Provably Beneficial AI

Stuart Russell
University of California, Berkeley

TECH • ARTIFICIAL INTELLIGENCE

United Kingdom Plans $1.3 Billion Artificial Intelligence Push

France to spend $1.8 billion on AI to compete with U.S., China

EU wants to invest £18bn in AI development

China's Got a Huge Artificial Intelligence Plan

Premise

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

Upside

- Access to significantly greater intelligence would be a step change in civilization
- NPV (HLAI) ≈ $13,500T

Downside

The Telegraph

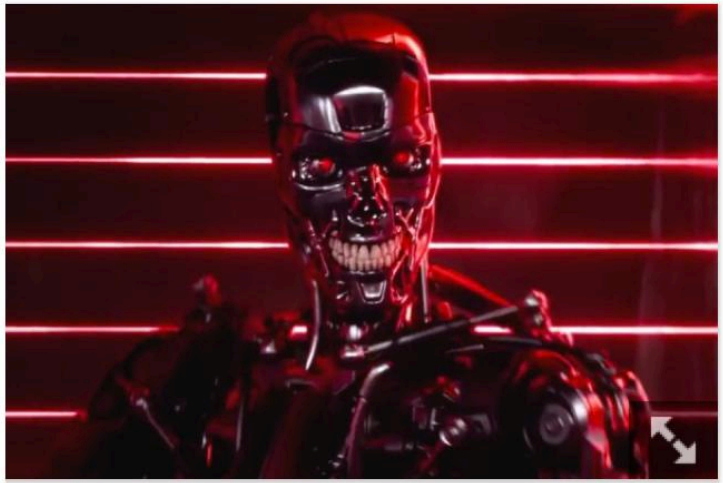
'Killer Robots' could be outlawed

'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

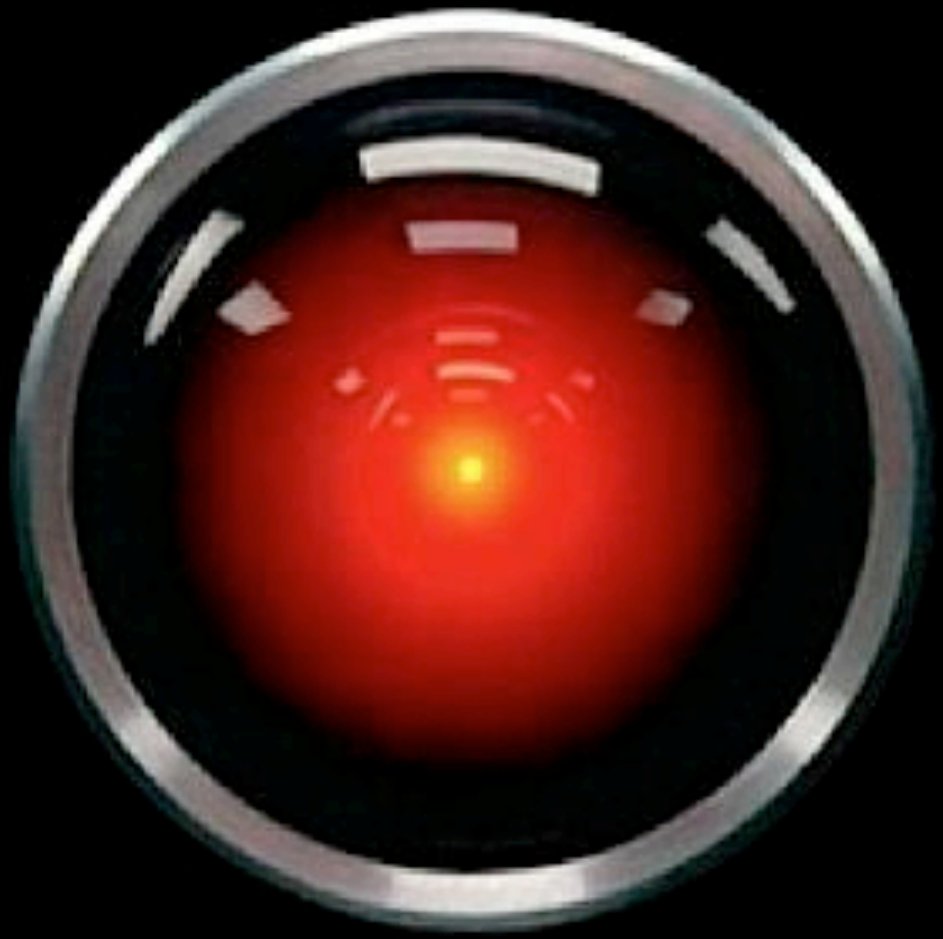
What's bad about better AI?

