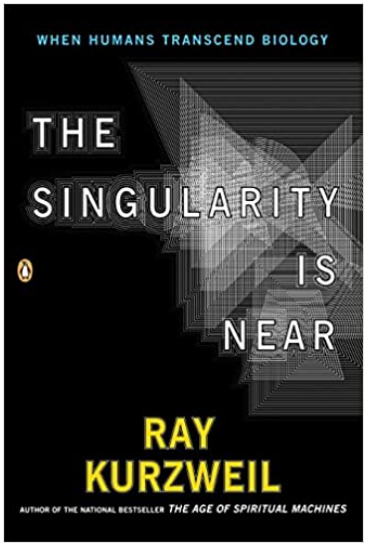
Andra samhällsaspekter på AI



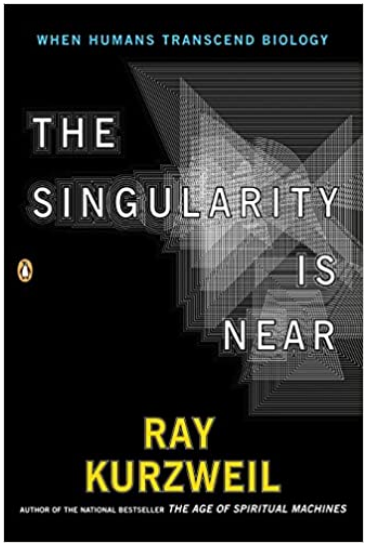


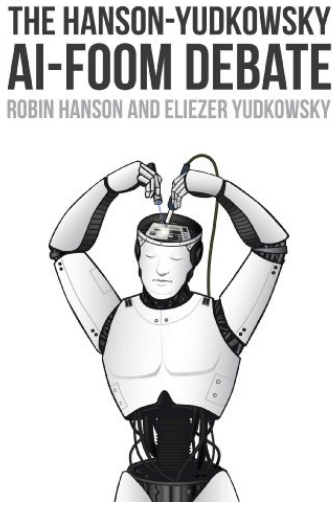
Det ultimata AI-genombrottet

Det ultimata AI-genombrottet

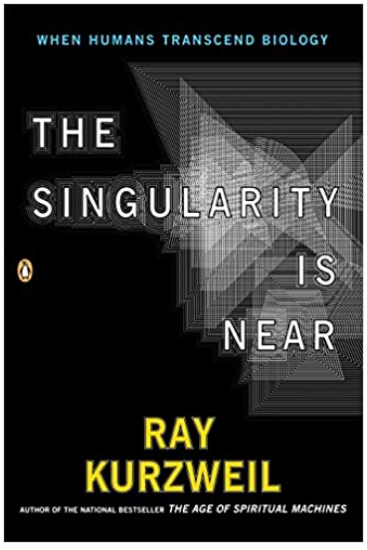


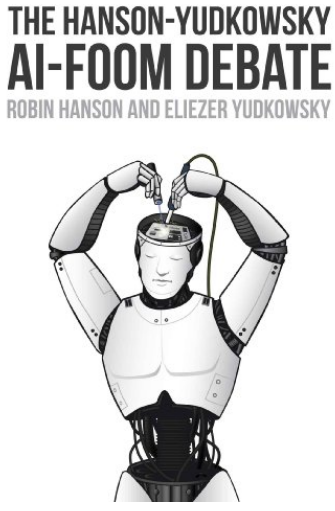
Det ultimata AI-genombrottet

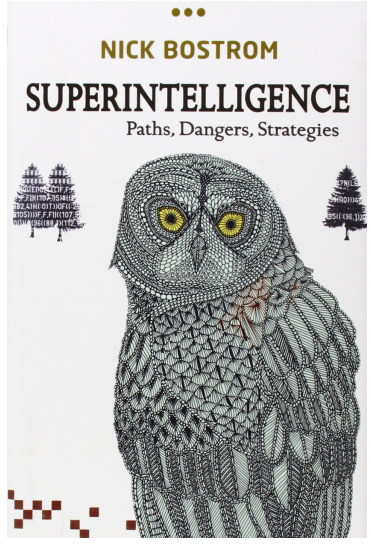




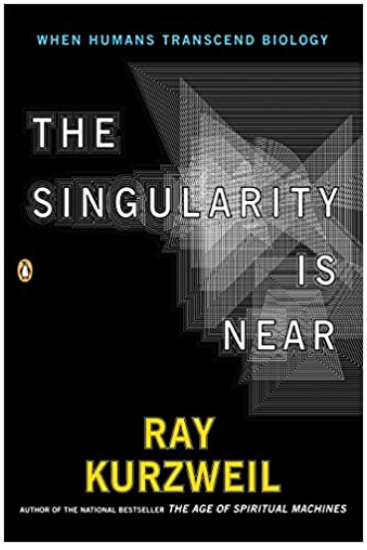
Det ultimata AI-genombrottet

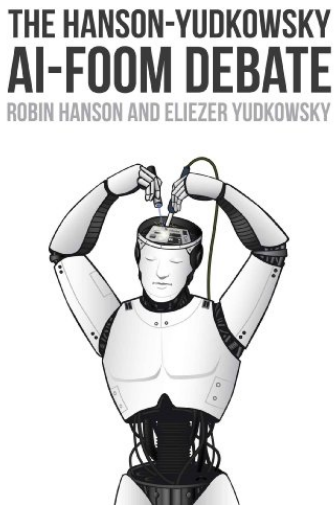


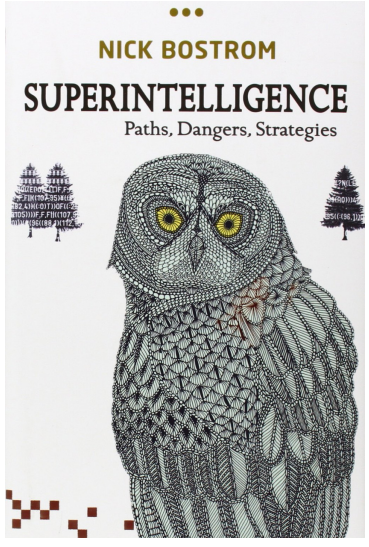


