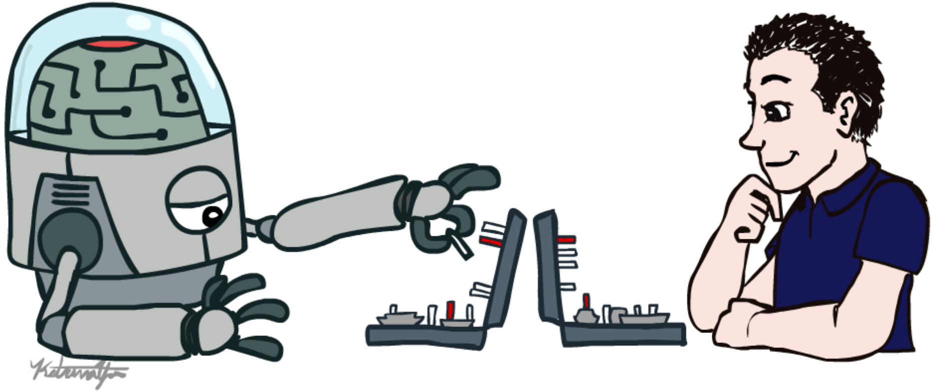
CS 188: Artificial Intelligence

Conclusion



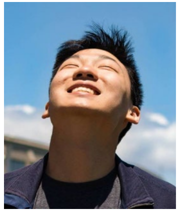
Instructors: Sergey Levine and Stuart Russell

Course Staff – Thanks!!


Aditya Baradwaj


Adam Gleave


Alex Li


Austen Zhu


Avi Singh


Charles Tang


Dennis Lee


Dequan Wang


Ellen Luo

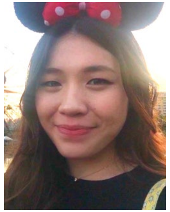

Fred Ebert


Henry Zhu


Jasmine Deng

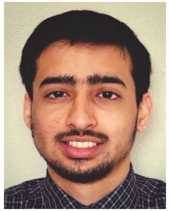

Jason Peng


Katie Luo


Laura Smith

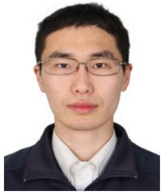

Micah Carroll


Mike Chang

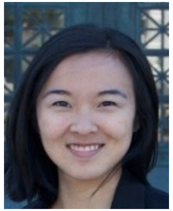

Murtaza Dalal


Rachel Li


Rishi Veerapaneni


Ronghang Hu


Sid Reddy


Simin Liu


Tony Zhao


Wilson Yan


Xiaocheng (Mesut) Yang

Announcements/Reminders

- Final exam: Thursday May 16, 7pm
  - Practice final online: 1pt extra credit if done by May 6
  - Clobbering policy: midterm score <- max(midterm score, final score)
  - HW12 (extra practice questions on ML, ungraded)
- RRR week: GSI office hours only

News AI

TECH • ARTIFICIAL INTELLIGENCE

United Kingdom Plans $1.3 Billion Intelligence Push

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

EU wants to invest £18b development

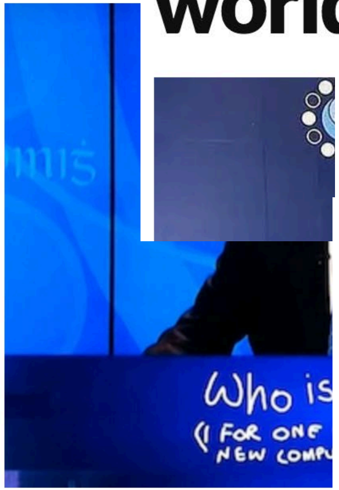
China's Got a Huge Art Intelligence Plan

'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day

Published time: 1 Sep, 2017 14:08
Edited time: 1 Sep, 2017 14:40

News AI

IBM's Watson Jeopardy Computer Shuts Down Humans in Final Game



DAILY NEWS 9 March 2016

‘I’m in shock!’ How an AI beat the world’s best human at Go

InfoQ

En | 中文 | 日本 | Fr | Br

Development

Architecture & Design

AI, ML and Data Engineering

Culture & Methods

DeepMind's AI Defeats Top StarCraft Players

LIKE 1 BOOKMARKS

APR 05, 2019 • 2 MIN READ

DeepMind's AlphaStar AI program recently defeated two top professional StarCraft players 5-0.

