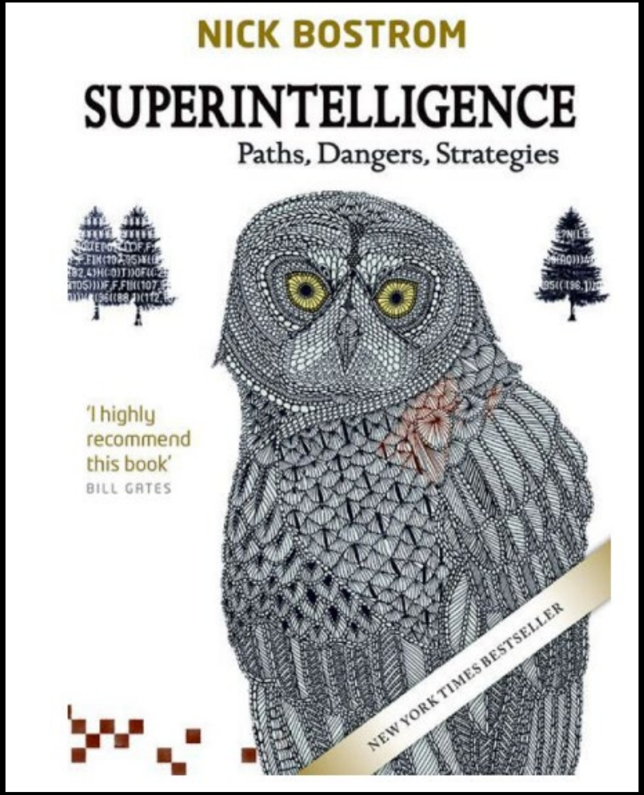
Will there be Superintelligence and will it hate us?


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies
'I highly recommend this book'
BILL GATES
NEW YORK TIMES BESTSELLER

Yorick Wilks
Gresham College
and
Florida Institute of Human and Machine Cognition
May 14th, 2019.

People are increasingly worried about AI

- A new poll of 1,004 registered voters in the U.S. found that 57 percent of them believe AI is a 'threat to the human race.’ (Daily Mail 3.5.19)

