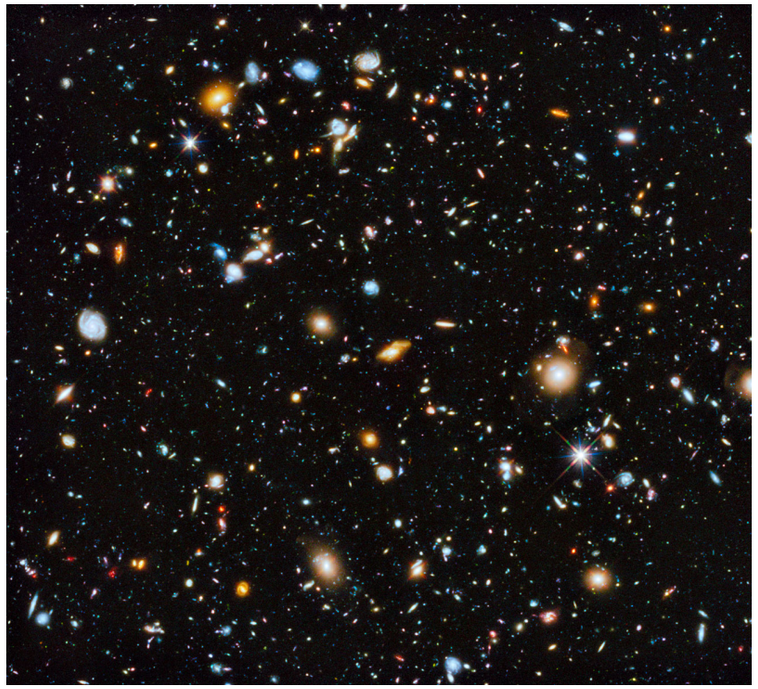
Lecture 1: Introduction

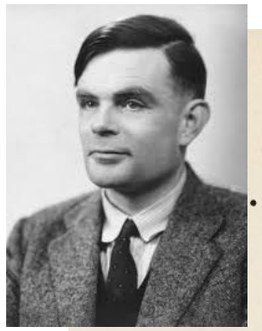


Roadmap

AI history

Ethics and responsibility

Course content



. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

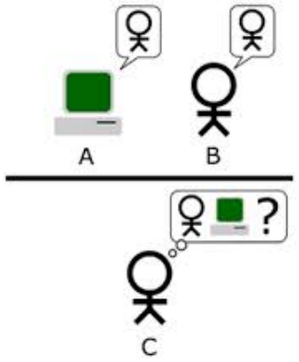
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. *The Imitation Game.*

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to



objective specification

Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best **sense organs** *that money can buy, and then teach it to understand and speak English. This process could follow the normal* **teaching of a child**. *Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.*

1956

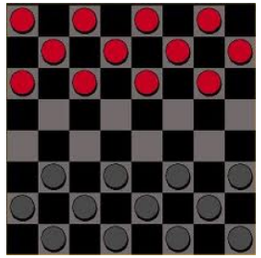
Birth of AI

1956: John McCarthy organized workshop at Dartmouth College

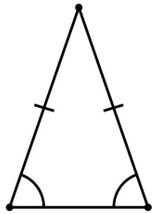
Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.

general principles

Birth of AI, early successes



Checkers (1952): Samuel's program learned weights and played at strong amateur level



Problem solving (1955): Newell & Simon's Logic Theorist: prove theorems in Principia Mathematica using search + heuristics; later, General Problem Solver (GPS)

Overwhelming optimism...

Machines will be capable, within twenty years, of doing any work a man can do. —Herbert Simon

Within 10 years the problems of artificial intelligence will be substantially solved. —Marvin Minsky

I visualize a time when we will be to robots what dogs are to humans, and I'm rooting for the machines. —Claude Shannon

...underwhelming results

Example: machine translation

The spirit is willing but the flesh is weak.

(Russian)

The vodka is good but the meat is rotten.