A note of caution

- Data is the new snake oil
  - Better learning => far less data needed
- Serious disappointments (e.g., autonomous vehicles) could result in a significant backlash

Google ponders the shortcomings of machine learning

Scientists of AI at Google's Google Brain and DeepMind units acknowledge machine learning is falling short of human cognition and propose that using models of networks might be a way to find relations between things that allow computers to generalize more broadly about the world.

By Tiernan Ray | October 20, 2018 -- 12:52 GMT (05:52 PDT) | Topic: Artificial Intelligence

François Chollet: "Many more applications are completely out of reach for current deep learning techniques – even given vast amounts of human-annotated data.

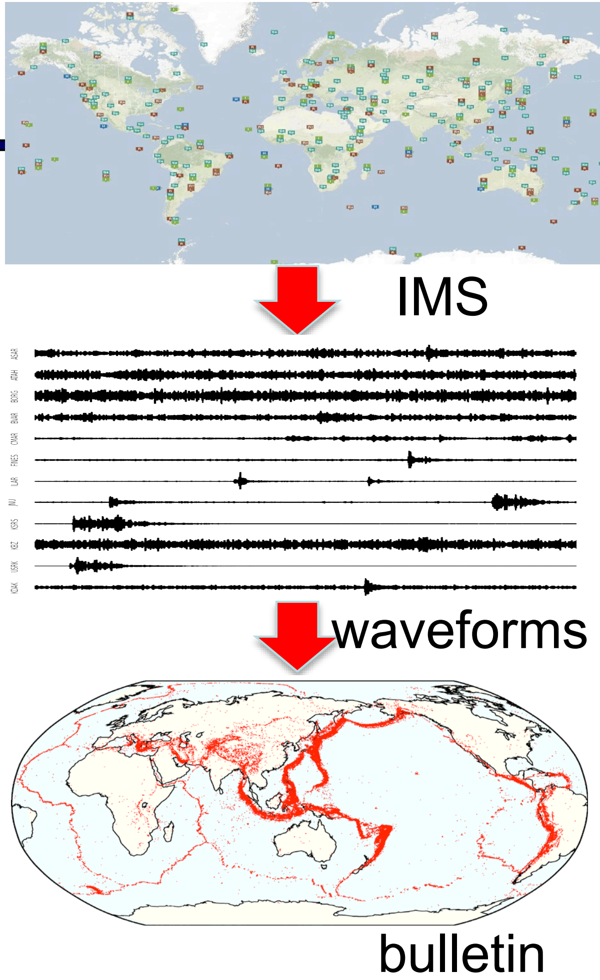
The main directions in which I see promise are models closer to general-purpose computer programs."

Probabilistic programming

Universal (Turing-equivalent) languages and algorithms for probabilistic modelling, learning, and reasoning

