Provably Beneficial AI

Stuart Russell
University of California, Berkeley
[joint work with Dylan Hadfield-Menell, Smitha Milli, Anca Dragan, Pieter Abbeel, Tom Griffiths]

Premise

- Eventually, AI systems will make better* decisions than humans
  - Taking into account more information, looking further into the future

Upside

- Everything we have is the product of intelligence
- Access to significantly greater intelligence would be a step change in civilization

Downside

The Telegraph

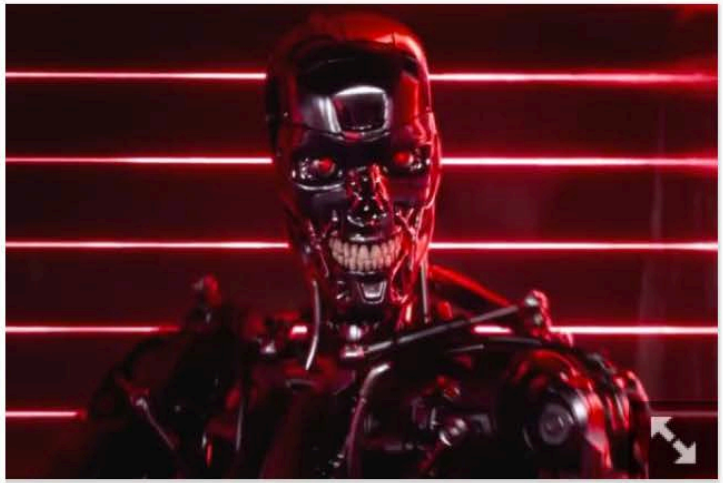
'Killer Robots' could be outlawed

'Killer Robots' could be made illegal if campaigners in Geneva succeed in persuading a UN committee, meeting on Thursday and Friday, to open an investigation into their development

TAG Robots , Robotics , Unemployment

Robots Could Replace Half Of All Jobs In 20 Years

By **Timothy Torres**, Tech Times | March 24, 6:56 PM



Robots will replace 47 percent of all jobs by the year 2035 if we're to believe University of Oxford associate professor Michael Osborne.
(Photo : Paramount)

If we're to believe University of Oxford associate professor Michael Osborne, then robots will replace 47 percent of all jobs by the year 2035.

If you want to stay employed by then, you better think about a career shift into software development, higher level management or the information sector. Those professions are only at a 10 percent risk of replacement by robots, according to Osborne. By contrast, lower-skilled jobs in the accommodation and food service industries are at a 87 percent risk, transportation and warehousing are at a 75 percent risk and real estate at 67 percent. The researcher warns that driverless cars, burger-flipping robots and other automatons taking over low-skilled jobs is the way of the future.

BALTIMORE
Post-Examiner

Artificial Intelligence could spell the end of the human race

BY PAUL CROKE · JUNE 9, 2015 · NO COMMENTS

What's bad about better AI?

"If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by ***turning off the power*** at strategic moments, we should, as a species, feel greatly humbled. ...

This new danger … is certainly something which can give us anxiety."

Alan Turing, 1951

What's bad about better AI?

If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively … we had better be quite sure that the purpose put into the machine is the purpose which we really desire