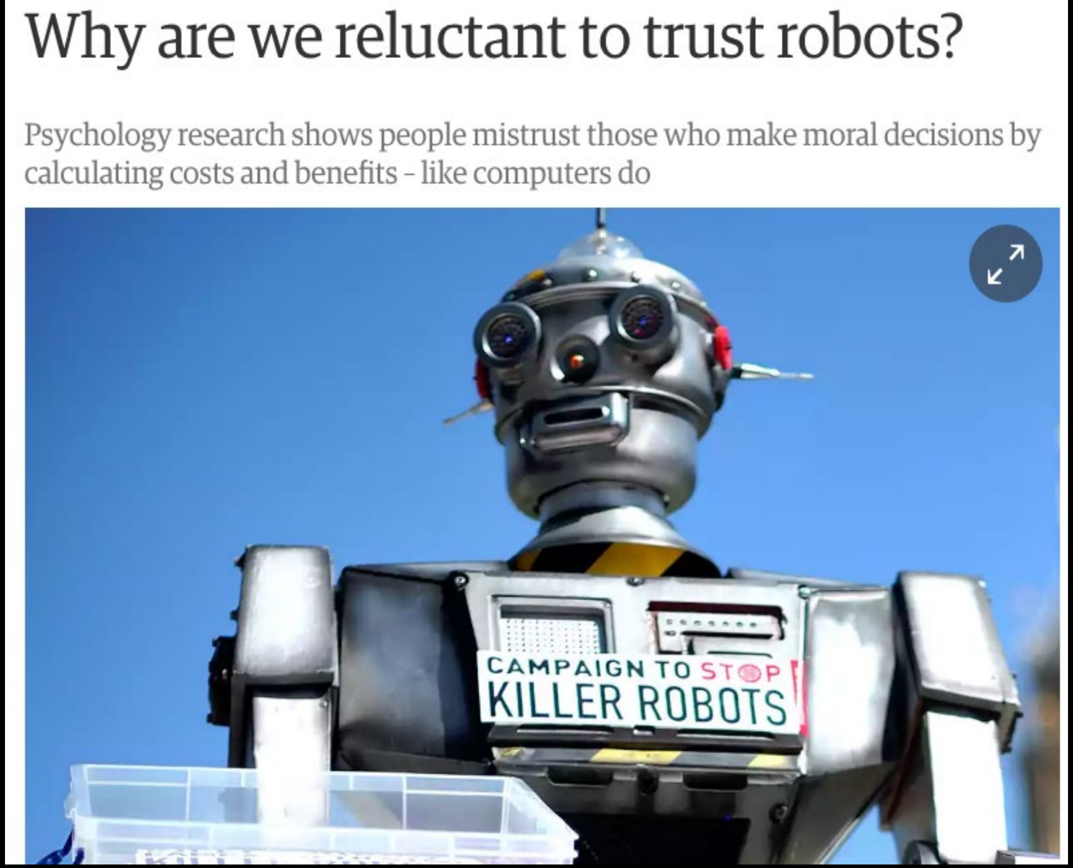
- *Are they right?*

- I shall argue that they are not, and that the benefits of AI to the human race hugely outweigh any possible harm. But………

Plan of the talk

- We shall examine arguments that AI might *itself* develop the capacity to harm or even destroy us all.
- And also mention the possibility that someone might design Malevolent AI (MAI) to that end.
- How can we control artificial intelligences?
- Perhaps by making them *ethical*?
- Let us begin with ethics…………………

We seem to prefer those who work with moral rules rather than calculating consequences



Why are we reluctant to trust robots?

Psychology research shows people mistrust those who make moral decisions by calculating costs and benefits - like computers do

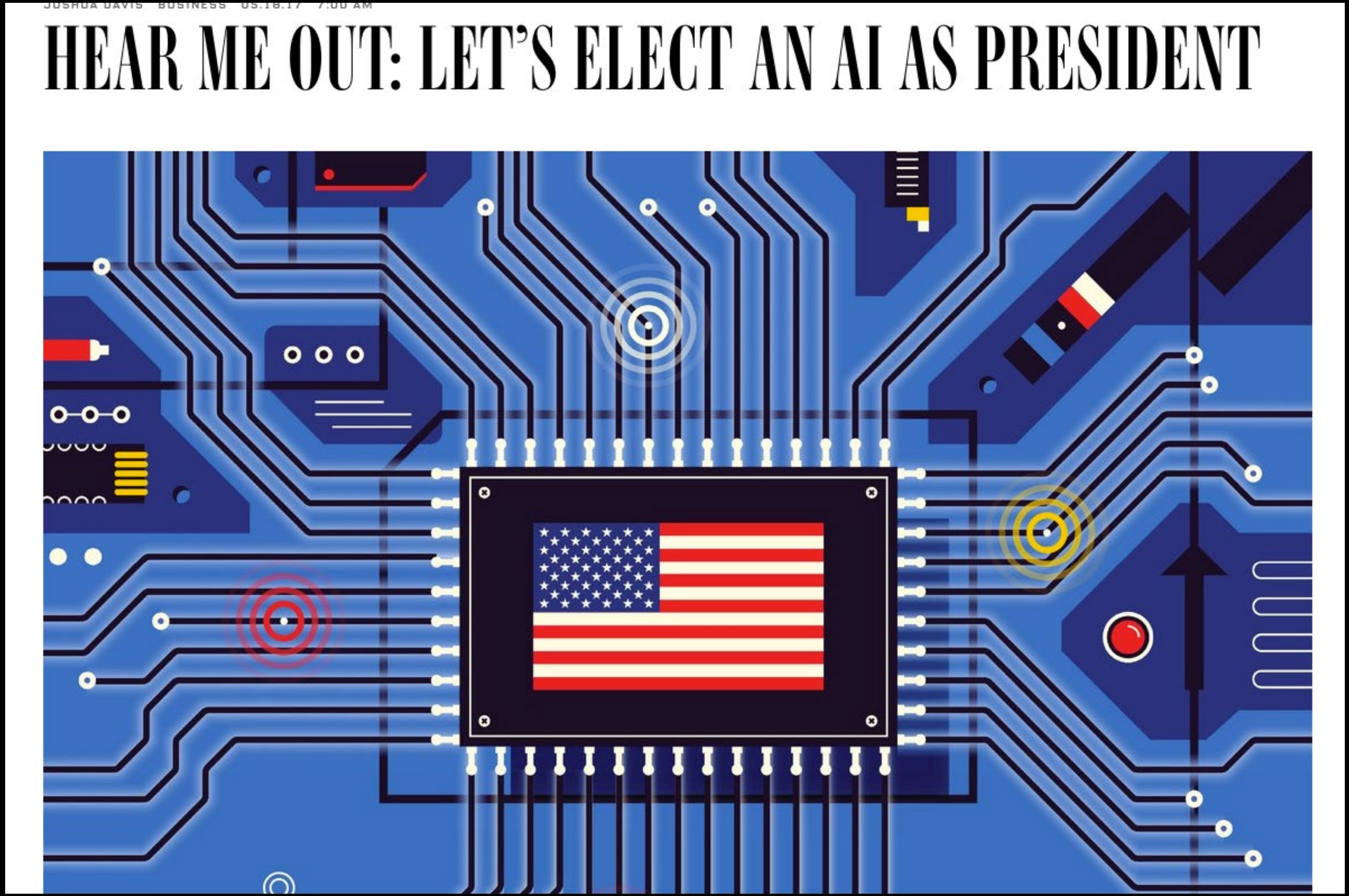
Molly Crockett (et al, of the Oxford Martin School) argued this based on surveys

