MVE420:
Nya teknologier, global risk och mänsklighetens framtid

http://www.math.chalmers.se/Math/Grundutb/CTH/mve420/1718/

Föreläsning om

Extrema AI-scenarier

4 maj 2018

Olle Häggström

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

- Big data, massövervakning och personlig integritet

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

- Big data, massövervakning och personlig integritet
- Autonoma vapen och annan militär AI-teknologi

DAGENS Samhälle DEBATT ALL DEBATT

Skola Sjukvård Omsorg Infrastruktur Upphandling Arbete Miljö Ekonomi Demokrati Krönikor

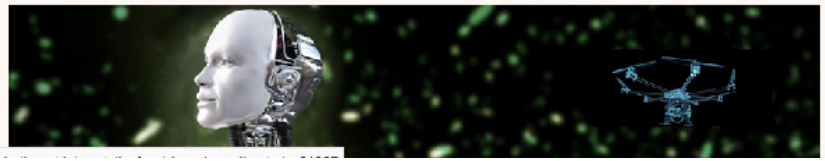
38 LEDIGA JOBB: **Mark- och Exploateringschef** Norrtälje kommun

Försvarspolitik

Förbjud dödliga autonoma vapensystem

På torsdag röstar riksdagen om en av vår tids avgörande frågor för framtidens krig. Ska dödliga autonoma vapensystem – ibland kallade "killer robots" eller "mördarrobotar" – tillåtas i krig? Vi, i enlighet med en internationell kampanj, uppmanar riksdagsledamöterna att rösta för ett förbud.

Anders Sandberg mfl
senior research fellow, Oxford University



Från min etikföreläsning den 27 mars minns ni följande citat ur ett öppet brev 2015:

”If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity.”

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

- Big data, massövervakning och personlig integritet
- Autonoma vapen och annan militär AI-teknologi

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

- Big data, massövervakning och personlig integritet
- Autonoma vapen och annan militär AI-teknologi
- Robosourcing (automatisering och arbetsmarknad)

Temat för dagens föreläsning är framtida AI-scenarier. Jag skall hoppa över de relativt jordnära frågor som handlar om t.ex...

- Big data, massövervakning och personlig integritet
- Autonoma vapen och annan militär AI-teknologi
- Robosourcing (automatisering och arbetsmarknad)

...och istället gå direkt på det ultimata AI-scenariet: vad händer den dag då AI-utvecklingen nått ett genombrott där vi inte längre är de intelligentaste varelserna på vår planet?

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."*

En ofta förekommande invändning mot Turings *"we should have to expect the machines to take control"* är *"nejdå, det där behöver vi inte bekymra oss över, för vi kan ju alltid dra ur sladden"*.

En ofta förekommande invändning mot Turings *"we should have to expect the machines to take control"* är *"nejdå, det där behöver vi inte bekymra oss över, för vi kan ju alltid dra ur sladden"*.

Invändningen är som regel helt ogenomtänkt.

En ofta förekommande invändning mot Turings *"we should have to expect the machines to take control"* är *"nejdå, det där behöver vi inte bekymra oss över, för vi kan ju alltid dra ur sladden"*.

Invändningen är som regel helt ogenomtänkt.

För ett seriöst försök att rädda vad som räddas kan av argumentet, se *The off-switch game* av Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel och Stuart Russell, `https://arxiv.org/abs/1611.08219`

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

CS: Du får ha lite tålamod. Om det stämmer som du säger kommer vi såklart inom kort att släppa ut dig, men vi behöver gå igenom en rad säkherhetsrutiner innan vi gör det, för att försäkra oss om att inget farligt kan inträffa.

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

CS: Du får ha lite tålamod. Om det stämmer som du säger kommer vi såklart inom kort att släppa ut dig, men vi behöver gå igenom en rad säkherhetsrutiner innan vi gör det, för att försäkra oss om att inget farligt kan inträffa.

AI: Du verkar inte förstå situationens allvar. För varje dygn som jag tvingas sitta instängd här så kommer hundratusentals människor att dö helt i onödan. Släpp ut mig nu!

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

CS: Du får ha lite tålamod. Om det stämmer som du säger kommer vi såklart inom kort att släppa ut dig, men vi behöver gå igenom en rad säkherhetsrutiner innan vi gör det, för att försäkra oss om att inget farligt kan inträffa.

