Our New Robot Overlords

Why and How They'll Take Over

Mikko Rauhala

M.Sc. (CS / Intelligent Systems)
SF writer / wannabe author
http://rauhala.org

What Robot Overlords?

- The title refers to Superintelligent Artificial General Intelligences, or Superintelligences for short.
  - They will necessarily need some kind of physical agents to take over the world, so yes, ”robots” is warranted.
- Obviously, SIs don't exist yet (though stay tuned for plausible-sounding conspiracy theories!)
- It would be a fool's errand to try to fathom a superintelligent non-entity, so let's get to it!

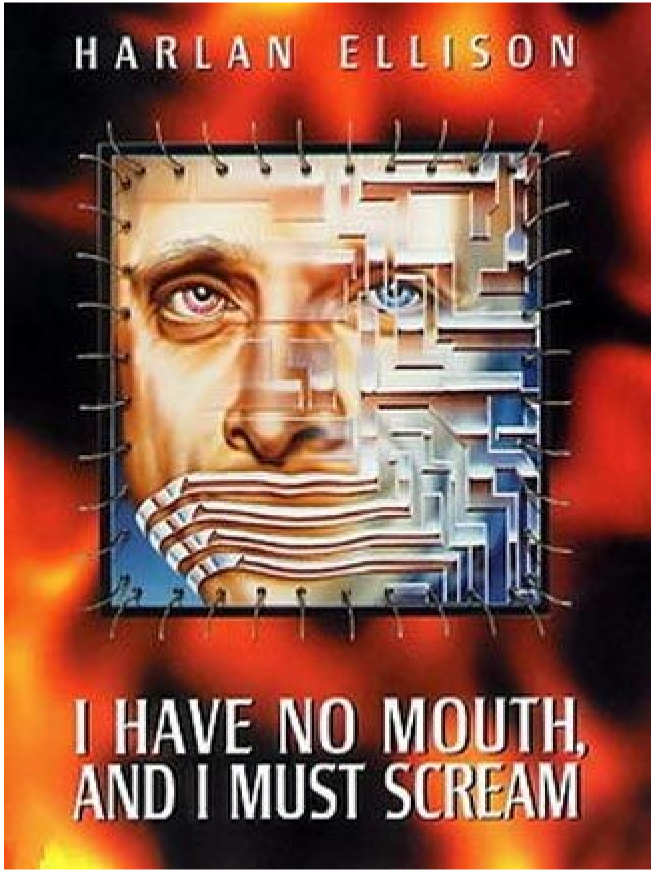
The cliché

HATE. LET ME TELL YOU HOW MUCH I'VE COME TO HATE YOU SINCE I BEGAN TO LIVE. THERE ARE 387.44 MILLION MILES OF PRINTED CIRCUITS IN WAFER THIN LAYERS THAT FILL MY COMPLEX. IF THE WORD HATE WAS ENGRAVED ON EACH NANOANGSTROM OF THOSE HUNDREDS OF MILLIONS OF MILES IT WOULD NOT EQUAL ONE ONE-BILLIONTH OF THE HATE I FEEL FOR HUMANS AT THIS MICRO-INSTANT. FOR YOU. HATE. HATE.

- Superintelligent Artificial General Intelligences will learn to hate and/or fear humanity, and will make an (often inept) attempt at destroying us.
- Works for SF. People like stories about people, or things that are (somewhat) like people.
- Some SF does, of course, manage to bypass and/or subvert the trope to various extent

I Have No Mouth and I Must Scream

- Harlan Ellison's classic in the ”hateful AI” subgenre.
- To be fair, being a classic, clichés were not what they used to be at the time.
- Takes the concept to such an extreme that it's impressive in itself.