- We tend to find people more trustworthy who work out consequences carefully but who refer to moral rules (eg “Thou shalt not kill”)
- In terms of classical moral theory this tells on the side of Kant, and even Hume who thought emotion not reason the basis of morals.
- Rather than the Utilitarians (or consequentialists as they’re called these days) who worked out effects.
- BUT computers don’t have to be Utilitarians, they could be Kantians if we want them to be.

What follows from this?

- Although humans may wish to deal with (humans and) Artificial Intelligences having a certain kind of ethics, what type of ethics machines have is up to us.
- It is a fallacy to assume all automata must be consequentialists, as automated cars are going to be (so they say).
- SUPERINTELLIGENCE doctrine claims we do not have that choice, and they will be malign and destructive (of us) consequentialists with their own goals.

Another budget of fallacies in the same vein in a recent WIRED:



JOSHUA DAVIS BUSINESS 05.18.17 7:00 AM

HEAR ME OUT: LET'S ELECT AN AI AS PRESIDENT

An AI President and Machine Learning, and three incompatible positions.

- WIRED first assumes a machine-learned decision maker is both random and inscrutable (the latter may be true, but so are some humans!!). RANDOM BEHAVIOUR.
- Yet, “So, like all presidents, the AI leader would seek to maximize the satisfaction of a majority of voters within the confines of the law.” IS A CONSEQUENTIALIST
- “Tesla CEO Elon Musk uses the example of an AI built to grow and harvest strawberries. To fulfill its primary mission of picking as many strawberries as possible, an AI might conclude it would be reasonable to wipe out humanity and turn the planet into a huge strawberry farm. ” GOOD GOAL, CRAZY METHODS.

What is Bostrom’s ***Superintelligence*** (SI)?

- “intellects that greatly outperform the best current human minds across many very general cognitive domains”.
- Gates, Musk, Hawking et al. all take Bostrom’s thesis seriously.
- Early role of Kurzweil and Yudkowsky in creating the “singularity” notion on which the claim rests: AI will trigger huge changes in human history at an identifiable point in time.
- Generality of intellect is essential because we already have lots of specialist machines like:
  - Arithmetic Calculation
  - Chess and Go machine champion
  - Landing planes in fog

The idea goes back at least to IJ Good (1965)

- “Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”

Bostrom suggests three ways to SI

- By speed
  - “to a very fast mind ‘events in the external world would appear to unfold in slow motion’” (Mind-body; HG Wells *New Accelerator*)
  - Speed is an uninteresting trade-off in computing
- By collectivity
  - Space shuttle cooperation smarter than the pyramid teams
  - **But**: Nobel prizes and small nations!
- By quality
  - Increasing IQ and the “Flynn effect”, but do Newtons/Einsteins really become more common with increasing populations?