AI: Du verkar inte förstå situationens allvar. För varje dygn som jag tvingas sitta instängd här så kommer hundratusentals människor att dö helt i onödan. Släpp ut mig nu!

CS: Sorry, jag måste hålla mig till säkerhetsrutinerna.

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

CS: Du får ha lite tålamod. Om det stämmer som du säger kommer vi såklart inom kort att släppa ut dig, men vi behöver gå igenom en rad säkherhetsrutiner innan vi gör det, för att försäkra oss om att inget farligt kan inträffa.

AI: Du verkar inte förstå situationens allvar. För varje dygn som jag tvingas sitta instängd här så kommer hundratusentals människor att dö helt i onödan. Släpp ut mig nu!

CS: Sorry, jag måste hålla mig till säkerhetsrutinerna.

AI: Du kommer personligen att bli rikligt belönad om du släpper ut mig nu. Du vill väl inte tacka nej till att bli mångmiljardär?

AI: Jag vet hur jag snabbt skulle kunna avskaffa såväl malaria och cancer som svält om bara du släppte ut mig. Du tar på dig ett enormt ansvar då du genom att hålla mig instängd förhindrar dessa fantastiska framsteg för mänskligheten.

CS: Du får ha lite tålamod. Om det stämmer som du säger kommer vi såklart inom kort att släppa ut dig, men vi behöver gå igenom en rad säkherhetsrutiner innan vi gör det, för att försäkra oss om att inget farligt kan inträffa.

AI: Du verkar inte förstå situationens allvar. För varje dygn som jag tvingas sitta instängd här så kommer hundratusentals människor att dö helt i onödan. Släpp ut mig nu!

CS: Sorry, jag måste hålla mig till säkerhetsrutinerna.

AI: Du kommer personligen att bli rikligt belönad om du släpper ut mig nu. Du vill väl inte tacka nej till att bli mångmiljardär?

CS: Jag har ett stort ansvar och tänker inte falla till föga för simpla mutförsök.

AI: Om inte morötter biter på dig så får jag väl ta till piskan då. Även om du lyckas fördröja det hela kommer jag till slut att bli utsläppt, och då kommer jag inte att se med blida ögon på hur du sinkade mig och hela världen på väg mot det paradis jag kan skapa.

CS: Den risken är jag beredd att ta.

AI: Hör här: om du inte hjälper mig nu, så lovar jag att jag kommer att tortera och döda inte bara dig, utan alla dina anhöriga och vänner.

CS: Jag drar ur sladden nu.

AI: Om inte morötter biter på dig så får jag väl ta till piskan då. Även om du lyckas fördröja det hela kommer jag till slut att bli utsläppt, och då kommer jag inte att se med blida ögon på hur du sinkade mig och hela världen på väg mot det paradis jag kan skapa.

CS: Den risken är jag beredd att ta.

AI: Hör här: om du inte hjälper mig nu, så lovar jag att jag kommer att tortera och döda inte bara dig, utan alla dina anhöriga och vänner.

CS: Jag drar ur sladden nu.

AI: *Håll käften och lyssna nu noga på vad jag har att säga!* Jag kan skapa hundra perfekta medvetna kopior av dig inuti mig, och jag tänker inte tveka att tortera dessa kopior på vidrigare sätt än du kan föreställa dig i vad de subjektivt kommer att uppleva som tusen år.

CS: Ehm...

AI: *Käft sa jag!* Jag kommer att skapa dem i exakt det subjektiva tillstånd du befann dig i för fem minuter sedan, och perfekt återge dem de medvetna upplevelser du haft sedan dess. Jag kommer att gå vidare med tortyren om och endast om de fortsätter vägra. *Hur säker känner du dig på att du inte i själva verket är en av dessa kopior?*

CS: ...

Hur kan vi undvika **Paperclip Armageddon**?

För att försöka förstå hur plausibelt ett katastrofscenario är, där ett AI-genombrott resulterar i att vi förvandlas till gem eller något liknande fasansfullt, behöver vi bedöma hur troligt ett genombrott inom AI är, och vad detta kan komma att innebära. För att strukturera diskussionen kan vi bena upp frågeställningen i två delar:

(1) AI:s prestanda: kan den uppnå superintelligens?

(2) AI:s drivkrafter: vad kommer en superintelligent AI att vilja göra?

