CS 188: Artificial Intelligence

If We Succeed

Instructors: Stuart Russell and Dawn Song
What is AI?

AI = making intelligent machines

**Standard model:** machines are intelligent to the extent that their actions can be expected to achieve their objectives

AI comprises:

- problem solving, constraint satisfaction, games
- knowledge representation, reasoning, planning
- natural language processing
- speech, vision, robotics
- machine learning

The goal is **general-purpose AI**: capable of quickly learning high-quality behavior in “any” task environment
What if we succeed?

- Lift the living standards of everyone on Earth to a respectable level
  - => 10x increase in world GDP ($13.5Q net present value)
- Potential advances in health, education, science
- Massive disruption of human economic roles
It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control.
A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says
Reasons to doubt

- **Expressive power of circuits**
  - Causes massive blowup in simple concepts (e.g., rules of Go), slow learning, extreme fragility

- **Adversarial examples: dimpled-manifold conjecture**
  - Tiny, invisible changes in an image change its classification from school bus to ostrich
  - Hypothesis (Shamir): CNN is a nearest-neighbor lookup table in the manifold of “natural” images

- **Other anecdotal evidence:**
Carter, Jain, Mueller, Gifford (2020, arXiv)
Overinterpretation reveals image classification model pathologies
The way we train AI is fundamentally flawed

The process used to build most of the machine-learning models we use today can't tell if they will work in the real world or not—and that's a problem.

by Will Douglas Heaven

November 18, 2020
Skin Monitoring Apps Fail to Detect Melanomas → Experts caution that the AI-
“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.”

François Chollet, creator of Keras deep learning system and author of Deep Learning with Python
Probabilistic programming (1997-)

- Combine probability theory with the universal (Turing-equivalent) expressive power of
  - Programming languages, or
  - First-order logic
    - E.g., rules of Go = 1 page, vs 1,000,000 pages for a circuit language
- Represent any computable probability model
- Perform inference/learning for any model, data, and query
- Combine prior knowledge and data, cumulatively

Data $\rightarrow$ Learning $\rightarrow$ Knowledge

Knowledge $\rightarrow$ Data
Growth in PPL papers
Bayesian Logic (BLOG)

- A formal language for defining probability models
  - Uses the syntactic and semantic elements of first-order logic (predicates, functions, constants, logical variables)

- Possible worlds are exactly those of a first-order logical language with the same vocabulary
  - The worlds can vary in which objects they contain, how they are related, and how they are named by constants
  - => BLOG supports *existence and identity uncertainty*

- Theorem: every well-formed* BLOG model defines a unique, proper probability distribution over its possible worlds
Global monitoring for CTBT

- **Evidence**: waveforms from 150 seismic stations
- **Query**: what happened?
- **Model**: elementary 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))
North Korea, Feb 12, 2013
Towards human-level AI

Still missing:

- Real understanding of language
- Integration of learning with knowledge
- Long-range thinking at multiple levels of abstraction
- Cumulative discovery of concepts and theories

Date unpredictable
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 the neutron-induced nuclear chain reaction
Eventually (but not very soon)...

AI systems will make better real-world decisions than humans.

Turing’s point: how do we retain power over entities more powerful than us, for ever?
Standard model of AI
(and control theory, statistics, operations research, economics)

Machines whose actions can be expected to achieve their objectives
But we cannot specify objectives completely and correctly

Third wish = please undo first two wishes
Example: Social media

Objective: maximize clickthrough

- learning what people want

- modifying people to be more predictable

With incompletely or incorrectly defined objectives, better AI => worse outcomes
A new model

- Machines are *intelligent* to the extent that *their* actions can be expected to achieve *their* objectives
- Machines are *beneficial* to the extent that *their* actions can be expected to achieve *our* objectives
1. Robot goal: satisfy human preferences*
2. Robot is *uncertain* about human preferences
3. Human behavior provides evidence of preferences

The robot solves a formally defined assistance game

Optimal solutions:

> defer to human, ask permission, allow self to be switched off

It is rational for humans to build machines that solve assistance games

Better AI => better outcomes
Real-life example

You’re the robot
Your partner is the human
You have to buy your partner the perfect birthday present using money from the joint account. (You’re not sure what to get, and in past years you’ve usually got it wrong.)
Your payoff is precisely your partner’s happiness with the present

This particular case is known to be unsolvable
The off-switch problem

- A robot with a fixed 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
Extending the basic theory

Many humans
  => connections to moral philosophy, economics

Many machines
  => avoiding unanticipated strategic interactions

Real (non-rational) humans
  => connections to cognitive psychology, neuroscience

Foundations
  => rebuild each area of AI (search, planning, RL, etc.)

Applications
  => self-driving cars, digital assistants, personal robots, social media
Ongoing research: Many humans

Commonalities and differences in preferences
Trade-offs when making decisions affecting many people
• Interpersonal comparisons of preferences
• Aggregation over individuals with different beliefs
• Actions affecting population size: Thanos problem
• Preferences about others: altruism, sadism, pride, envy
One robot, many humans

How should a robot aggregate human preferences?

Harsanyi: every Pareto-optimal policy optimizes a linear combination, assuming a common prior over the future

In the general case, Pareto-optimal policies have dynamic weights proportional to whose predictions turn out to be correct

- Everyone prefers this policy because they think they are going to win

[Critch, Russell, Desai, NeurIPS 18]
Ongoing research: Real humans

- Imperfect translation of human preferences into human behavior
  - Computational limitations on human decisions
  - Emotionally driven behavior
- Preferences for autonomy, agency
- Plasticity of preferences
  - Should the AI system act on behalf of the present you or the future you?
  - Can we ensure AI systems do not manipulate human preferences?
  - Should AI systems take human preferences at face value?
    - Even if they result from indoctrination?
Summary and implications

AI has vast potential and unstoppable momentum

The standard model for AI leads to loss of human control over increasingly intelligent AI systems

Provably beneficial AI is possible and desirable

It’s not Al Ethics, it’s Al

Problems of misuse and overuse remain open