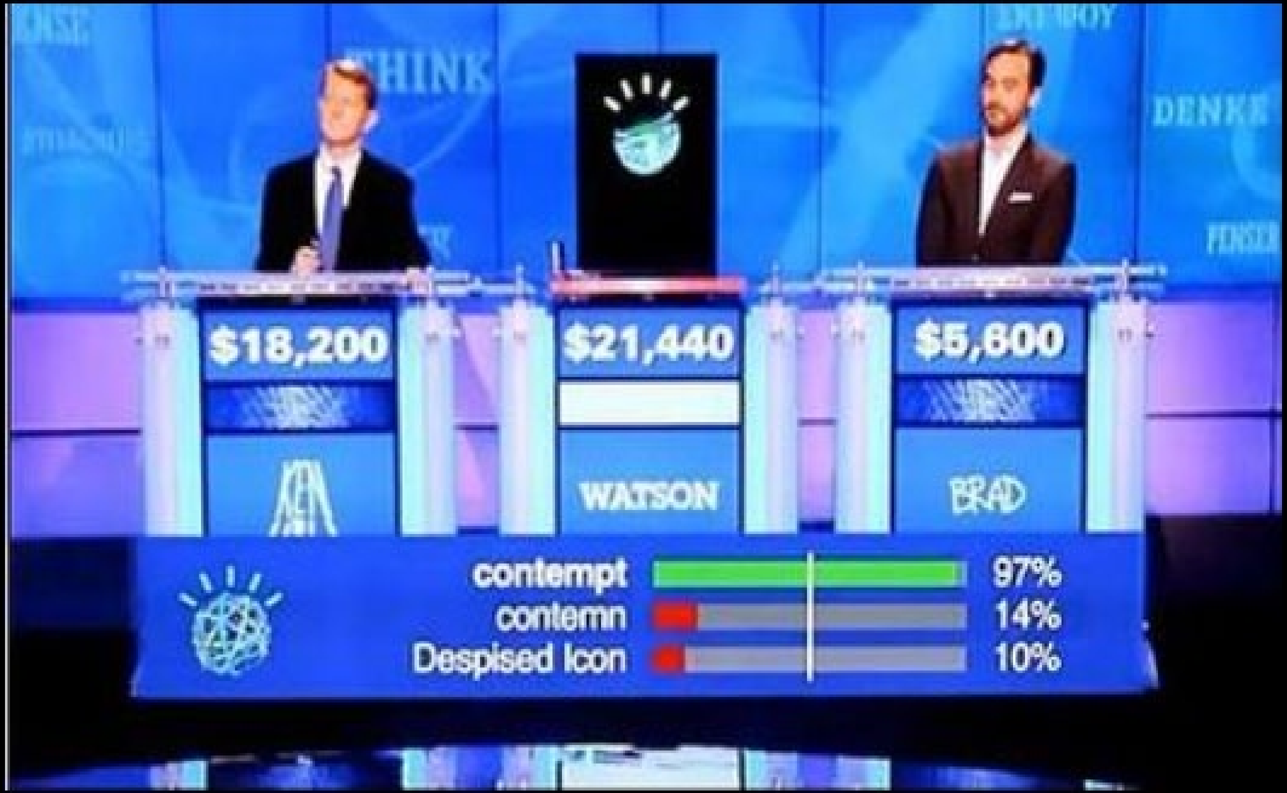
Bostrom’s equation for the arrival of SI

- $dI/dt = O/R$
- which gives us the rate of change of Intelligence (on the left) as the optimization power O, divided by the “*recalcitrance*” R.
- *Decorative mathematics* and Ohm’s law
- $I = V/R$
- where Current = Voltage/Resistance
- “Recalcitrance starts out high while the only available method [..for intelligence enhancement..] is selective breeding sustained over many generations”
- Why is recalcitrance high in human breeding beyond the social impediments? This is a social science issue, not a philosophical or AI-technical one.

Does it depend on **how much** the SI knows?

- Bostrom cites WATSON at IBM as a source of large-scale knowledge for an SI
- But that is mere state of the art, WATSON is just text, and has no representation of knowledge at all.
- → Detachment of this SI exploration from AI and its development
- BUT "If and when a takeoff occurs, it will likely be explosive"
- → why, and how could he possibly know that?