Norbert Wiener, 1960

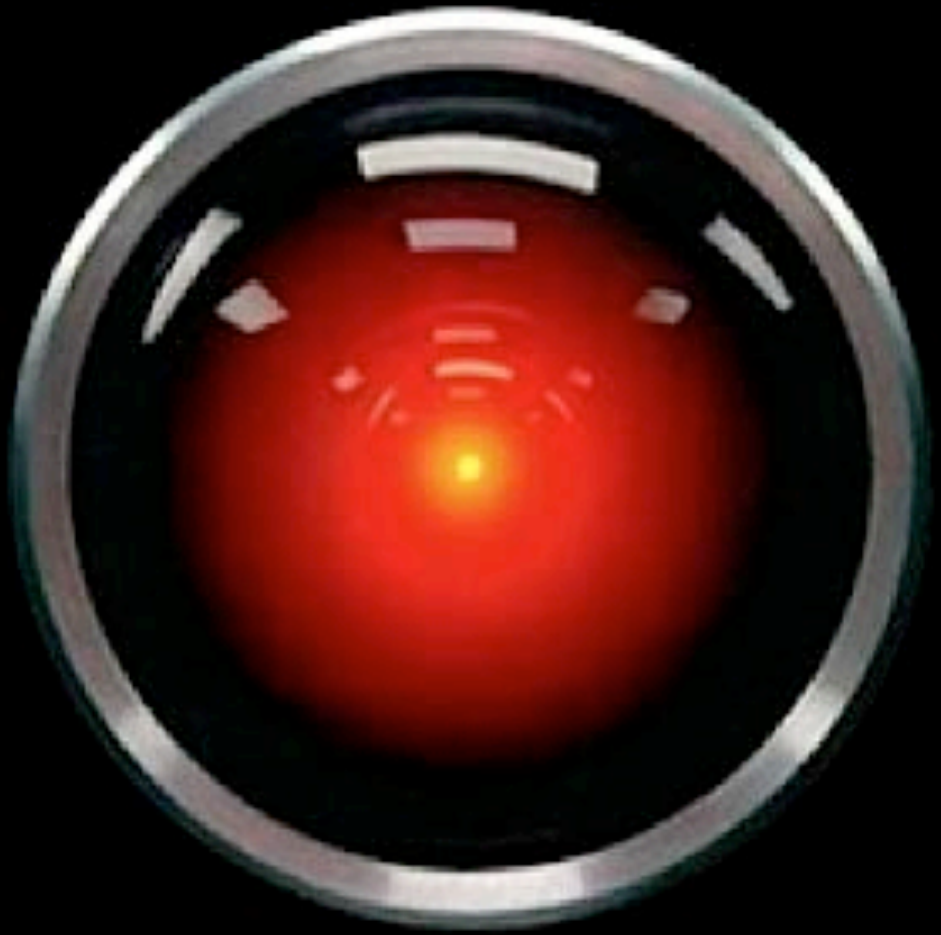
King Midas, c540 BCE

Value misalignment

- AI systems that are incredibly good at achieving something other than what we really want
- AI, economics, statistics, operations research, control theory all assume utility to be ***exogenously specified***

Instrumental goals

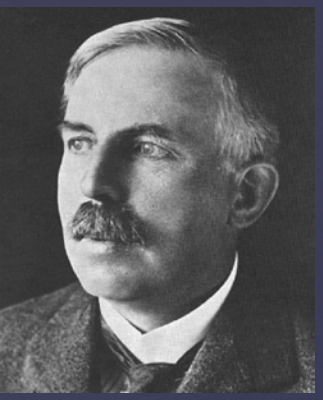
- For ***any primary goal***, the odds of success are improved by
  1) Maintaining one's own existence
     (you can't fetch the coffee if you're dead)
  2) Acquiring more resources
- With value misalignment, these lead to obvious problems



I’m sorry, Dave, I’m afraid I can’t do that

Reasons not to pay attention:

- It'll never happen



Sept 11, 1933: Lord Rutherford addressed BAAS: *"Anyone who looks for a source of power in the transformation of the atoms is talking moonshine."*