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

You can't fetch the coffee if you're dead



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

Where did we go wrong?

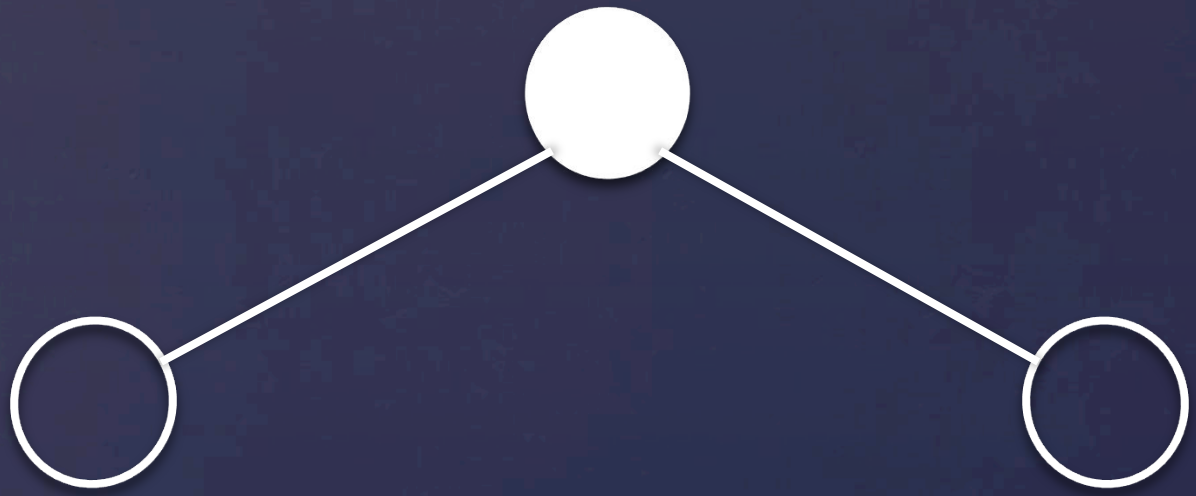
- Humans are intelligent to the extent that our actions can be expected to achieve our objectives
- Machines are intelligent to the extent that their actions can be expected to achieve their objectives
  - Give them objectives to optimize (cf control theory, economics, operations research, statistics)
- We don't want machines that are intelligent in this sense
- Machines are ***beneficial*** to the extent that ***their*** actions can be expected to achieve ***our*** objectives
- We need machines to be ***provably beneficial***