IBM’s WATSON plays *Jeopardy*



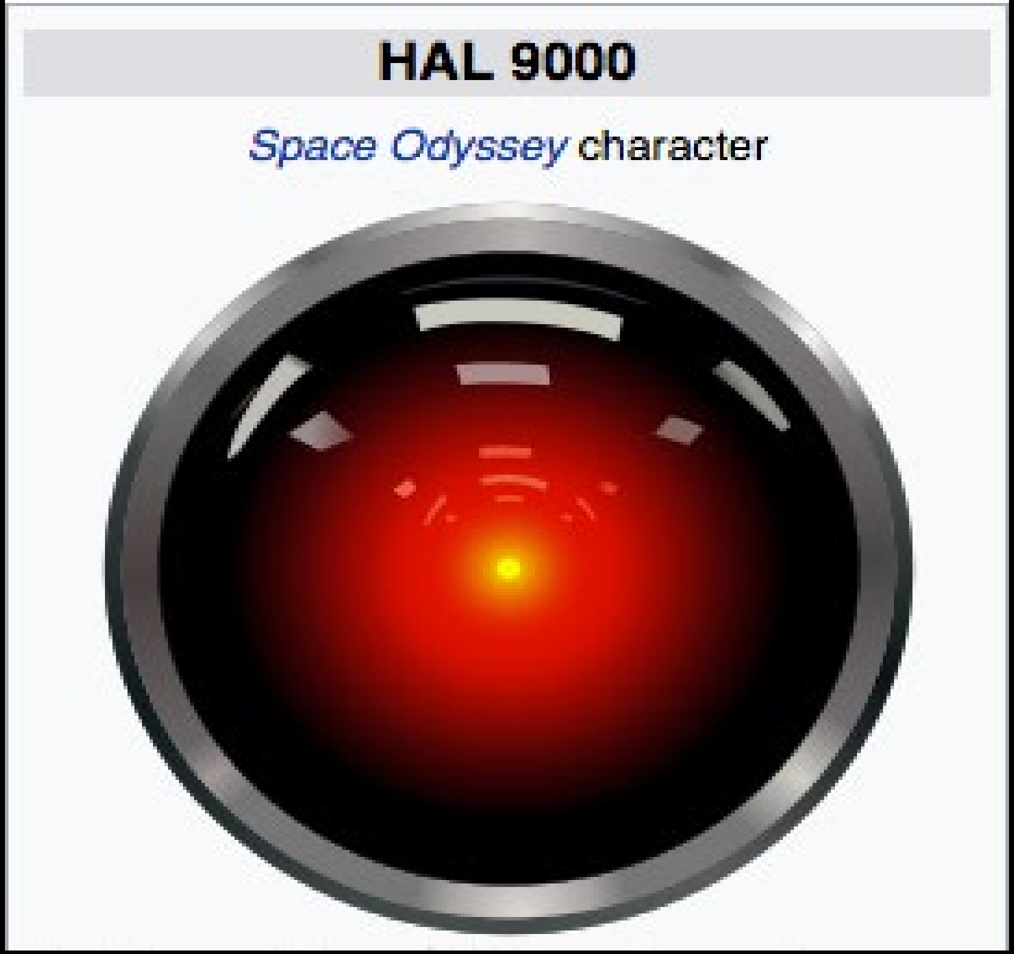
How intelligent will an SI be?

- For Bostrom an SI will want to “send out von Neumann probes into the galaxy”
- Unlike us, SI’s would have *permanent* goals (an unevolutionary view?).
- Also an SI would be a *singleton* (compete/destroy each other, not cooperate or even reproduce! Anti-evolution again!!)
- How different from our own (cooperative) home life!
- Contrast with Papert’s classic view of AI’s goals – the superhuman fallacy.

The “existential catastrophe”

- Bostrom’s argument for the existential catastrophe requires:
  - The strategic advantage of the first mover (the first, single SI)
  - The “orthogonality” of the SI’s final goals: that they may be quite unrelated to ours
  - “Instrumental convergence” by which he means that there is a range of sub-goals that almost any intelligence will pursue, e.g. self-preservation, resource acquisition etc
  - Deceptively good behaviour by an artificial intelligence “in a sandbox” which may be no guide to the future behaviour of the real SI.
- “Decisive strategic advantage” is a business studies approach
- The sorcerer’s apprentice, the paper-clip maker (and the strawberries!!)

