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
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 to develop
China’s Got a Huge Artificial Intelligence Plan
IBM's Watson Jeopardy Computer Shuts Down Humans in Final Game

DAILY NEWS 9 March 2016

Sili

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

DeepMind's AI Defeats Top StarCraft Players

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 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.

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
#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

Global expert consensus location

NET-VISA location

Tunnel entrance
Fraction of events missed

Previous UN system
NET-VISA

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

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

Human objective

Human behaviour

Machine behaviour
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
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

\[ R(s,a,s'; \theta) \]
Basic CIRL game

Preferences $\theta$
Acts roughly according to $\theta$
CIRL equilibria: Human teaches robot
Robot asks questions, permission; defers to human; allows off-switch

Maximize unknown human $\theta$
Prior $P(\theta)$

Solve by reduction to POMDP in $[s, \theta]$
[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 \(\theta\) on \([0,1]\)

\[
\begin{array}{c}
\text{H} \\
\text{R} \quad \text{R} \\
\quad \text{R} \\
[2,0] \quad [1,1] \quad [0,2] \\
\text{\$0.98} \quad \text{\$1.00} \quad \text{\$1.02} \\
[90,0] \quad [50,50] \quad [0,90]
\end{array}
\]

\([1,1]\) is optimal for \(\theta\) 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?

So you must be quite hungry!

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.

Yes, terrible, not even time for lunch.

Starving! Anything for 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...