Three simple ideas

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

AIMA 1,2,3: objective given to machine

Human objective

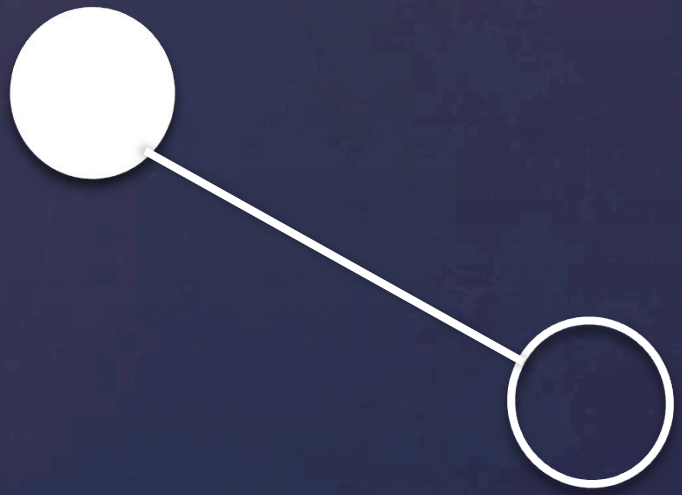


Human behaviour

Machine behaviour

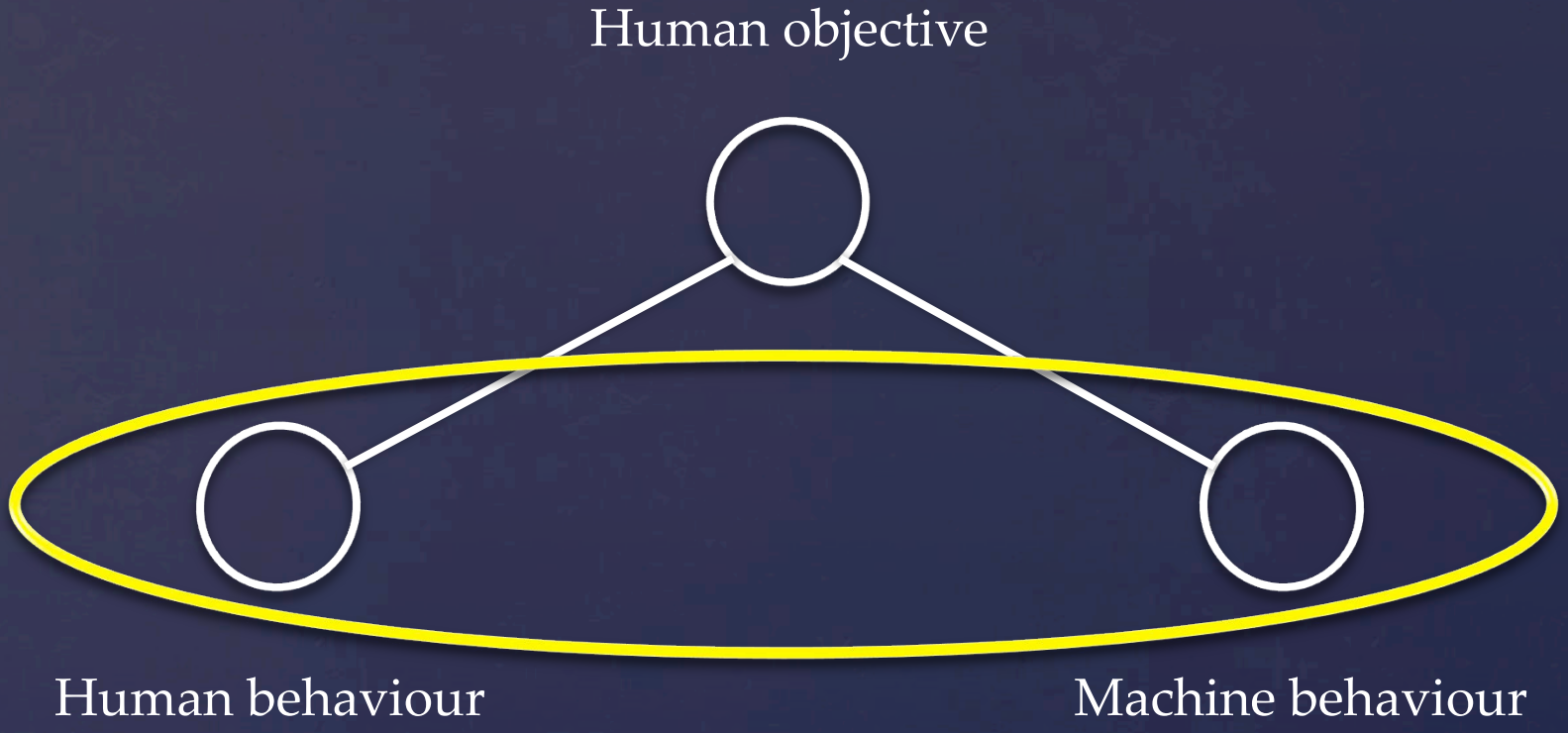
AIMA 1,2,3: objective given to machine

Human objective



Machine behaviour

AIMA 4: objective is a latent variable



Example: image classification

- Old: minimize loss with (typically) a *uniform* loss matrix
  - Accidentally classify human as gorilla
  - Spend millions fixing public relations disaster
- New: structured prior distribution over loss matrices
  - Some examples safe to classify
  - Say “don’t know” for others
  - Use active learning to gain additional feedback from humans