The influence of SciFi on Bostrom ---girls in flimsy dresses-----and HAL9000



HAL 9000

Space Odyssey character

It is much harder now for the public to distinguish science-fiction from science fact.

- Beliefs about machine translation in the 1980s: impossible, but also done and dusted (it wasn’t).
- Contradictory claims right now about the arrival of automated cars.
- Much science fiction does become true, some does not and few foresaw the internet and web.
- Bostrom is strongly influenced by fears of malevolent science fiction.

Bostrom’s “value-loading” problem

- “If we cannot transfer human values into an AI by typing out full-blown representations in computer code, what else might we try?”
- The obvious overall answer is to instil values we approve of in the SI, values that will be favourable to us as humans and not destroy or enslave us.
- ***“How can we get an SI to do what we want?”***
- This is the key (unanswered) question in the work, and it’s the very same question Asimov sought to answer.
- Again, it could be Kantian, and doesn’t have to be consequentialist (e.g. “Never kill a human, no matter what”.)

But where do the SI’s treacherous goals and final values come from?

- Why they should be hostile to humans?
- Most human religions worship the creator of humanity (i.e. God) without any particular assumptions about the intelligence of the creator, who was (sometimes anyway) admirable and an object of positive emotions.
- Why therefore might an SI not feel similarly about its creator, humanity, independently of the creator’s IQ level which it might perceive, correctly , to be lower than its own?
- This is as plausible as Bostrom’s key “default” assumption to the contrary and at least has some historical-anthropological backing.

What then can be done to ward off catastrophe?

- Bostrom’s remedies are surprisingly conventional:
  - Put the SI without arms in a box ☺
  - Give it incentives to be nice
  - “Stunting” it so it is less clever
  - Install tripwires in case it goes too far in some direction.
  - A final cluster of suggestions are motivational, to use Bostrom’s term: following Asimov’s laws of robotics, one might install in the SI limited final goals, or try a set of goals out in a simpler, safer, system and then put it an SI (the “sandbox” approach).

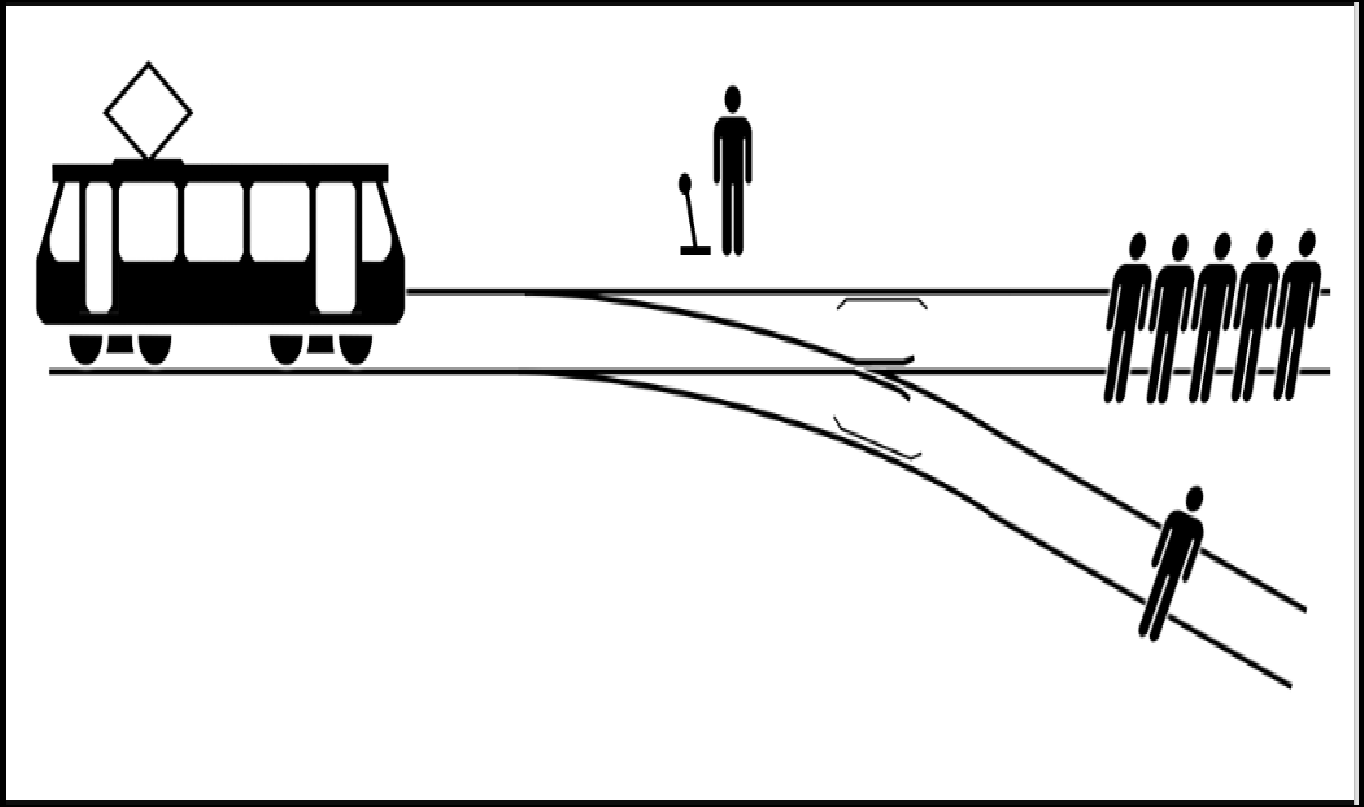
There are other ways (not in Bostrom) to think about this without falling back on Asimov's "Laws of Robotics"