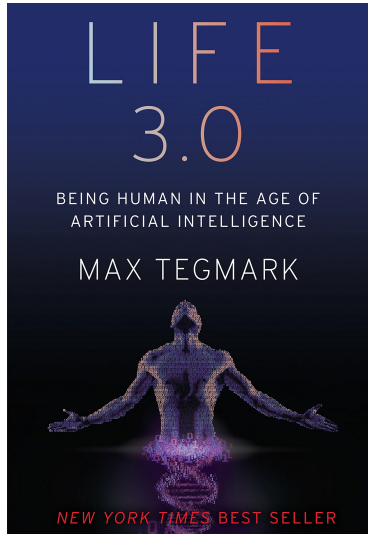


Det ultimata AI-genombrottet

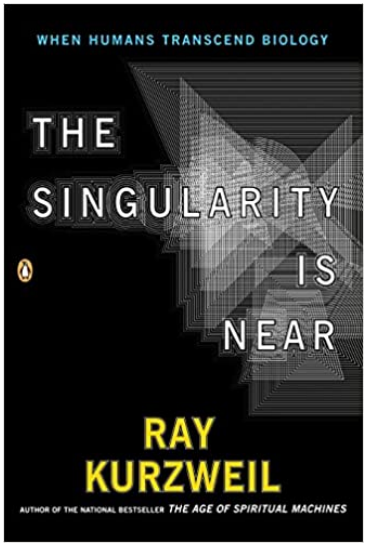


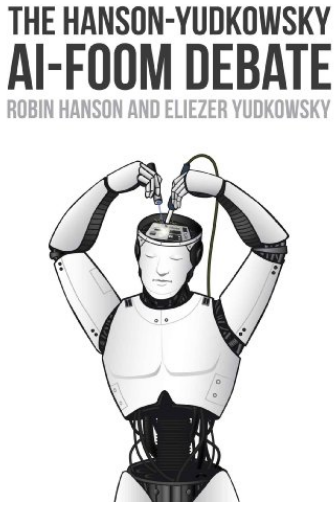


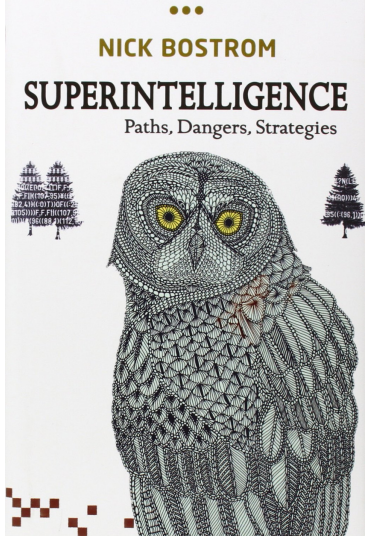


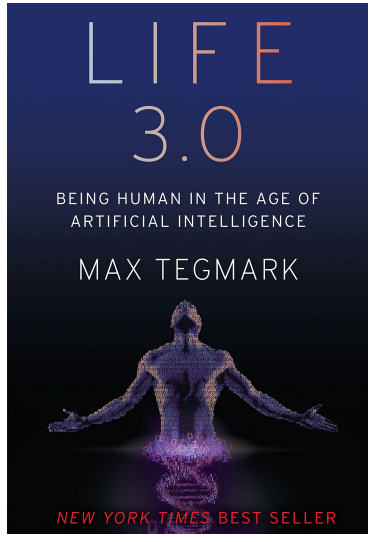


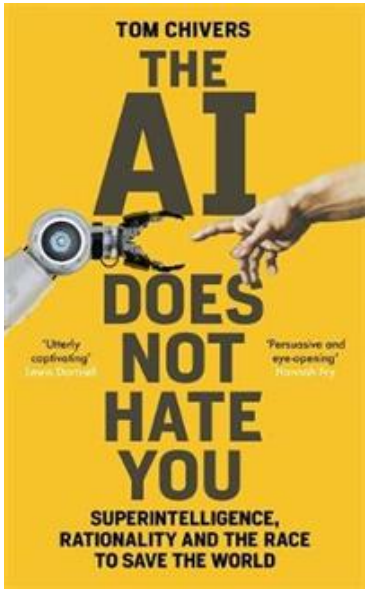
Det ultimata AI-genombrottet



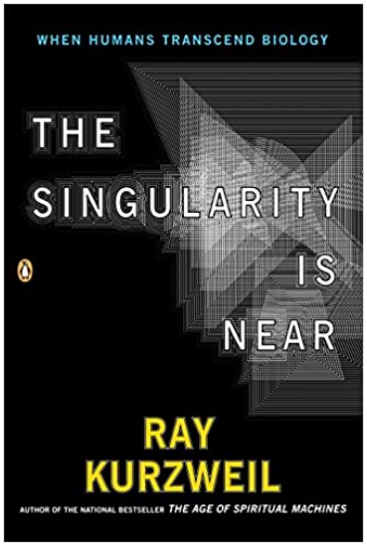


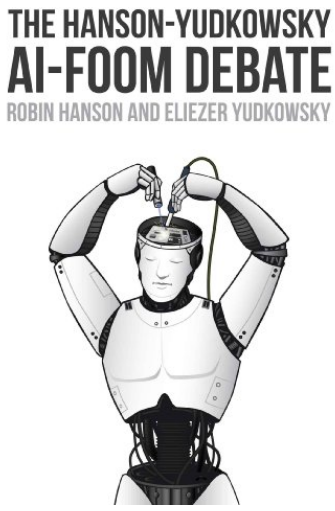


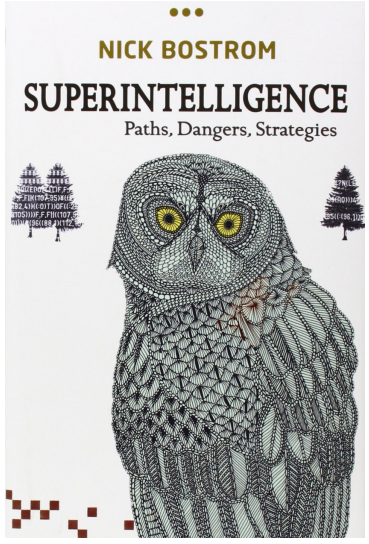


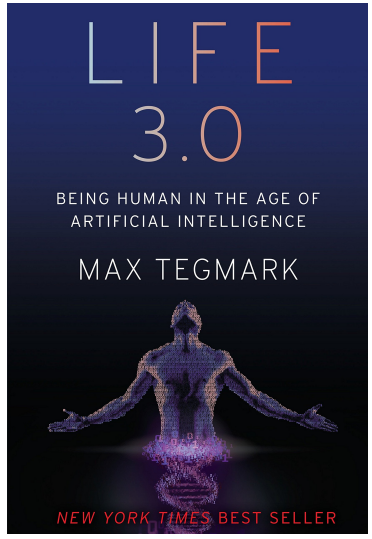


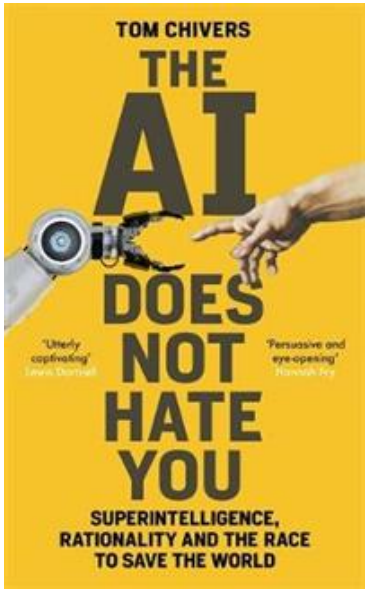
Det ultimata AI-genombrottet