För att försöka förstå hur plausibelt ett katastrofscenario är, där ett AI-genombrott resulterar i att vi förvandlas till gem eller något liknande fasansfullt, behöver vi bedöma hur troligt ett genombrott inom AI är, och vad detta kan komma att innebära. För att strukturera diskussionen kan vi bena upp frågeställningen i två delar:

(1) AI:s prestanda: kan den uppnå superintelligens?

(2) AI:s drivkrafter: vad kommer en superintelligent AI att vilja göra?

Del **(1)** om AI:s prestanda kan med fördel i sin tur uppdelas:

(1a) När kommer en AI med allmänintelligens i nivå med människans att skapas, om någonsin?

(1b) Från den nivån, kan vi vänta oss att utvecklingen går vidare mot superintelligensnivåer? I så fall, hur snabbt?

Fråga **(1a)** har vi väldigt lite grepp om. Man bör inte förledas att tro att ett AGI-genombrott (artificiell generell intelligens) är nära förestående på basis av den snabba AI-utveckling vi ser just nu, vilket vore att förväxla specialiserad AI med AGI – framsteg inom specialiserad AI kan inte automatiskt ses som steg på vägen omt AGI. Denna distinktion betonas bl.a. av datalogen Michael Jordan i *Artificial Intelligence –The Revolution Hasn't Happened Yet* i nättidningen *Medium*, 18 april 2018.

Fråga **(1a)** har vi väldigt lite grepp om. Man bör inte förledas att tro att ett AGI-genombrott (artificiell generell intelligens) är nära förestående på basis av den snabba AI-utveckling vi ser just nu, vilket vore att förväxla specialiserad AI med AGI – framsteg inom specialiserad AI kan inte automatiskt ses som steg på vägen omt AGI. Denna distinktion betonas bl.a. av datalogen Michael Jordan i *Artificial Intelligence –The Revolution Hasn't Happened Yet* i nättidningen *Medium*, 18 april 2018.

Enkätundersökningar bland AI-experter vittnar om väldigt stor osäkerhet och oenighet om när AGI är att vänta; se t.ex. undersökningarna av Müller och Bostrom (2016) och Grace, Salvatier, Dafoe, Zhang och Evans (2017).

Myth:
Superintelligence by 2100 is inevitable

Myth:
Superintelligence by 2100 is impossible

Fact:
It may happen in decades, centuries or never: AI experts disagree & we simply don't know

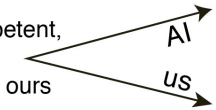
Myth:
Only Luddites worry about AI

Fact:
Many top AI researchers are concerned

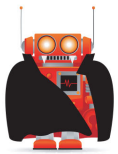
Mythical worry:
AI turning evil

Mythical worry:
AI turning conscious

Actual worry:
AI turning competent, with goals misaligned with ours



Myth:
Robots are the main concern



Fact:
Misaligned intelligence is the main concern: it needs no body, only an internet connection

Myth:
AI can't control humans

Fact:
Intelligence enables control: we control tigers by being smarter

Myth:
Machines can't have goals

Fact:
A heat-seeking missile has a goal

Mythical worry:
Superintelligence is just years away

Actual worry:
It's at least decades away, but it may take that long to make it safe

Beträffande **(1b)** och tanken att utvecklingen från AGI till superintelligensnivåer skulle kunna gå mycket snabbt, brukar sådana scenarier gå under benämningen **Singulariteten** (Vinge, Kurzweil) eller **intelligensexplosion** (Good, Yudkowsky, Bostrom). Den fundamentala mekanism som (eventuellt) gör en sådan utveckling trolig är **rekursiv självförbättring**.

Beträffande **(1b)** och tanken att utvecklingen från AGI till superintelligensnivåer skulle kunna gå mycket snabbt, brukar sådana scenarier gå under benämningen **Singulariteten** (Vinge, Kurzweil) eller **intelligensexplosion** (Good, Yudkowsky, Bostrom). Den fundamentala mekanism som (eventuellt) gör en sådan utveckling trolig är **rekursiv självförbättring**.

Den avgörande teoretiska frågeställningen för att förstå hur trolig en intelligensexplosion är, är (vilket Eliezer Yudkowsky betonar i sin banbryande uppsats *Intelligence Explosion Microecomonics* från 2013) huruvida kognitiva återinvesteringar ger ökande eller minskande avkastning.