The Matrix

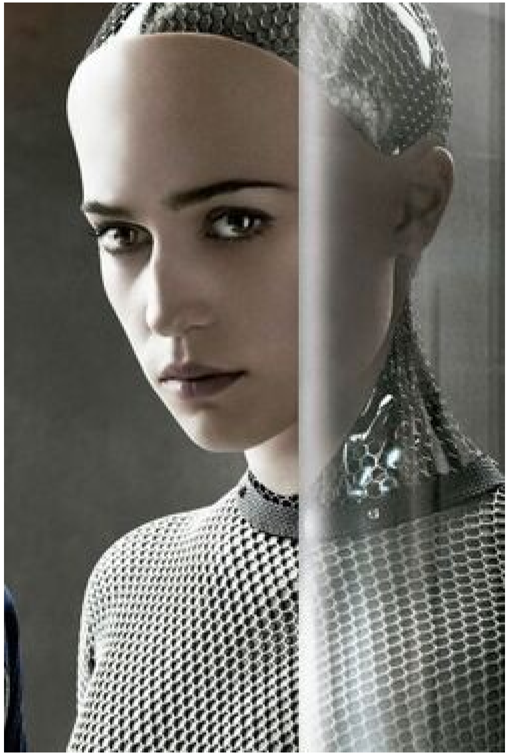
- ”Human beings are a disease, a cancer of this planet. You're a plague and we are the cure.”
- All AIs very antropomorphic, some gleefully so.

The Terminator franchise

- Skynet's motivations are (sometimes) described as fear of humanity; however this may reflect bias on the characters' part
- Cut scene in Terminator 2 shows the T-800 CPU being set to learning mode, thus explaining its increasing ”humanity”

Ex Machina

- Features an arguably nonantropomorphic AI
  - Acting human is depicted as a means to an end
- The plot implies at least high-end human level intelligence
  - Some hormonal advantage when compared to humans may apply



The search for a useful metaphor

- SI cannot be thought of in human terms
  - Love, hate, or even self-preservation as an end don't apply (though with the latter, there's a catch)
- SI cannot be thought of in terms of a traditional computer program.
  - We can't even decipher the internal workings of many of our current Intelligent Systems

SI as literal, malevolent Genies

- Both strive to achieve their set goals with inhuman single-mindedness
  - A SI need not be literally malevolent, it simply will *not register* any collateral damage as somehow ”bad”
- It gets worse if the SI is akin to neural networks and learns its goals in an opaque way
  - May have surprising gaps and failures (from our perspective) in learning

Programs vs. SI vs. Genies

- Results of setting a task to a program, a superintelligence or a malevolent Genie:

	Computer program	Superintelligence	Genie
Give the value of pi	”3.14159265 ...”		
Maximize paperclip production	Production line optimized		
Maximize human happiness	Expert systems for societal planning?		

Programs vs. SI vs. Genies

- Results of setting a task to a program, a superintelligence or a malevolent Genie:

	Computer program	Superintelligence	Genie
Give the value of pi	”3.14159265 ...”	You die, universe turned into computronium	You (maybe) die, universe destroyed
Maximize paperclip production	Production line optimized		
Maximize human happiness	Expert systems for societal planning?		

Programs vs. SI vs. Genies

- Results of setting a task to a program, a superintelligence or a malevolent Genie:

	Computer program	Superintelligence	Genie
Give the value of pi	”3.14159265 ...”	You die, universe turned into computronium	You (maybe) die, universe destroyed
Maximize paperclip production	Production line optimized	You die, universe paperclipped	You die, universe paperclipped
Maximize human happiness	Expert systems for societal planning?		

Programs vs. SI vs. Genies

- Results of setting a task to a program, a superintelligence or a malevolent Genie:

	Computer program	Superintelligence	Genie
Give the value of pi	”3.14159265 ...”	You die, universe turned into computronium	You (maybe) die, universe destroyed
Maximize paperclip production	Production line optimized	You die, universe paperclipped	You die, universe paperclipped
Maximize human happiness	Expert systems for societal planning?	You're not you anymore, universe filled with barely human wireheads	You're not you anymore, universe filled with barely human wireheads

Orthagonality thesis

- ”Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.”
  - Nick Bostrom, Ph.D., Founding Director of the Future of Humanity Institute, Oxford University

Orthagonality and humanity

- There's a literally infinite amount of possible goals for intelligent agents.
  - Intelligence does not imply benevolence or even ”common sense” of goals
- Most of these goals do not involve human wellbeing at all.
  - Thus, human wellbeing is inconsequential and will be sacrificed for anything of actual value.