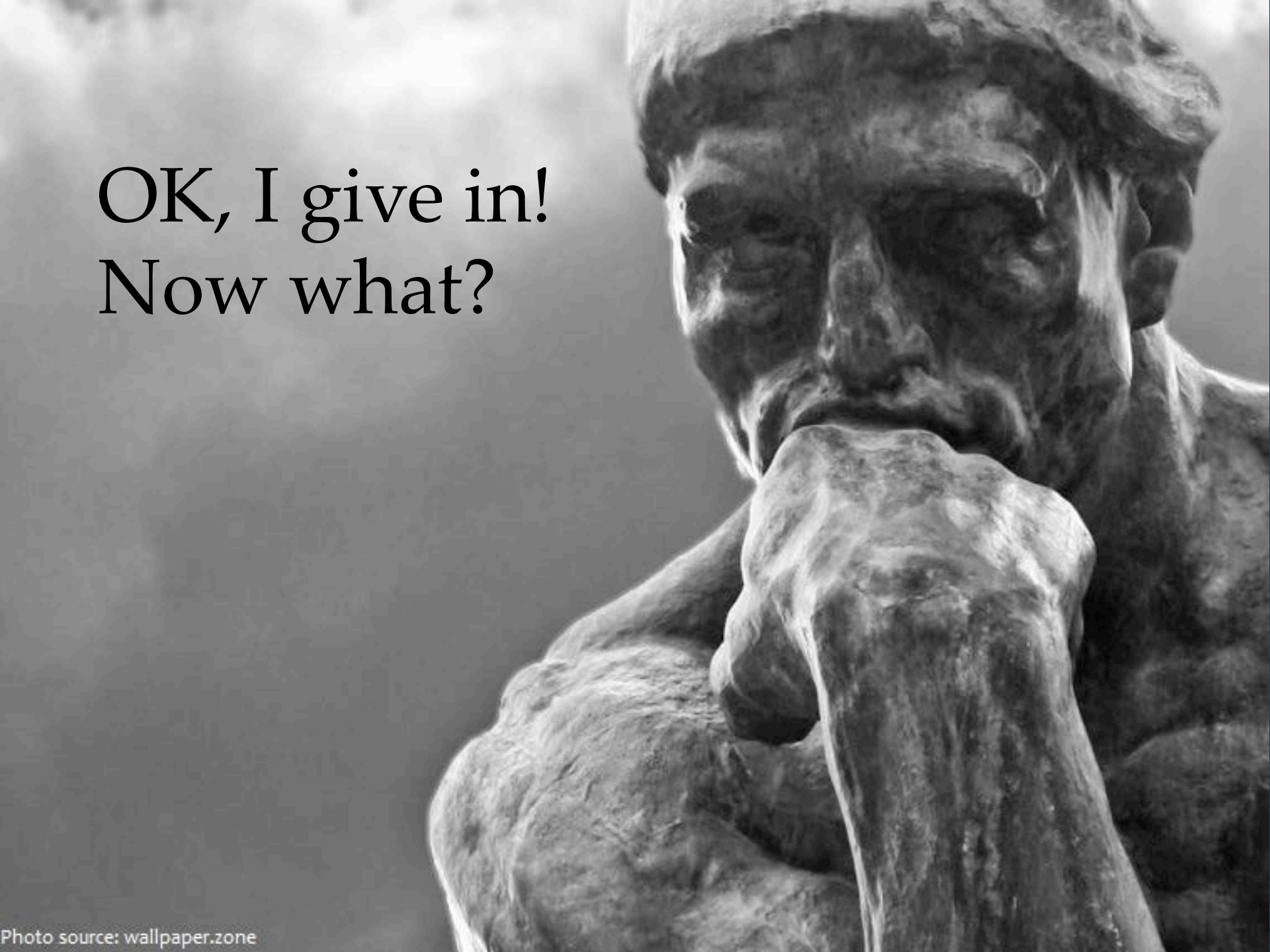
Sept 12, 1933: Leo Szilard invented neutron-induced nuclear chain reaction

Reasons not to pay attention:

- It'll never happen
  - See Rutherford, 9/11/33, Szilard 9/12/33
- It's too soon to worry about it
  - 2066 asteroid collision: when exactly do we worry?
  - When should we have worried about climate change?
- It's like worrying about overpopulation on Mars
  - No, it's as if we were spending billions moving humanity to Mars with no plan for what to breathe
- Just don't have explicit goals for the AI system
  - We need to steer straight, not remove the steering wheel
- Don't worry, we'll just have human-AI teams
  - Value misalignment precludes teamwork

Reasons not to pay attention:

- You can't control research
  - Yes, we can: we don't genetically engineer humans
- You're just Luddites/anti-AI
  - Fusion researchers are Luddites if they point out the need for containment?
  - Turing, Wiener, Minsky, Gates, and Musk are Luddites?
- Don't worry, we can just switch it off
  - As if a superintelligent entity would never think of that
- Don't put in "human" goals like self-preservation
  - Death isn't bad per se. It's just hard to fetch the coffee after you're dead
- **Don't mention risks, it might be bad for funding**
  - **See nuclear power, GMOs, tobacco, global warming**