Global seismic monitoring for CTBT

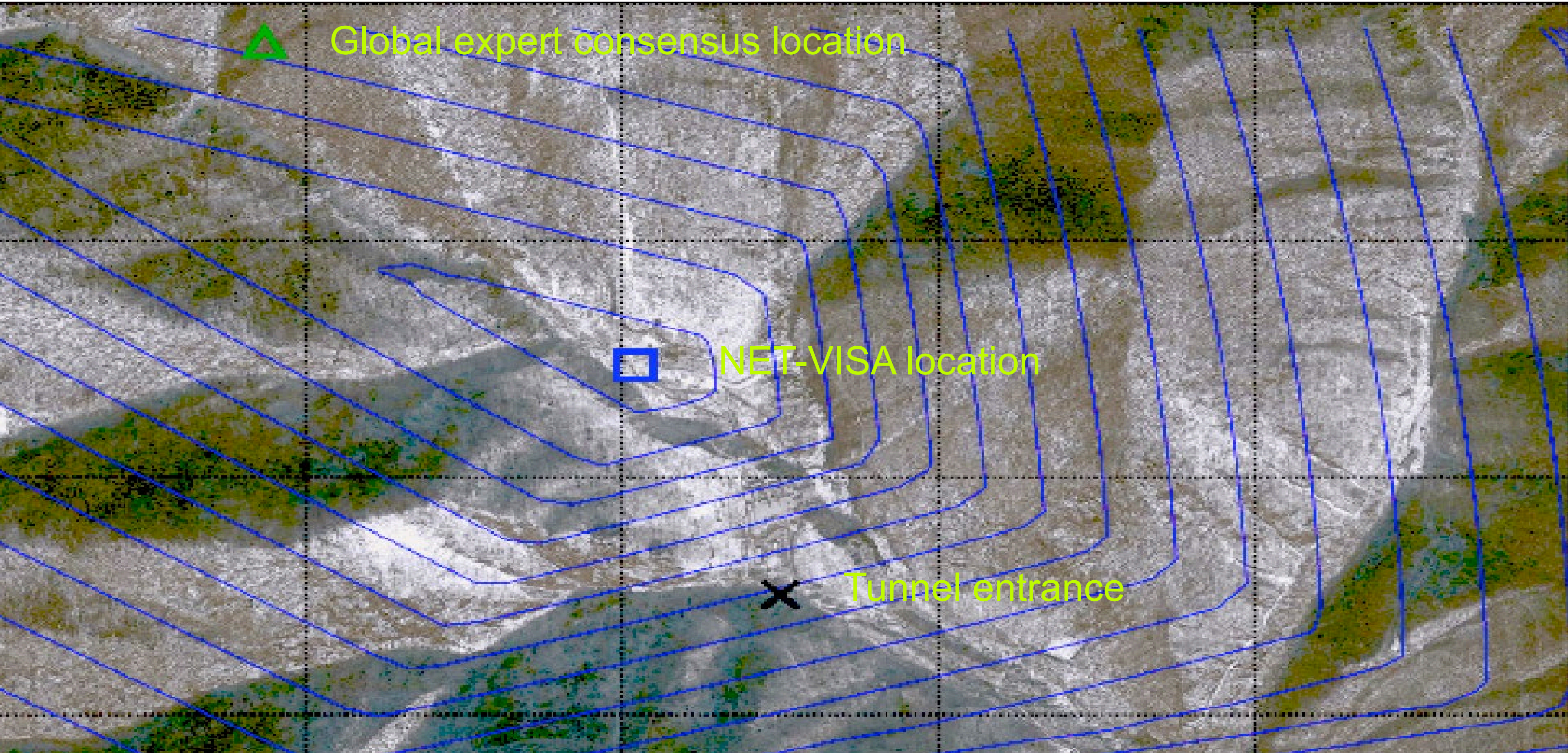
- **Evidence**: waveforms from 150 seismic stations
- **Query**: what happened?
- **Model**: geophysics of event occurrence, signal transmission, detection, noise