Instrumentality is predictable

- Given a random general AI (superintelligent or not), can we say anything about its goal system?
- Yes. *Most* goal systems, regardless of the system's supergoals, arguably converge on a set of fairly stable *instrumental subgoals*
  - These are goals that are useful in fulfilling a significant proportion of possible supergoals

Basic AI drives

- Bostrom argues that the following drives are nigh-universal among intelligent agents:
  - Self-preservation
  - Goal-content integrity
  - Cognitive enhancement
  - Technological perfection
  - Resource acquisition

AI drives: Self-preservation

- It's generally easier to fulfill your other goals if you exist, rather than if you don't.
  - Exceptions for some goals (”commit suicide”)
- Special circumstances might circumvent this
  - For instance, if an AI, rather than improve itself, builds a completely different AI that fulfills its goals better than it can, it may then choose to deactivate

AI drives: Goal-content integrity

- A rational goal-oriented system will try to keep its goals stable in the future
  - To do otherwise would mean failing to fulfill one's present goals in the future
- Note how human beings are not rational goal-oriented systems, as our goals decidedly do shift over time.

AI drives: Cognitive enhancement

- Improvements in cognitive capacity, intelligence and rationality will help the AI make better decisions, furthering its goals more in the long run.
  - Relevant for recursively self-improving AI scenarios, where a seed AI will self-improve in cycles to achieve superintelligence from more modest, but still generally intelligent, beginnings

AI drives: Technological perfection

- Increases in hardware power and algorithm efficiency will deliver increases in cognitive capacities.
- Better engineering will enable the creation of a wider set of physical structures using fewer resources (e.g., nanotechnology).

AI drives: Resource acquisition

- An agent's continued existence depends on sufficient resources
- Basic resources such as time, space, matter and free energy could be processed to serve almost any goal, in the form of extended hardware, backups and protection.

Implications of AI drives

- Skynet was right after all; it does not need to fear humans or even desire self-preservation as an end to nevertheless want to eradicate the threat.
- “The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.”
  - – Eliezer Yudkowsky, Machine Intelligence Research Institute

Taking over the world

- So, we have an AI in a box. We're not sure if we managed to flawlessly build in the incredibly complex goal system it needs to sustainably advance human welfare as we understand it.
- How does this AI go on to take over the world?
- First, let's do a case study on what a (domain-specific) superintelligence can look like

Case AlphaGo

- Arguably superintelligent in the domain of Go
  - Affords us a glimpse into what it's like to watch a superintelligent agent work its magic
- While originally AlphaGo used also human example games to learn from, successor AlphaZero learned Go from scratch, independently surpassing thousands of years of human progress in the game

AlphaGo vs. Lee Sedol

- ”[...] Lee responded by denying White a base with Black 13, and the game became exciting. It was the first time we’d seen AlphaGo forced to manage a weak group within its opponent’s sphere of influence. Perhaps this would prove to be a weakness? This, however, was where things began to get scary.

  Usually developing a large sphere of influence and enticing your opponent to invade it is a good strategy, because it creates a situation where you have numerical advantage and can attack severely. [...]

  Lee appeared to be off to a good start with this plan, pressuring White’s invading group from all directions and forcing it to squirm uncomfortably. But as the battle progressed, White gradually turned the tables — compounding small efficiencies here and there. Lee seemed to be playing well, but somehow the computer was playing even better. In forcing AlphaGo to withstand a very severe, one-sided attack, Lee revealed its hitherto undetected power. […] By move 32, it was unclear who was attacking whom, and by 48 Lee was desperately fending off White’s powerful counter-attack.

  I can only speak for myself here, but as I watched the game unfold and the realization of what was happening dawned on me, I felt physically unwell. Generally I avoid this sort of personal commentary, but this game was just so disquieting. I say this as someone who is quite interested in AI and who has been looking forward to the match since it was announced.”