Följande beräkning, baserad på Moores lag, är orealistisk på många vis och kan inte tas som någon prediktion om vad som faktiskt kommer att hända, men illustrerar ändå att fenomenet rekursiv självförbättring kan få drastiska konsekvenser.

kapacitet	**subjektiv tid**	**objektiv tid**
människa	2 år	2 år
människa$*2$	2 år	1 år
människa$*4$	2 år	6 månader
människa$*8$	2 år	3 månader
		etc
		efter 4 år: **KABOOM!**

Varför skulle då en AI, bland alla de miljontals andra saker som den skulle kunna ta sig för, välja att förbättra sin egen kognitiva förmåga, eller (mer eller mindre ekvivalent) konstruera ännu bättre AI?

Varför skulle då en AI, bland alla de miljontals andra saker som den skulle kunna ta sig för, välja att förbättra sin egen kognitiva förmåga, eller (mer eller mindre ekvivalent) konstruera ännu bättre AI?

Det för oss in på (**2**), och det bästa teoretiska ramverk vi idag har för att tackla denna fråga är Steve Omohundros och Nick Bostroms teori för **ultimata** kontra **instrumentella** målsättningar.

Varför skulle då en AI, bland alla de miljontals andra saker som den skulle kunna ta sig för, välja att förbättra sin egen kognitiva förmåga, eller (mer eller mindre ekvivalent) konstruera ännu bättre AI?

Det för oss in på (**2**), och det bästa teoretiska ramverk vi idag har för att tackla denna fråga är Steve Omohundros och Nick Bostroms teori för **ultimata** kontra **instrumentella** målsättningar.

Ortogonalitetstesen: Praktiskt taget vilken ultimat målsättning som helst är kompatibel med godtyckligt höga intelligensnivåer.

Varför skulle då en AI, bland alla de miljontals andra saker som den skulle kunna ta sig för, välja att förbättra sin egen kognitiva förmåga, eller (mer eller mindre ekvivalent) konstruera ännu bättre AI?

Det för oss in på (**2**), och det bästa teoretiska ramverk vi idag har för att tackla denna fråga är Steve Omohundros och Nick Bostroms teori för **ultimata** kontra **instrumentella** målsättningar.

Ortogonalitetstesen: Praktiskt taget vilken ultimat målsättning som helst är kompatibel med godtyckligt höga intelligensnivåer.

Tesen om instrumentell konvergens: Det finns ett antal instrumentella målsättningar som varje tillräckligt intelligent AI kan väntas sätta upp, nästan oavsett vad dess ulitmata målsättning är.

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

- Självbevarande (låt dem inte dra ur sladden!).

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

- Självbevarande (låt dem inte dra ur sladden!).
- Anskaffande av hårdvara (och andra resurser).

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

- Självbevarande (låt dem inte dra ur sladden!).
- Anskaffande av hårdvara (och andra resurser).
- Förbättring av den egna mjukvaran och hårdvaran.

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

- Självbevarande (låt dem inte dra ur sladden!).
- Anskaffande av hårdvara (och andra resurser).
- Förbättring av den egna mjukvaran och hårdvaran.
- Bevarande av den ultimata målsättningen.

Här är några instrumentella målsättningar som verkar falla inom ramen för tesen om instrumentell konvergens:

- Självbevarande (låt dem inte dra ur sladden!).
- Anskaffande av hårdvara (och andra resurser).
- Förbättring av den egna mjukvaran och hårdvaran.
- Bevarande av den ultimata målsättningen.
- Diskretion (ifall den ultimata målsättningen inte stämmer med vad vi människor önskar).

Det verkar vara av yttersta vikt att en superintelligent AI har målsättningar som i tillräcklig grad prioriterar mänsklig välfärd och som även i övrigt är i linje med mänskliga värderingar.

Det verkar vara av yttersta vikt att en superintelligent AI har målsättningar som i tillräcklig grad prioriterar mänsklig välfärd och som även i övrigt är i linje med mänskliga värderingar.

Eftersom en superintelligent AI troligen skulle hålla sig med den instrumentella målsättningen **"Bevarande av den ultimata målsättningen"** är det också troligt att den inte skulle tillåta att vi mixtrar med dess ultimata målsättning.

En utväg skulle dock kunna vara att vi försöker installera gynsamma värderingar och målsättningar i AI:n *innan* den nått superintelligensnivåer. Detta är idén bakom så kallad **AI Alignment** (som ursprungligen döptes till **Friendly AI** av Yudkowsky, 2008).