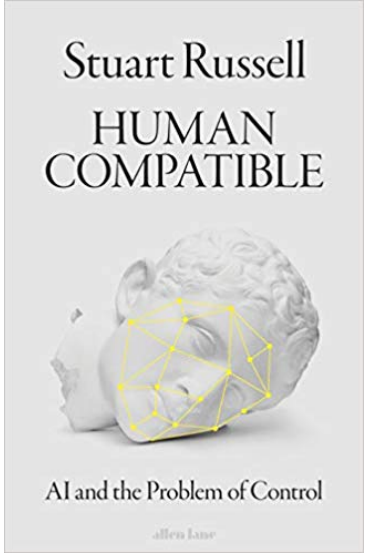




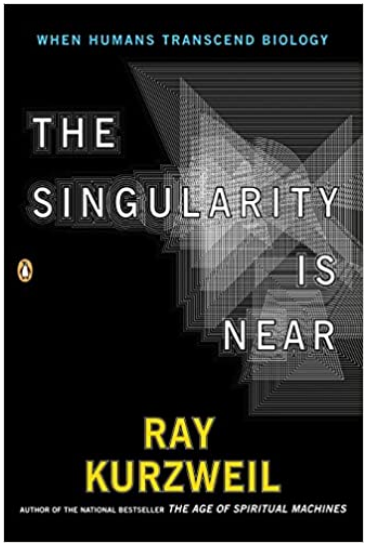


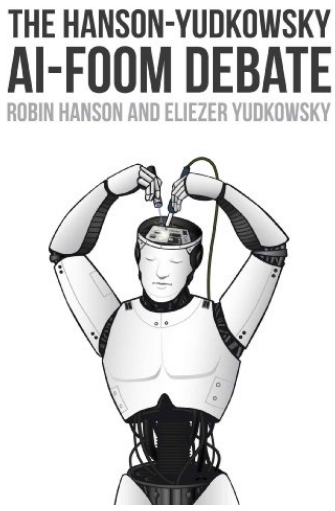


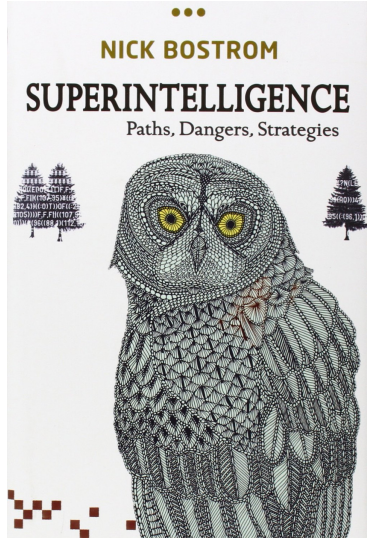


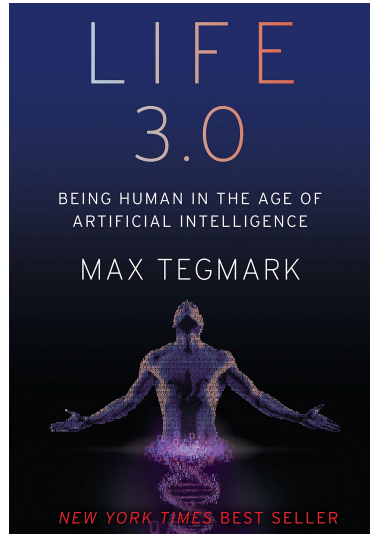


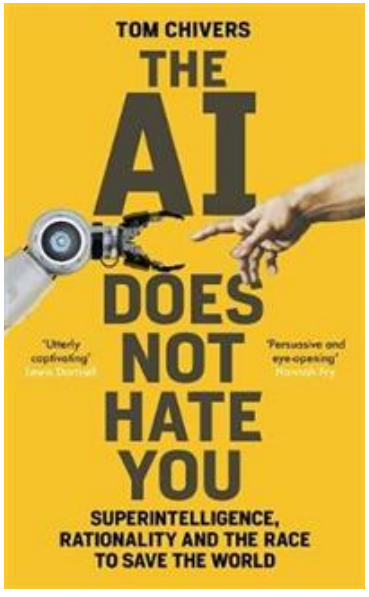
Det ultimata AI-genombrottet

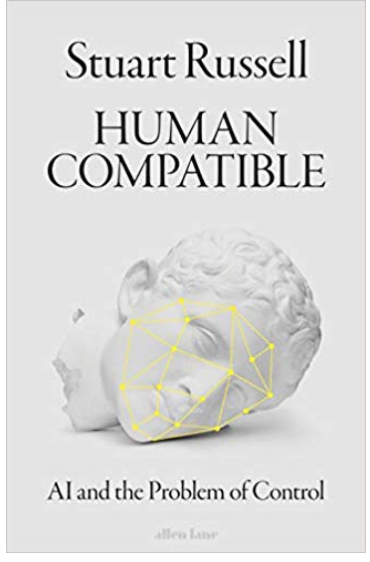


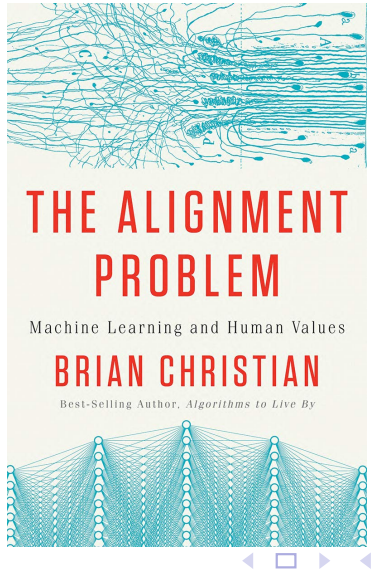




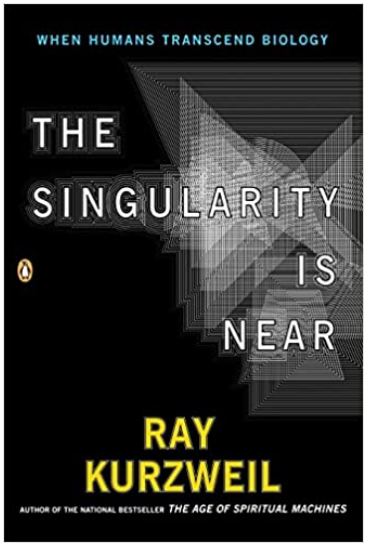


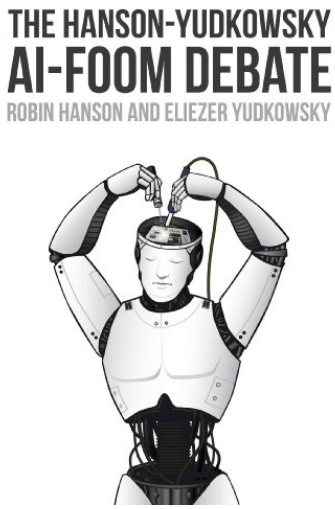


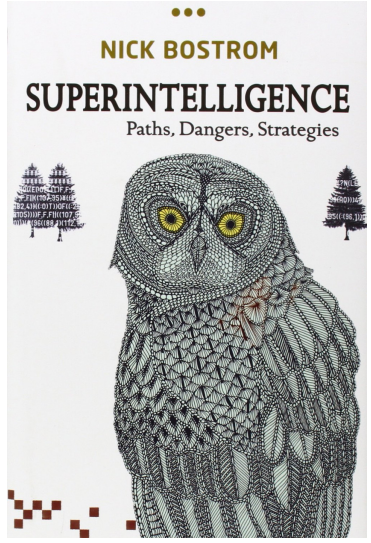


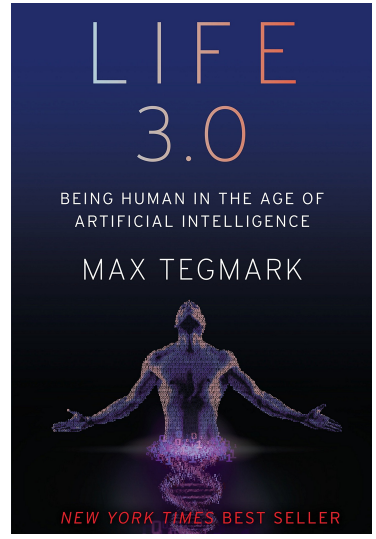