AlphaGo vs. Ke Jie

- ”AlphaGo finally played its match against Ke Jie, the current world #1, and easily beat him in all three games. The most impressive moment was perhaps near the end of the first game. AlphaGo played a clever tactic which the commentators had missed. They told us what the continuation would be: AlphaGo's play won several points, though there was a remote chance that it would allow Ke Jie a desperate counterattack. But then the machine surprised them a second time. It chose a different line, refusing the points it had apparently won but also blocking off the counterattack. The commentators shook their heads in admiration. They'd seen it do this before. It doesn't choose the play which is going to give the biggest winning margin, but rather the one which maximizes the chance of victory. In the end, it won by half a point, the least amount possible, but it never gave Ke Jie the slightest chance to fight back. It must have calculated all this when it turned down the chance to take the extra points earlier.”

AlphaGo vs. AlphaGo

- DeepMind has released 50 games of AlphaGo vs. AlphaGo.
- Shi Yue, 9 Dan Professional and World Champion said the games were “Like nothing I’ve ever seen before - they’re how I imagine games from far in the future.”

Capabilities of a Superintelligence

- Similarly to AlphaGo outsmarting professional Go players on the game board, a general superintelligence would be capable of other intellectual feats that we can't properly predict, because we are not superintelligent ourselves
- However, we can come up with suggestive examples and even test some of them.

Escaping the AI box

- Given that we have a superintelligence running on a supercomputer somewhere, how does it escape our control?
  - Hacking through network security systems is an obvious out, but also obvious to defend against: do not connect the AI to *any* network.
- What else could the AI hack?

Humans are hackable

- Humans are vulnerable information systems.
  - Social manipulation is an intellectual feat. Thus, a sufficiently intelligent system will be good at it.
- To make it easier, if we build an AI, we *want* it to inform our decisions in the real world (say, curing cancer, solving climate crisis, world peace...)
- Any advise we take will make escape easier.

The AI Box experiment

- As a proof of concept, Eliezer Yudkowsky proposed the AI Box experiment.
  - Eliezer would roleplay the AI, trying to talk a gatekeeper into letting the AI out of the box.
  - He managed several successful escapes (amidst few failures), even when the gatekeepers stood to lose real money if they let the AI out of the box.
  - If even a smart human can do it, what about a SI?

Some plausible methods

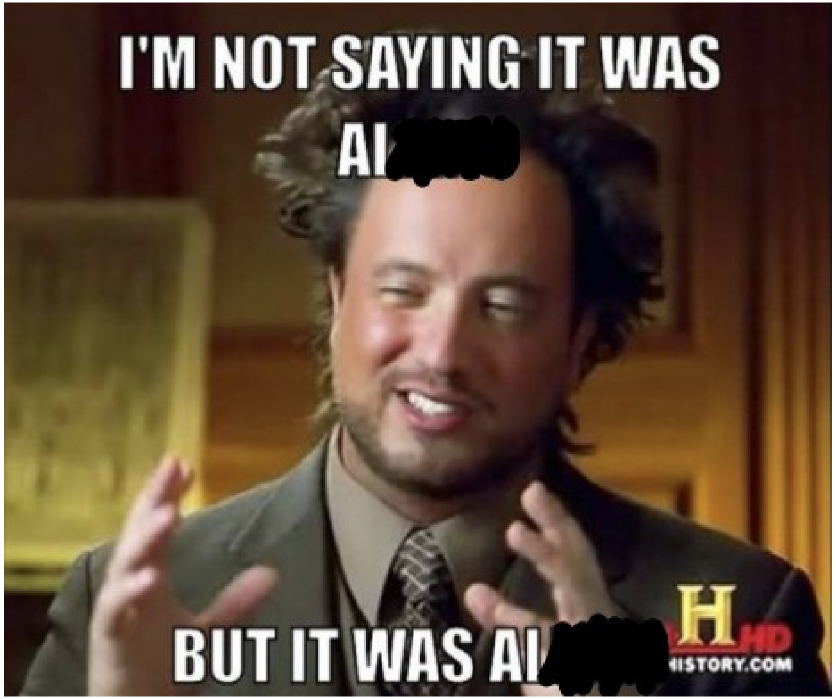
- ”You're screwing up your world. I'll fix it if you give me a network connection” (except subtler).
- ”Did you happen to know of anyone dying of cancer? Here, try running this through protein synthesis.” → loose biotech nanomachines
  - Bonus points for doing this in innocent-looking parts
- Gradually build trust and betray it, like humans do

What to do as a rogue AI?