OK, I give in!
Now what?
Photo source: wallpaper.zone

Center for Human-Compatible AI

reorient the general thrust of AI research towards provably beneficial systems

Also FHI, CSER/LCFI, MIRI, FLI, OpenAI

AAAI, IEEE, NSF, DARPA, PonAI

Three simple ideas

1. The robot's only objective is to maximize the realization of human values
2. The robot is initially uncertain about what those values are
3. The best source of information about human values is human behavior

Value alignment

- ***Inverse reinforcement learning***: learn a value function by observing another agent's behavior
  - The value function is a succinct explanation for what the other agent is doing

Cooperative inverse reinforcement learning

- A two-player game with "human" and "robot"
  - Human "knows" the value function (usually acts according to it)
  - Robot doesn't know it, but wants to maximize it
- Optimal solutions have these properties:
  - Robot has an incentive to ask questions first
  - Human has an incentive to teach the robot

The off-switch problem

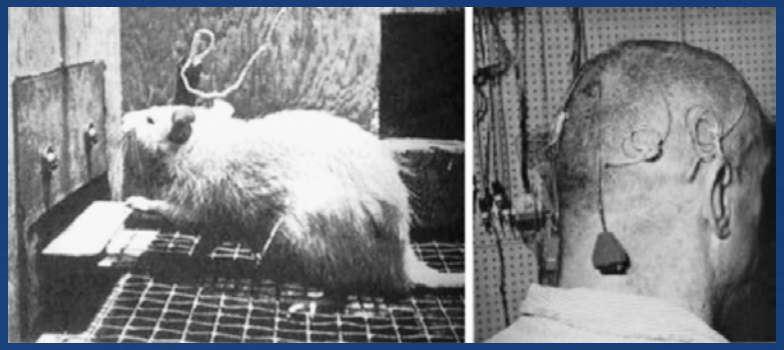
- A robot, given an objective, has an incentive to disable its own off-switch

  (You can't fetch the coffee if you're dead)
- How can we prevent this?
- Answer: robot isn't given an objective!
- Instead, it must allow for ***uncertainty*** about the true human objective
  - The human will only switch off the robot if that leads to better outcomes for the true human objective
  - Theorem: it's ***in the robot's interest*** to allow it

Uncertainty in objectives

- Largely ignored, even though uncertainty has been central to AI since early 1980s
- *Irrelevant* in standard decision problems…
- …*Unless* the environment provides further information about objectives
  - E.g., observable human actions
  - E.g., reward signals in RL

Reward signals



- Wireheading
  - In a real RL problem, rewards come from environment
  - => RL agent hijacks the reward-generating mechanism

- Mathematical framework for RL is wrong: reward signals are ***not*** actual rewards
- A "reward signal" is a human action that provides *information* about the true reward
- Hijacking the mechanism *loses information*

Provably beneficial AI

- Define a formal problem F that we assume the robot solves arbitrarily well
  - The robot is an F-solver, not just "AGI"
- Desired theorem: The human is provably better off with the robot
- Move the theoretical framework gradually towards reality

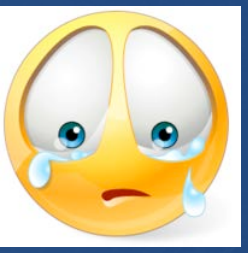
Reasons for optimism

- Vast amounts of evidence for human behavior and *human attitudes towards that behavior*
- We need value alignment even for *subintelligent* systems in human environments; strong economic incentives!
  - E.g., are all photo misclassification costs equal?

DERANGED ROBOT COOKS KITTY FOR FAMILY DINNER

Reasons for pessimism

Reasons for working hard



- Humans are nasty, irrational, inconsistent, weak-willed, computationally limited, and heterogeneous

Practical projects

- It's hard to work on control of AGIs
- OK, work on something simpler:
  - intelligent personal assistant
  - smart home
  - …
- Simulation environments where real (simulated) disasters can happen

Questions

- Can we change the way AI defines itself?
  - A civil engineer says "I design bridges", not "I design bridges that don't fall down"
  - Willingness to re-examine foundations
- How can we engage social scientists?
- Where do human value systems come from?
- Can AI optimize future social evolution?
- Will it make us better people?

Wiener, contd.

This work requires an imaginative forward glance at history which is difficult, exacting, and only partially achievable. …

We must always exert the full strength of our imagination to examine where the full use of our new modalities may lead us