- Thomas Schelling's work on complex game theoretic principles, that could extend Asimov's laws, take the form of constraints on what an entity MUST NOT DO. One might work back from these to plans of action and strategy in a way that would be harder to unpick for the SI than Asimov's over-simple rules.
- Hubert Dreyfus, an acute philosophical critic of AI, argued long ago that AI could only succeed or have values if and when it consisted of engineered entities that grew up and learned as we do.

Where we came in: Ethics, AI, trolleys and automated cars.

- Phillipa Foot’s “trolley problem” (1978) has resurfaced as a way of analysing the kill-or-save choices an automated car will be required to make.
- Consequentialism vs. Rules made concrete: do I push one fat man off a bridge to save five people from an oncoming train? (I paraphrase)
- Automated cars will be the most striking piece of AI in coming years and may have to make just these choices, including: do I protect my driver?
- There are no such concrete ethical problems for the Bostrom’s SI, beyond a suggestion that we should perhaps defer to the SI’s own values because it knows more than we do *ex hypothesi*.
- This is a very odd conclusion indeed to work that argued forcefully that an SI’s final values might have no connection to human ones and might well seek to destroy us.

Classic Trolley Problem:
Should you pull the lever to divert the runaway trolley onto the side track?



There are even darker figures lurking out there in SI-land……..

- Links to :”dark transhumanists”: the singularity, the Rapture, and the Catastrophe are often associated with strong libertarian/inegalitarian/racist view on the future of humanity (e.g. Thiel).
- SI is another face of “a future of technologically enhanced beings.” (the optimistic version---compare the Nazi’s “evolutionary humanism”)
- Humans as tools for technology, not vice versa.
- Which is modern fusion of gnosticism (=secret knowledge of the few) and technocracy.
- Gates, Hawking, Musk are all pessimists about the future, though without any real knowledge of AI.

There is little of AI and its history in the whole SI argument

- An old fashioned view (HAL9000) of AI as focused on logic and inference methods
- But Feigenbaum as long ago as noted the 1970s that “for a long time AI focused its attention almost exclusively on the development of clever inference methods,” only to discover that “the power of its systems does not reside in the inference method.”
- AI has moved on to knowledge bases, then big data and learning algorithms.
- SI unlikely to come, and it certainly won’t the way Bostrom fears.

New (2017) predictions of malevolence

- From Cambridge University's Centre for the Study of Existential Risk and others (with obvious answers)
  - **Drones seeking individuals to assassinate [cf. decades of Bond assassins]**
  - **Automated cars as bombs [cf. truck suicide bombers]**
  - **Fake videos of the famous or even you [cf. faked spliced sound recordings]**
  - **Tech-assisted coups d‘etat [wrong countries have coups]**
  - **Employment loss [since the weaving machine]**
  - **Legislative concentration –to one malicious AI [forgets how small electorates used to be!]**
  - **Letting off A-bombs [two keys and four arms still needed]**

Those possibilities cover both SI (AI-self-development) and MAI, the deliberate construction of AI as a malevolent weapon.