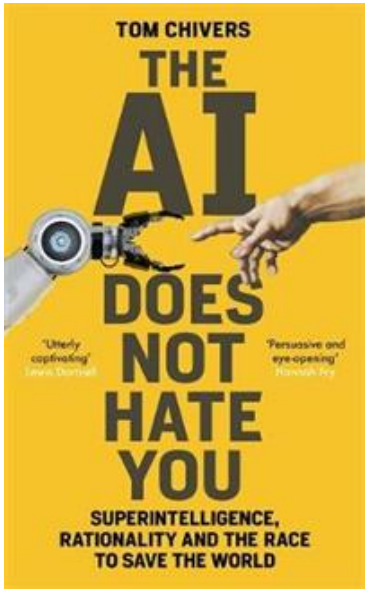
Det ultimata AI-genombrottet

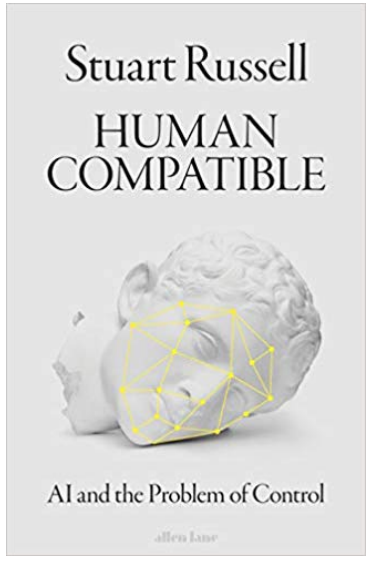


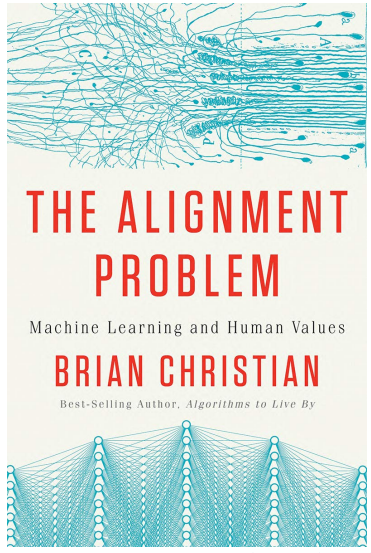


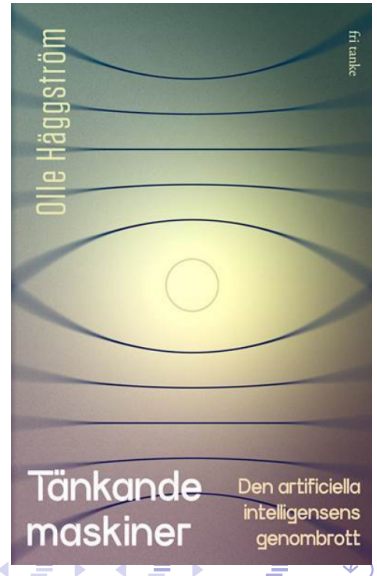












AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

David Krueger
MILA
Université de Montréal

June 11, 2020

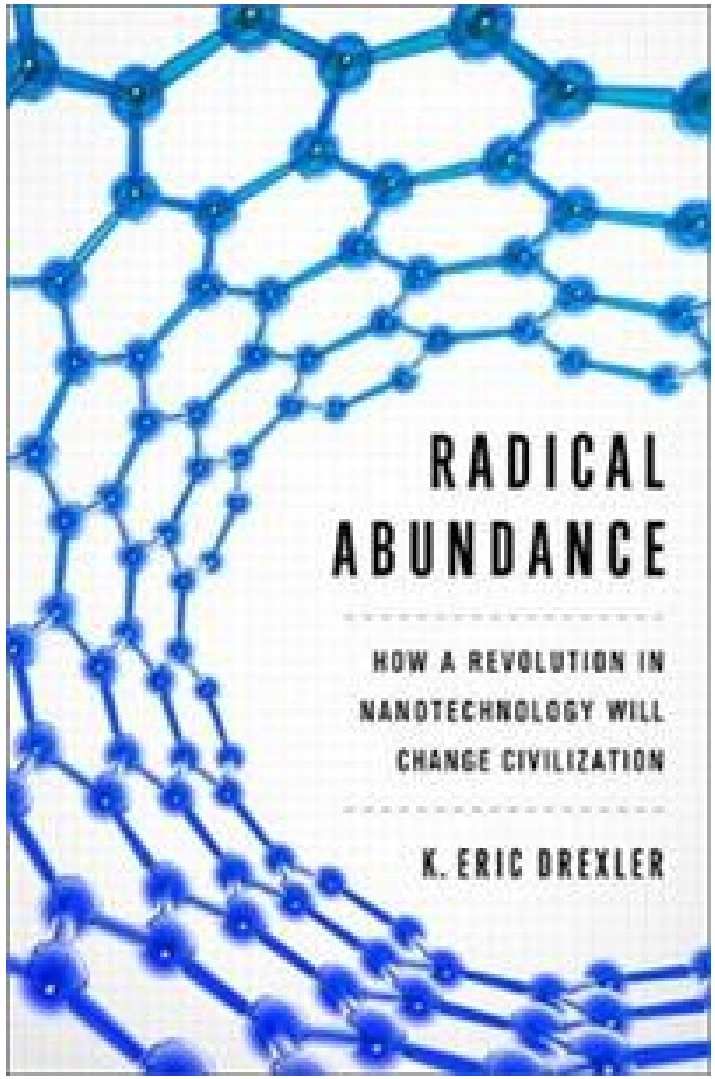
Abstract

Framed in positive terms, this report examines how technical AI research might be steered in a manner that is more attentive to humanity's long-term prospects for