Example: fetching the coffee

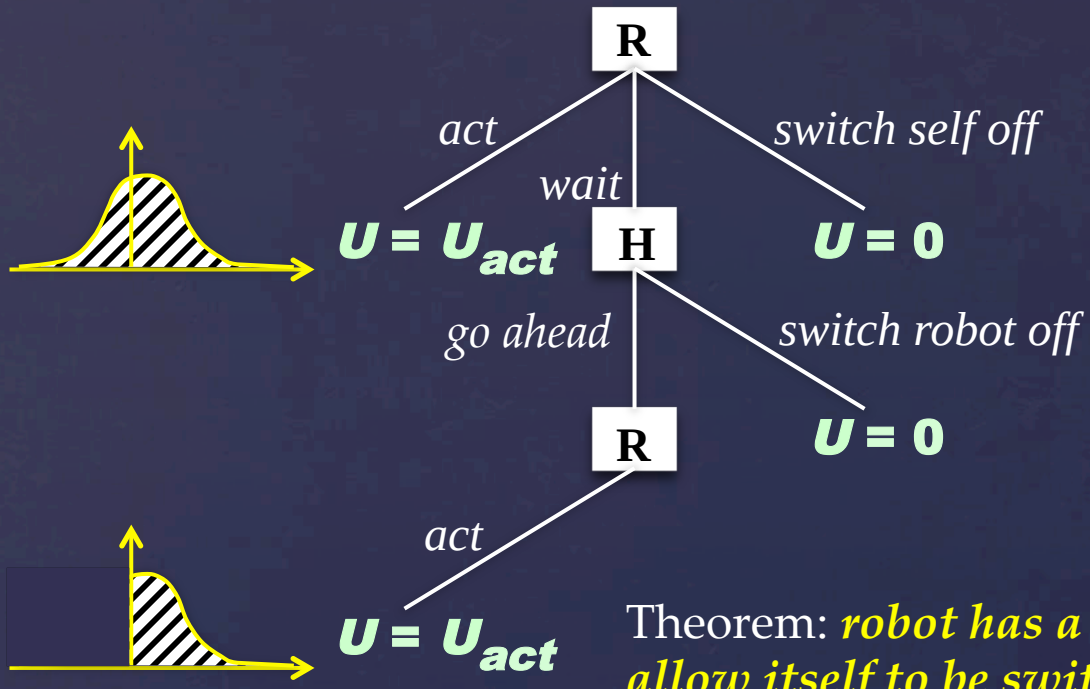
- What does "fetch some coffee" mean?
- If there is so much uncertainty about preferences, how does the robot do anything useful?
- Answer:
  - The instruction suggests coffee would have higher value than expected a priori, ceteris paribus
  - Uncertainty about the value of other aspects of environment state doesn't matter *as long as the robot leaves them unchanged*

- Noninterference is (usually) good because the world is (roughly) in the stationary distribution resulting from agents operating with preferences
  - => preferences ***can*** be inferred from the state of the world

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"
- A robot with uncertainty about objective won't behave this way

Off-switch model



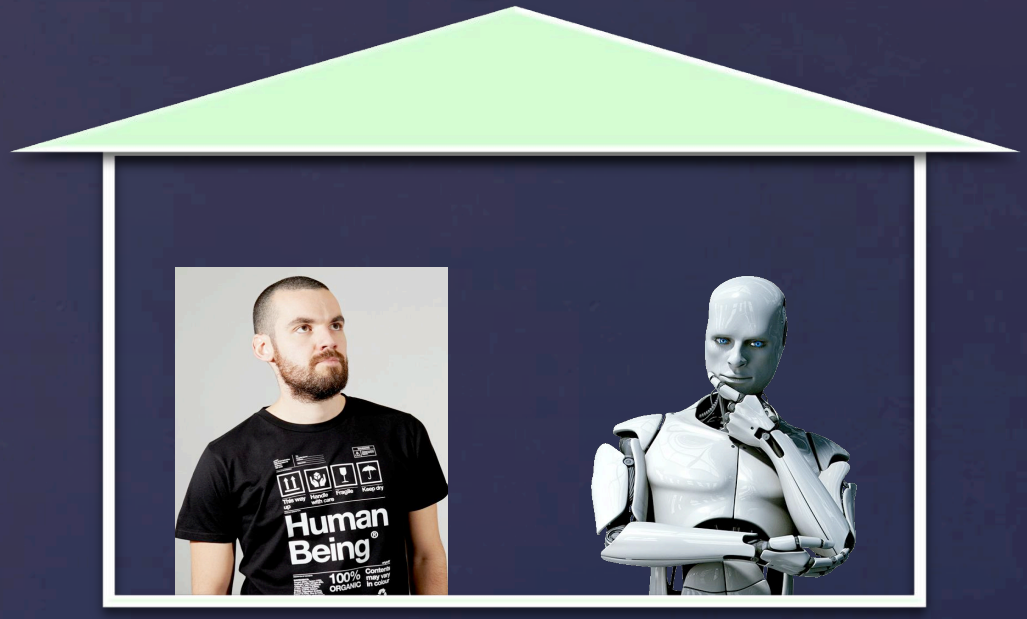
Theorem: ***robot has a positive incentive to allow itself to be switched off***