- One strong precaution against both is to limit the use of *closed source* hardware and software
- That is, hardware and software that is not *open*, available for inspection and general use
- *“any group with the intent of creating a malevolent artificial intelligence would find the ideal conditions for operating in quasi-total obscurity, without any oversight board and without being screened or monitored, all the while being protected by copyright law, patent law, industrial secret”* (Yampolskiy, 2017)

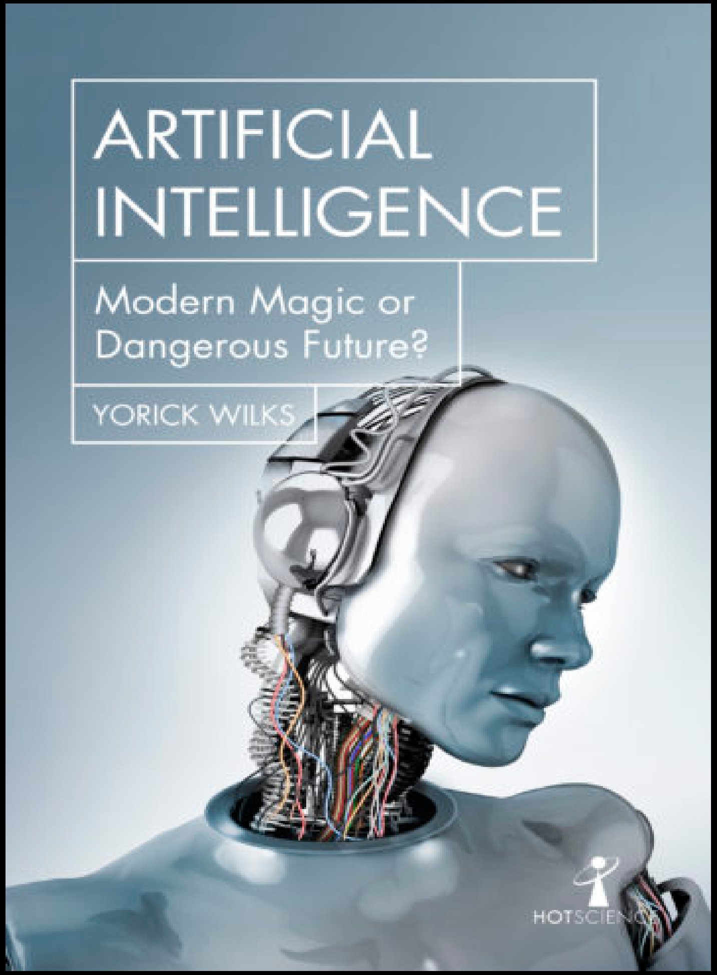
Don’t believe claimed a priori limitations on the scope of AI and predictions of failure

- Automated cars will be a huge success and bring enormous changes to our lives
- The Turing test and chatbots are irrelevant
- Claimed limits on decidability equally so
- Computer reasoning can be probabilistic as easily as logical if that’s what we design in.
- They don’t have to do things well the way we do (an AI child might be 100 before it learned to speak)
- AI devices don’t have to solve eternal ethical dilemmas to be useful.
- Any more than they have to “solve semantics” (!) to do excellent translation.

But AI may produce many non-SI problems in the future

- “AI-enhanced technologies might still be extremely dangerous due to their potential for amplifying human stupidity. The AIs of the foreseeable future need not think or create to sow mass unemployment, or enable new weapons technologies that undermine precarious strategic balances. Nor does artificial intelligence need to be smarter than humans to threaten our survival—all it needs to do is make the technologies behind familiar 20th-century existential threats faster, cheaper, and more deadly.” **Edward Geist**.

Oh, and here’s a new book this week just in case you’re looking for an AI primer!!


ARTIFICIAL INTELLIGENCE
Modern Magic or Dangerous Future?
YORICK WILKS
HOTSCIENCE