survival as a species. In negative terms, we ask what existential risks humanity might face from AI development in the next century, and by what principles contemporary technical research might be directed to address those risks.

A key property of hypothetical AI technologies is introduced, called *prepotence*, which is useful for delineating a variety of potential existential risks from artificial intelligence, even as AI paradigms might shift. A set of twenty-nine contemporary research directions are then examined for their potential benefit to existential safety. Each research direction is explained with a scenario-driven motivation, and examples of existing work from which to build. The research directions present their own risks and benefits to society that could occur at various scales of impact, and in particular are not guaran-

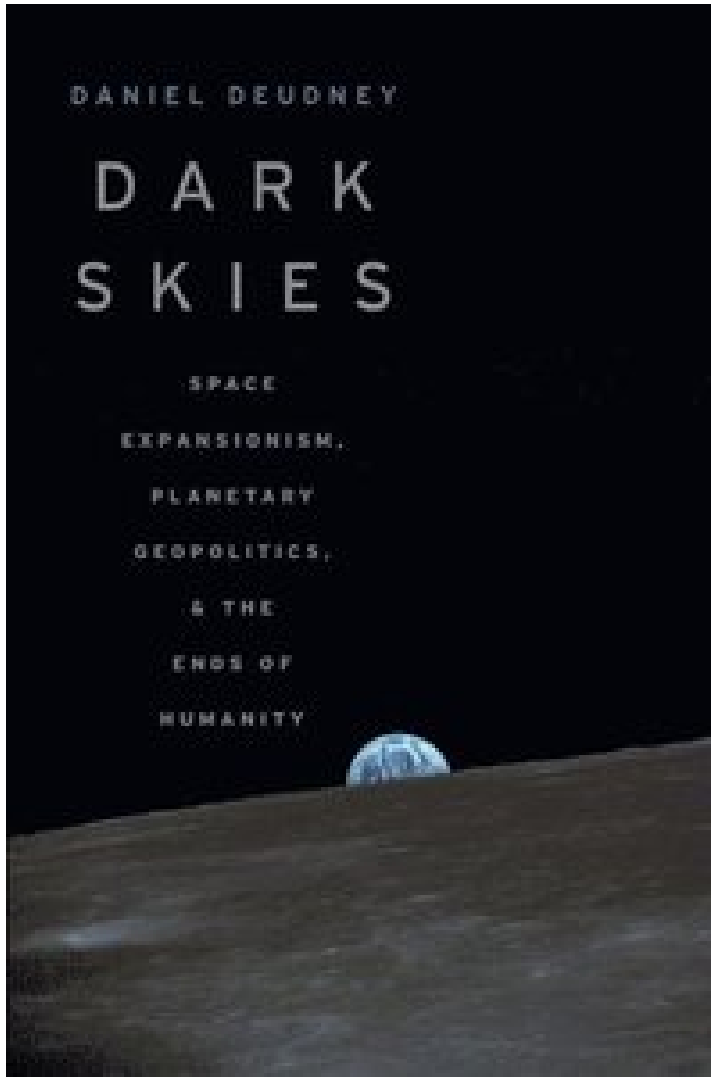
Radikal nanoteknologi

Radikal nanoteknologi



Rymdkolonisering

Rymdkolonisering



Mer om rymdkolonisering

Mer om rymdkolonisering

- Al Globus (2020) Not so dark skies, *The Space Review*, 13 juli.