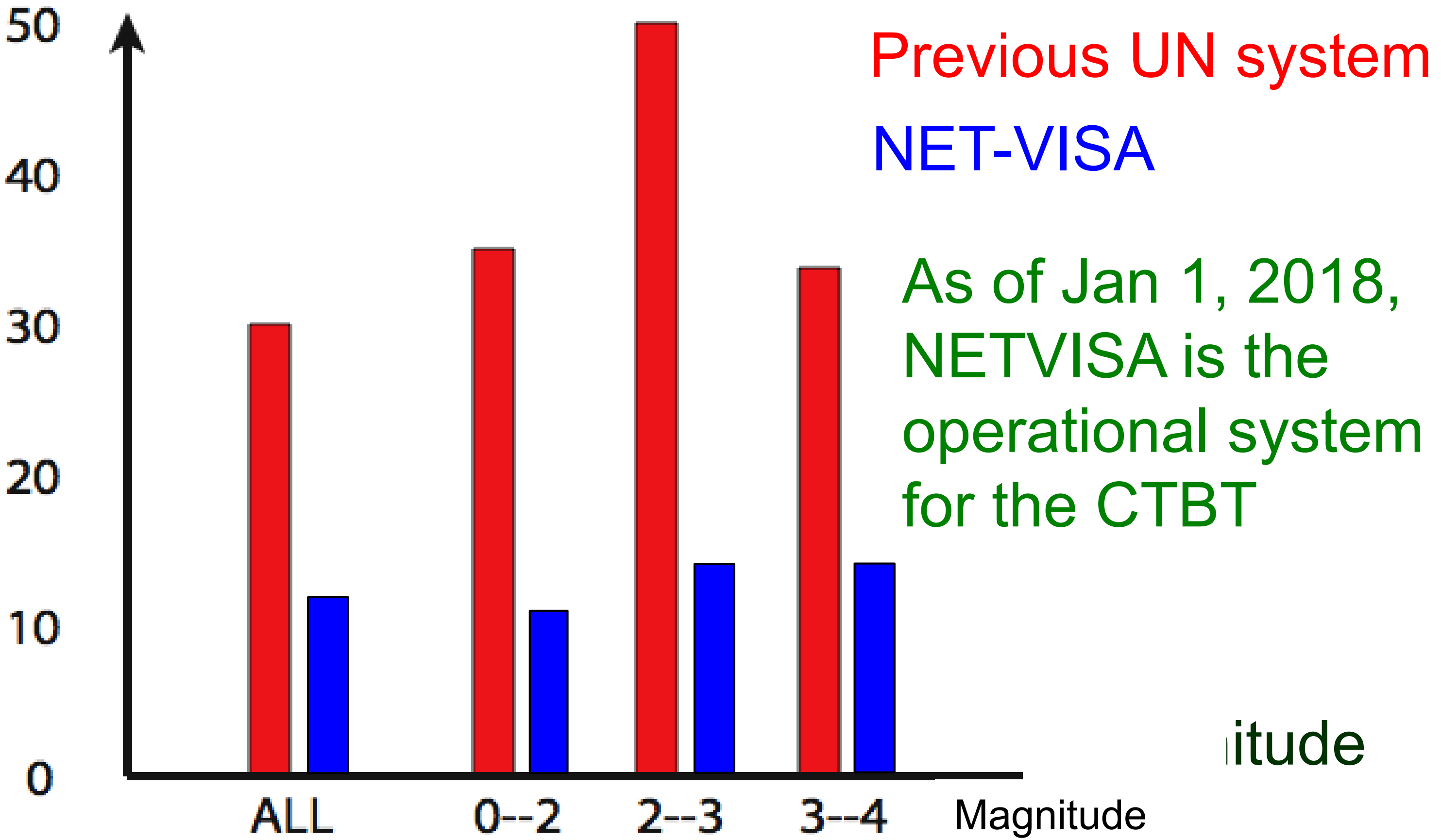
NET-VISA model

#SeismicEvents ~ Poisson[T*λ_e];
Time(e) ~ Uniform(0,T)
IsEarthQuake(e) ~ Bernoulli(.999);
Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();
Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;
Magnitude(e) ~ Exponential(log(10));
IsDetected(e,p,s) ~ Logistic[weights(s,p)](Magnitude(e), Depth(e), Distance(e,s));
#Detections(site = s) ~ Poisson[T*λ_f(s)];
#Detections(event=e, phase=p, station=s) = if IsDetected(e,p,s) then 1 else 0;
OnsetTime(a,s) ~ if (event(a) = null) then Uniform[0,T] else
  Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
    + Laplace(μ_t(s), σ_t(s))
Amplitude(a,s) ~ If (event(a) = null) then NoiseAmplitudeDistribution(s)
  else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))
Azimuth(a,s) ~ If (event(a) = null) then Uniform(0, 360)
  else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))
Slowness(a,s) ~ If (event(a) = null) then Uniform(0,20)
  else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))
ObservedPhase(a,s) ~ CategoricalPhaseModel(phase(a))

February 12, 2013 DPRK test



Fraction of events missed



As of Jan 1, 2018, NETVISA is the operational system for the CTBT

itude

Future

- We are doing AI…
  - To create intelligent systems
    - The more intelligent, the better
  - To gain a better understanding of human intelligence
  - To magnify those benefits that flow from it
    - E.g., net present value of human-level AI $\geq$ \$13,500T
    - Might help us avoid war and ecological catastrophes, achieve immortality and expand throughout the universe
- What if we succeed?

DEFENSE

Killer robots await Trump's verdict

The new president will have to decide how aggressively the U.S. pursues military technology that could let machines make life-or-death decisions.