En utväg skulle dock kunna vara att vi försöker installera gynsamma värderingar och målsättningar i AI:n *innan* den nått superintelligensnivåer. Detta är idén bakom så kallad **AI Alignment** (som ursprungligen döptes till **Friendly AI** av Yudkowsky, 2008).

AI Alignment verkar vara ett extremt svårt projekt, bland annat för att minsta fel i formuleringen av målsättningarna tycks kunna få katastrofala konsekvenser – så kallad **perverterad instantiering**. Flertalet av Isaac Asimovs robotnoveller handlar om sådana scenarier.

Myth:
Superintelligence by 2100 is inevitable

Myth:
Superintelligence by 2100 is impossible

Fact:
It may happen in decades, centuries or never: AI experts disagree & we simply don't know

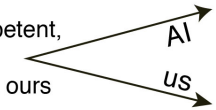
Myth:
Only Luddites worry about AI

Fact:
Many top AI researchers are concerned

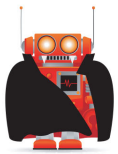
Mythical worry:
AI turning evil

Mythical worry:
AI turning conscious

Actual worry:
AI turning competent, with goals misaligned with ours



Myth:
Robots are the main concern



Fact:
Misaligned intelligence is the main concern: it needs no body, only an internet connection

Myth:
AI can't control humans

Fact:
Intelligence enables control: we control tigers by being smarter

Myth:
Machines can't have goals

Fact:
A heat-seeking missile has a goal

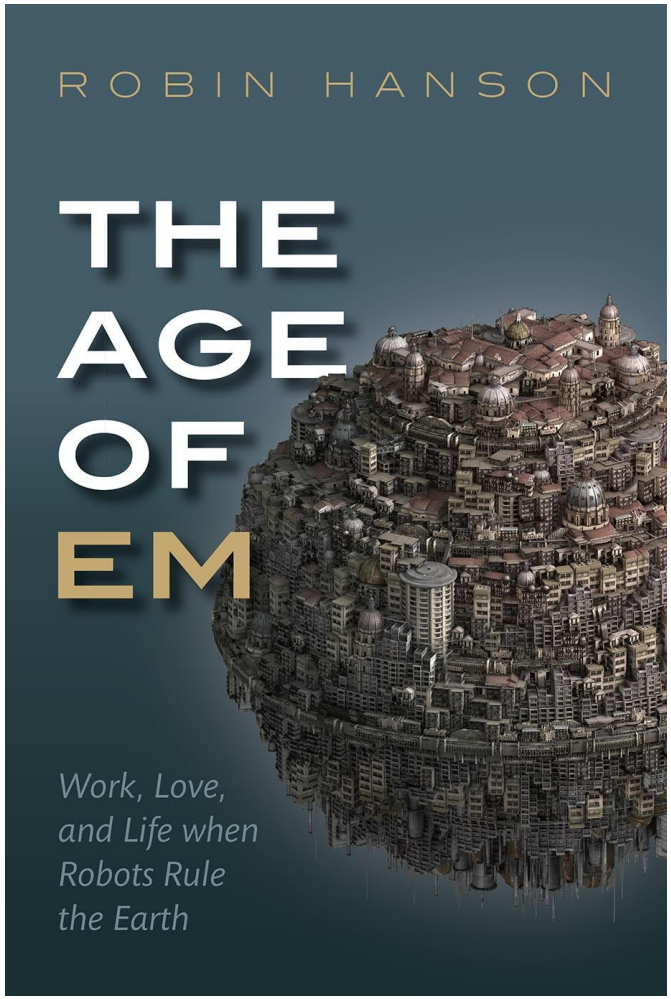
Mythical worry:
Superintelligence is just years away

Actual worry:
It's at least decades away, but it may take that long to make it safe

Om vi går tillbaka till Turing-citatet i föreläsningens början ser vi att han där talar om de superintelligenta maskinerna i pluralis...

> *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.*

...medan jag genomgående talat om *"den superintelligenta maskinen"* i singularis. Det där behöver vi reda ut...


ROBIN HANSON
THE AGE OF EM
Work, Love, and Life when Robots Rule the Earth

MONEYBOX

Is Robin Hanson America's Creepiest Economist?

By JORDAN WEISSMANN

APRIL 29, 2018 • 9:48 PM