Mer om rymdkolonisering

- Al Globus (2020) Not so dark skies, *The Space Review*, 13 juli.
- Phil Torres (2018) Space colonization and suffering risks: Reasessing the "maxipok" rule, *Futures* **100**, 74–85.

Mer om rymdkolonisering

- Al Globus (2020) Not so dark skies, *The Space Review*, 13 juli.
- Phil Torres (2018) Space colonization and suffering risks: Reasessing the "maxipok" rule, *Futures* **100**, 74–85.
- Milan Cirkovic (2018) Space colonization remains the only long-term option for humanity: a reply to Torres, *Futures* **105**, 166–173.

Mer om rymdkolonisering

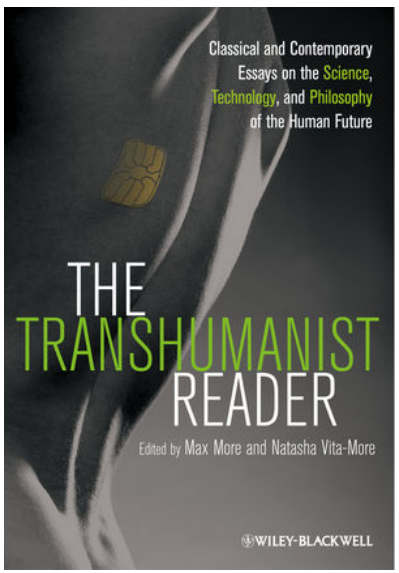
- Al Globus (2020) Not so dark skies, *The Space Review*, 13 juli.
- Phil Torres (2018) Space colonization and suffering risks: Reasessing the "maxipok" rule, *Futures* **100**, 74–85.
- Milan Cirkovic (2018) Space colonization remains the only long-term option for humanity: a reply to Torres, *Futures* **105**, 166–173.
- Tom Murphy (2011) Stranded resources, *Do the Math*, 25 oktober.

Mer om rymdkolonisering

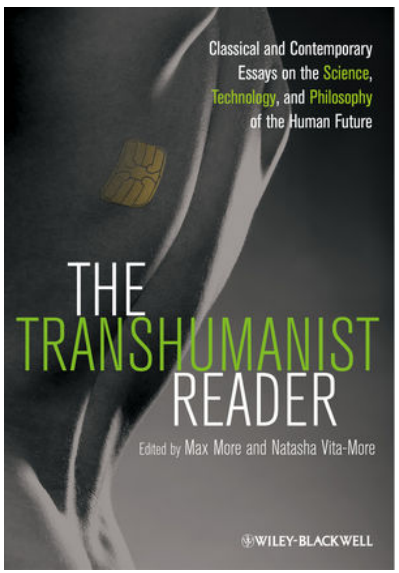
- Al Globus (2020) Not so dark skies, *The Space Review*, 13 juli.
- Phil Torres (2018) Space colonization and suffering risks: Reasessing the "maxipok" rule, *Futures* **100**, 74–85.
- Milan Cirkovic (2018) Space colonization remains the only long-term option for humanity: a reply to Torres, *Futures* **105**, 166–173.
- Tom Murphy (2011) Stranded resources, *Do the Math*, 25 oktober.
- Stuart Armstrong och Anders Sandberg (2013) Eternity in six hours: intergalactic spreading of intelligent life and sharpening the Fermi paradox, *Acta Astronautica* **89**, 1–13.

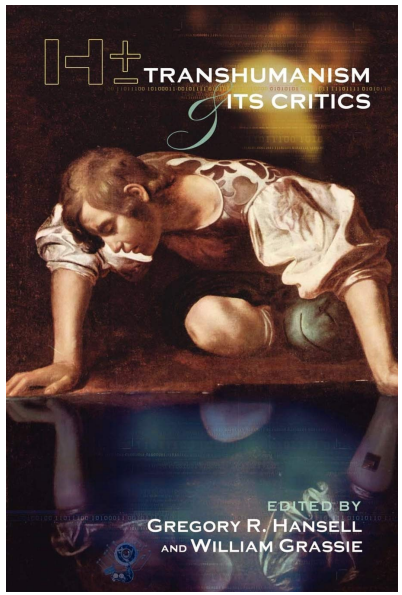
Transhumanistisk modifiering av människan