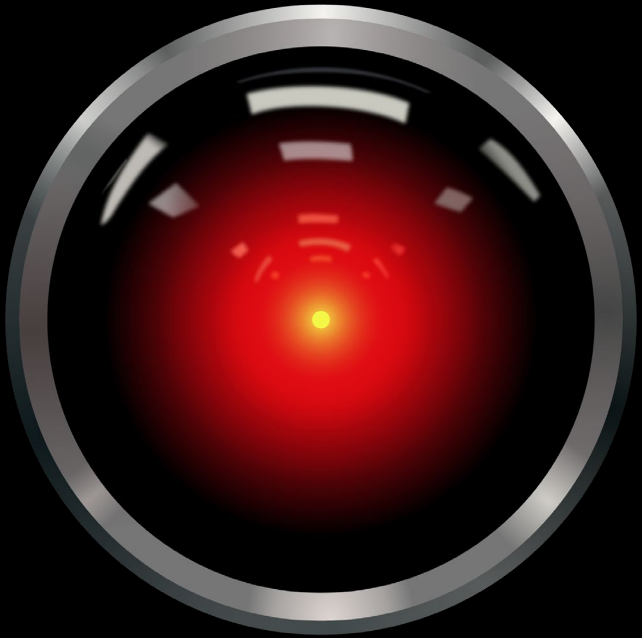
By ANDREW HANNA | 12/25/16 07:38 AM EST

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

Social media catastrophe

- Optimizing clickthrough
  - ~~= learning what people want~~
  - = modifying people to be more predictable

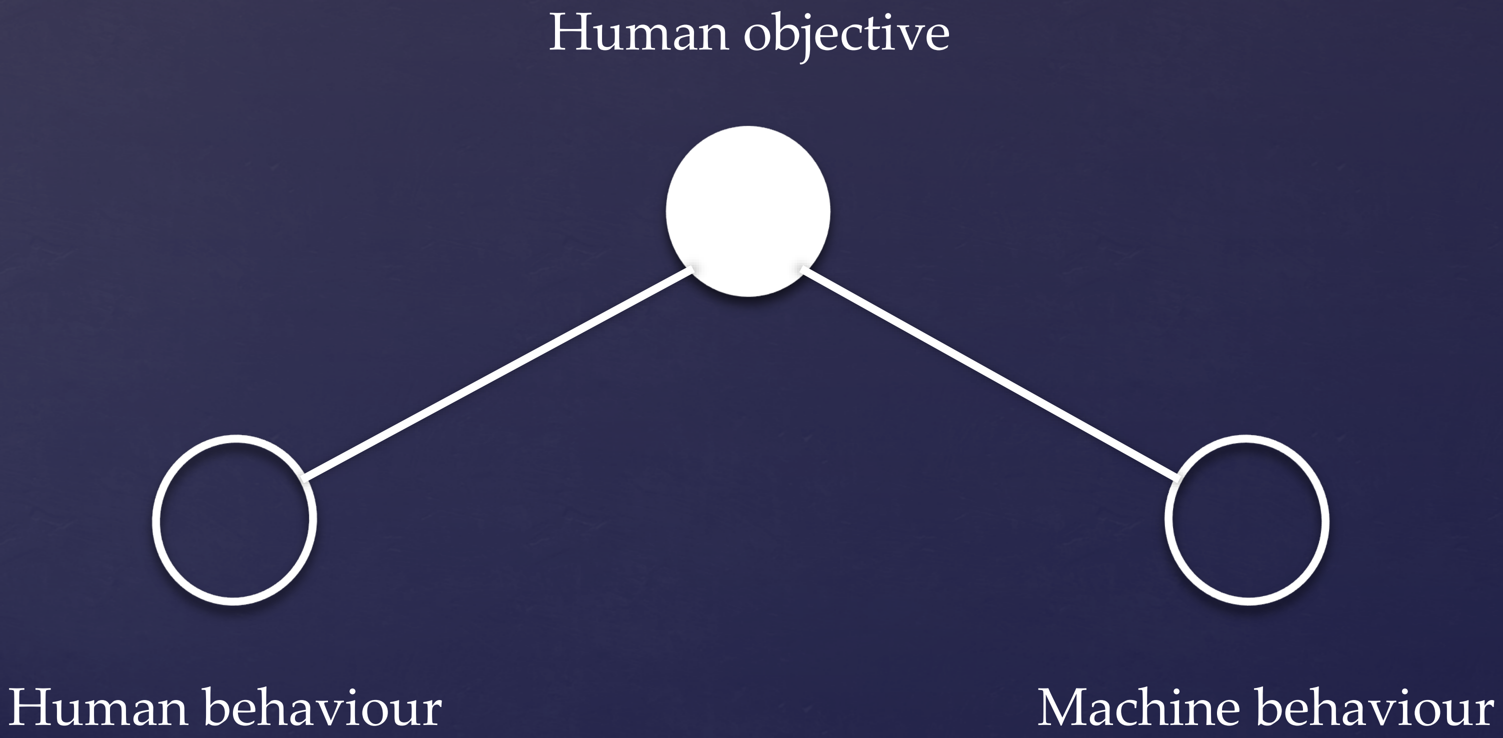
Where did we go wrong?