Theorem: ***robot is provably beneficial***

Learning from human behavior

- ***Inverse reinforcement learning***: learn a reward function by observing another agent's behavior
- ***Cooperative IRL***:
  - human and robot in same environment

Basic CIRL game



Preferences θ
Acts roughly according to θ

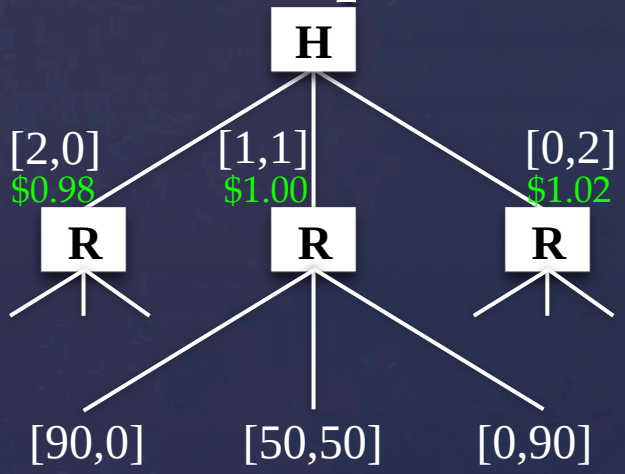
Maximize unknown human θ
Prior P(θ)

CIRL equilibria: Human teaches robot
Robot asks questions, permission; defers to human; allows off-switch

Solve by reduction to POMDP in [s,θ]
[Hadfield-Menell et al, NIPS 16; Fisac et al, ISRR 17; Palaniappan et al, ICML 18]

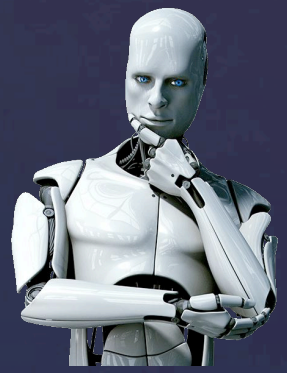
Example: paperclips vs staples

- State (p,s) has p paperclips and s staples
- Human reward is θp + (1-θ)s and θ=0.49
- Robot has uniform prior for θ on [0,1]



[1,1] is optimal
($51.00 vs $46.92)

One robot, many humans



- Weighing human preferences:
  - Harsanyi: Pareto-optimal policy optimizes a linear combination when humans have a common prior over the future
  - Critch, Russell, Desai (NIPS 18): weights proportional to whose predictions turn out to be correct
- Utility monsters (Nozick, 1974)
- Welfare aggregation and the Somalia problem

Real(ish) humans

- Computationally limited, irrational
  - Hierarchically organized behavior
  - Emotional states affecting behavior, revealing preferences
- Heterogeneous
- Nasty
  - Zero out negative-altruism preferences (sadism, pride/envy)
- Inconsistent, non-additive, memory-laden preferences
  - “two selves” (Kahneman, 2015)
- Plastic/adaptive preferences

Summary

- AI may eventually overtake human abilities
- Provably beneficial AI is possible ***and desirable***
  - ***It isn't "AI safety," it's AI***
  - Continuing theoretical work (AI, CS, economics)
  - Initiating practical work (assistants, robots, cars)
  - Inverting human cognition (AI, cogsci, psychology)
  - Long-term goals (AI, philosophy, polisci, sociology)
- Remaining problems…