Transhumanistisk modifiering av människan

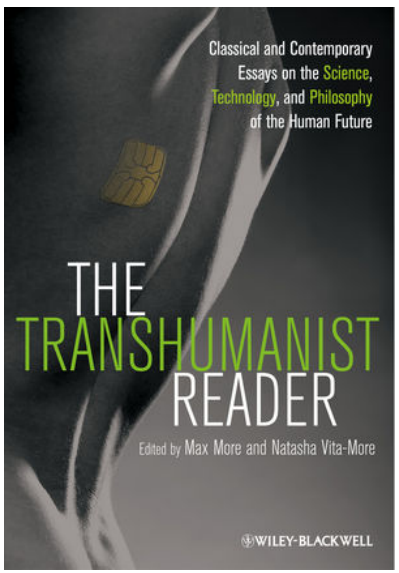


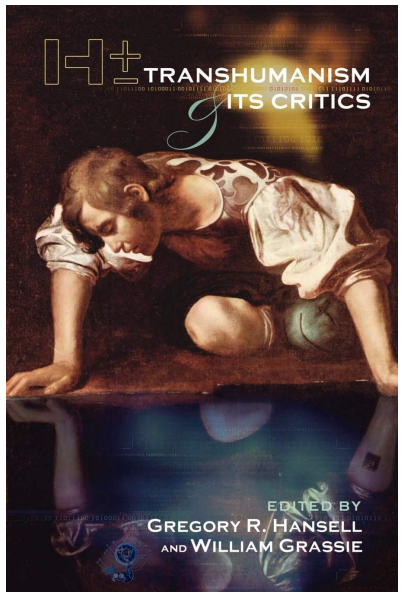
Transhumanistisk modifiering av människan

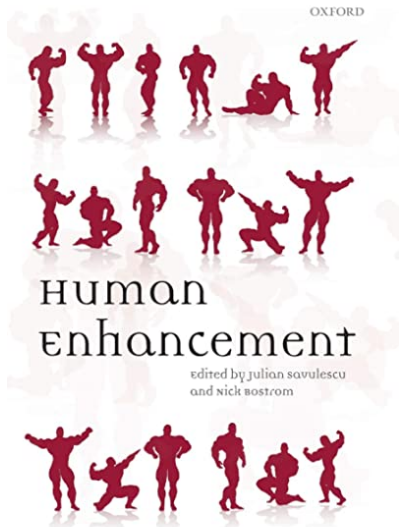




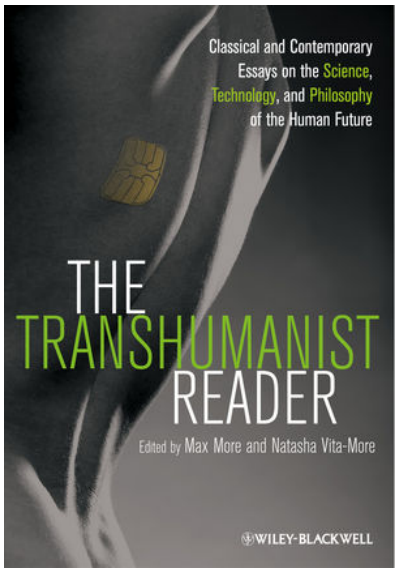
Transhumanistisk modifiering av människan

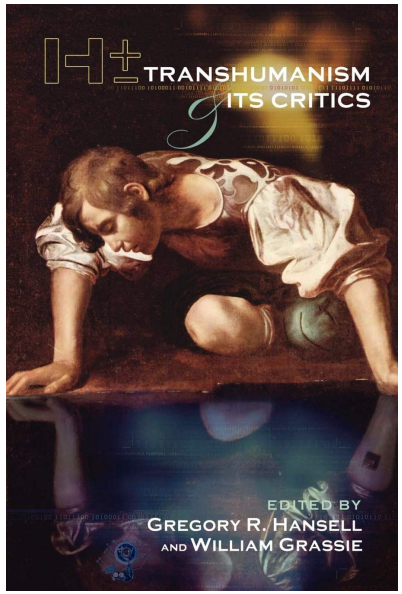


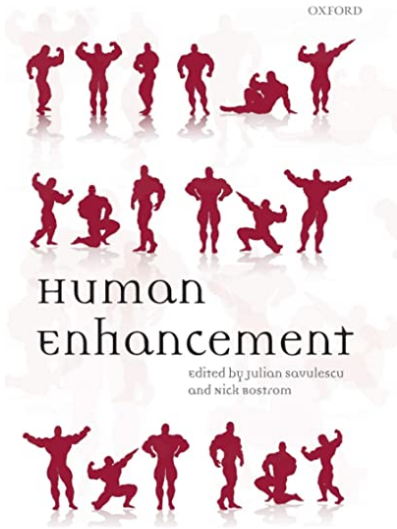


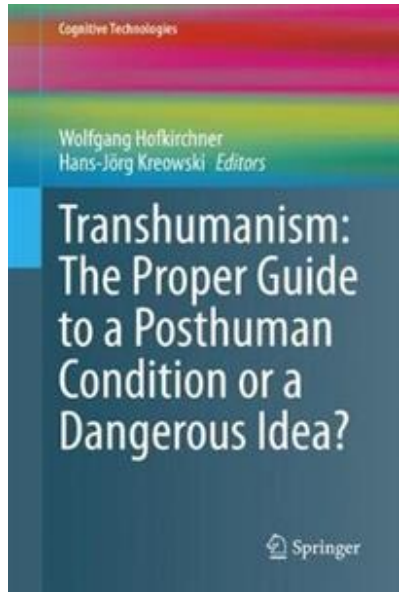


Transhumanistisk modifiering av människan









Uppladdning av mänskligt medvetande

Uppladdning av mänskligt medvetande