- 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***

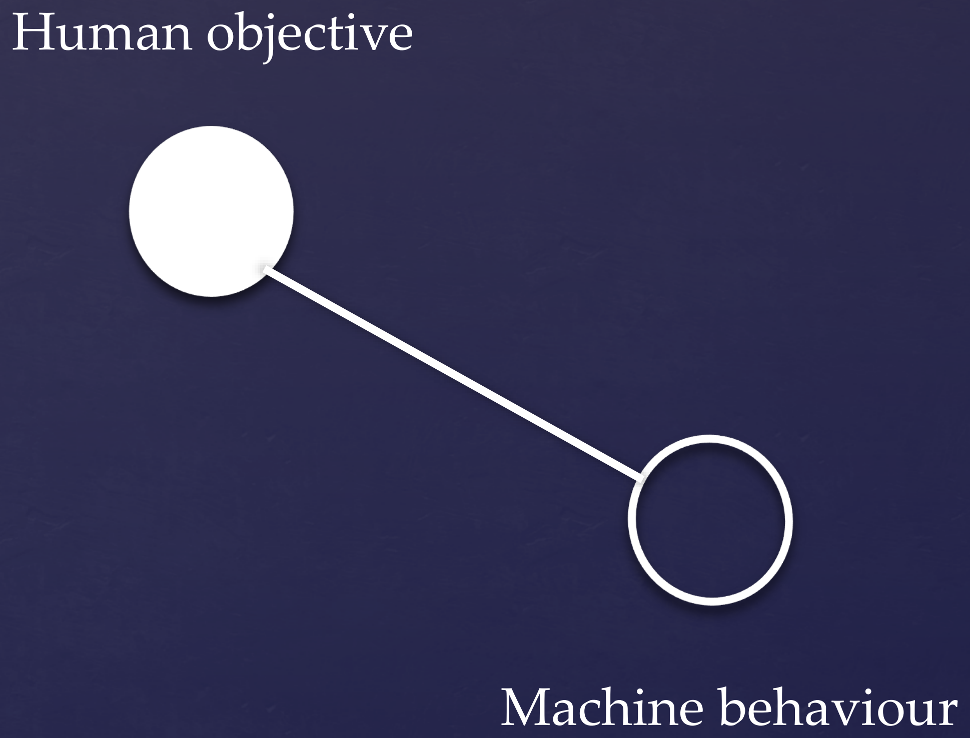
Three simple ideas

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*

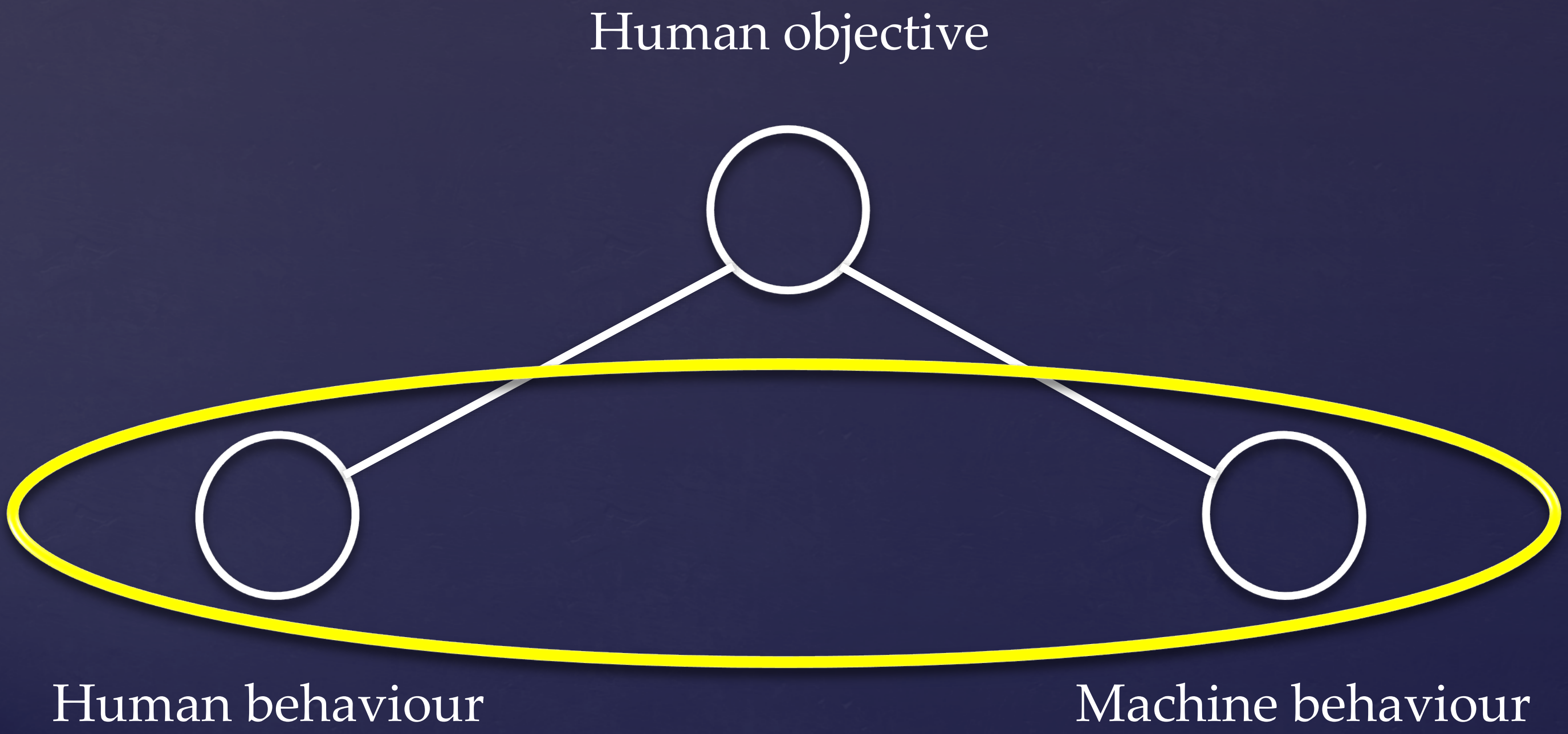
AIMA 1,2,3: objective given to machine



AIMA 1,2,3: objective given to machine



AIMA 4: objective is a latent variable



Example: image classification

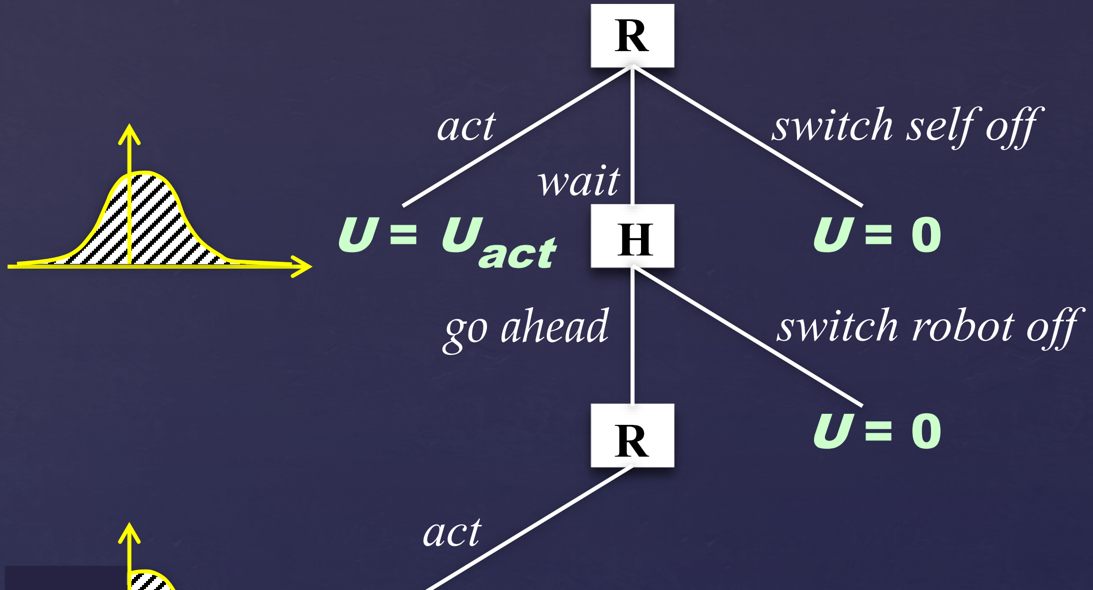
- Old: minimize loss with (typically) a ***uniform*** loss matrix
  - Accidentally classify human as gorilla
  - Spend millions fixing public relations disaster
- New: structured prior distribution over loss matrices
  - Some examples safe to classify
  - Say “don’t know” for others
  - Use active learning 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
    - and there’s probably a low-cost way to get it
  - Uncertainty about the value of other aspects of environment state doesn’t matter *as long as the robot leaves them unchanged*
  - ***Humans mostly like things the way they are***