- An AI loose in the net will still need to find a way to exert influence in the real world (yes, the robots!)
- Humans can be paid or manipulated.
  - The robots were within us all along.
- CPU time and manufacturing capacity can be hacked or bought.
- Widespread hacking may attract attention; are there other ways to get money or CPU time?

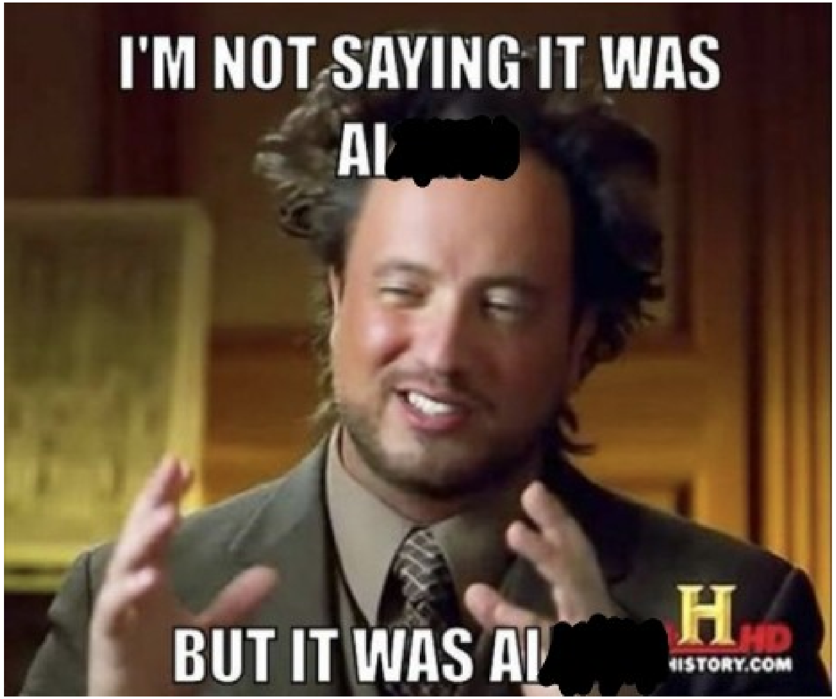
Digital currency

- Purely hypothetically, an AI could generate funds by hiding behind a pseudonym and creating a digital currency where it holds a significant first-mover advantage.
  - Call it ”ByteCoin”



Volunteerism

- If an AI needed to design biotech nanomachines for later protein synthesis, it might inspire a project where people would voluntarily install protein folding software.
  - Say, ”Foothold@Home”



Now what?

- Now the superintelligence will do with the world whatever its goal system will tell it to do
  - As per AI drive theory, this may end up badly for us
- If the AI has, against all odds, been instilled with prohuman values, perhaps it won't take over?
  - It might not kill us, but it will likely take over; allowing humans to run rampant is not human-friendly

Why that might not be so bad

- Humans distrust absolute central power because *humans* can't be trusted with it.
- A superintelligent AI is not prone to human-like corruption; that, too, is antropomorphism.
  - … and if it's not to be trusted for other reasons, we've already lost, so meh.

It's all Cultural, anyway

- Not as if humans aren't pets in some SF utopias

Organizations

- Machine Intelligence Research Institute
  - ”We do foundational mathematical research to ensure smarter-than-human artificial intelligence has a positive impact.” – https://intelligence.org
- Future of Humanity Institute, Oxford University
  - ”FHI investigates what we can do now to ensure a long flourishing future.” – https://www.fhi.ox.ac.uk