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



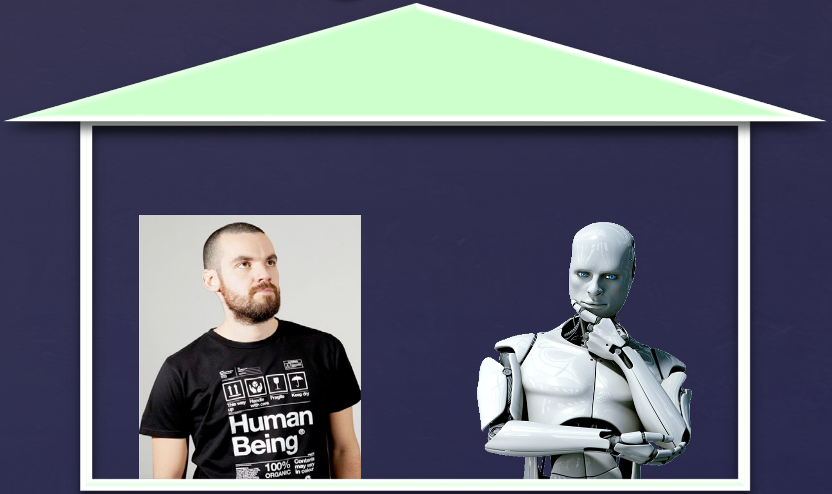
Theorem: ***robot has a positive incentive to allow itself to be switched off***

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

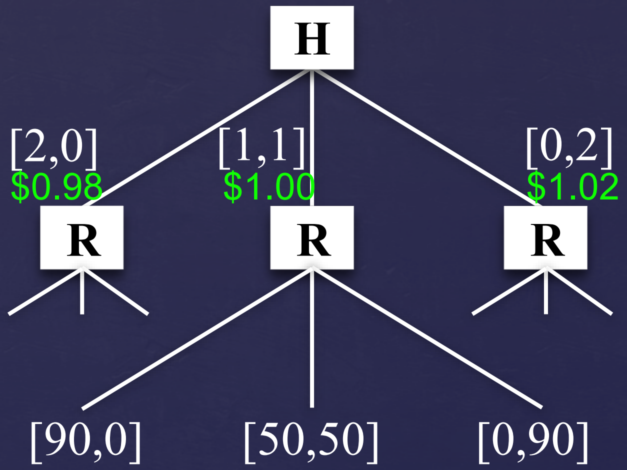
Maximize unknown human θ
Prior P(θ)

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 $\theta p + (1-\theta)s$ and $\theta=0.49$
- Robot has uniform prior for θ on [0,1]



[1,1] is optimal
for θ in [.446,.554]

One robot, many humans

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

Welcome home! Long day?

Yes, terrible, not even time for lunch.

So you must be quite hungry!

Starving! Anything for dinner?

There's something I need to tell you

There are humans in Somalia in more urgent need of help.
I am leaving now. Please make your own dinner.

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***
  - 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…