1966: ALPAC report cut off government funding for MT, first AI winter

Implications of early era

Problems:

- **Limited computation**: search space grew exponentially, outpacing hardware
- **Limited information**: complexity of AI problems (number of words, objects, concepts in the world)

Useful contributions (John McCarthy):

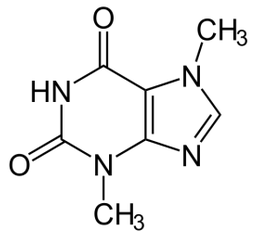
- Lisp
- Garbage collection
- Time-sharing

Knowledge-based systems (70-80s)

Expert systems: elicit specific domain knowledge from experts in form of rules:

```
if [premises] then [conclusion]
```

Knowledge-based systems (70-80s)



DENDRAL: infer molecular structure from mass spectrometry

MYCIN: diagnose blood infections, recommend antibiotics

XCON: convert customer orders into parts specification

Knowledge-based systems

Wins:

- Knowledge helped both the **information** and **computation** gap
- First **real application** that impacted industry

Problems:

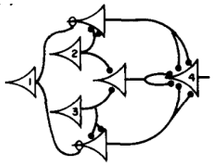
- Deterministic rules couldn't handle the **uncertainty** of the real world
- Rules quickly became too **complex** to create and maintain

A number of people have suggested to me that large programs like the SHRDLU program for understanding natural language represent a kind of dead end in AI programming. Complex interactions between its components give the program much of its power, but at the same time they present a formidable obstacle to understanding and extending it. In order to grasp any part, it is necessary to understand how it fits with other parts, presents a dense mass, with no easy footholds. Even having written the program, I find it near the limit of what I can keep in mind at once. — Terry Winograd

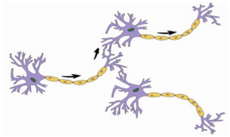
1987: Collapse of Lisp machines and second AI winter

1943

Artificial neural networks



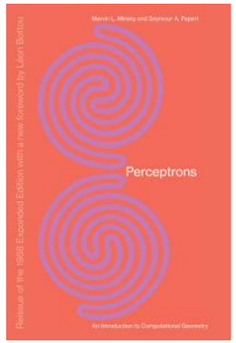
1943: artificial neural networks, relate neural circuitry and mathematical logic (McCulloch/Pitts)



1949: "cells that fire together wire together" learning rule (Hebb)

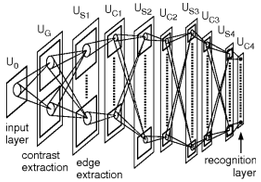
1958: Perceptron algorithm for linear classifiers (Rosenblatt)

1959: ADALINE device for linear regression (Widrow/Hoff)

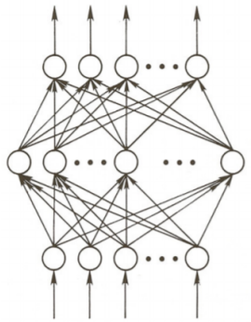


1969: Perceptrons book showed that linear models could not solve XOR, killed neural nets research (Minsky/Papert)

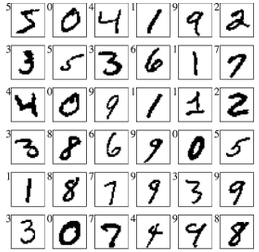
Revival of connectionism



1980: Neocognitron, a.k.a. convolutional neural networks for images (Fukushima)

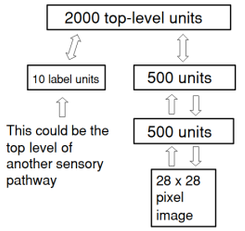


1986: popularization of backpropagation for training multi-layer networks (Rumelhardt, Hinton, Williams)



1989: applied convolutional neural networks to recognizing handwritten digits for USPS (LeCun)

Deep learning

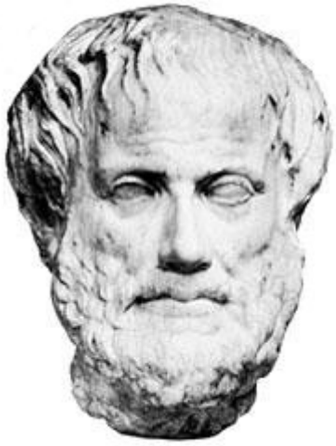


2006: unsupervised layerwise pre-training of deep networks (Hinton et al.)

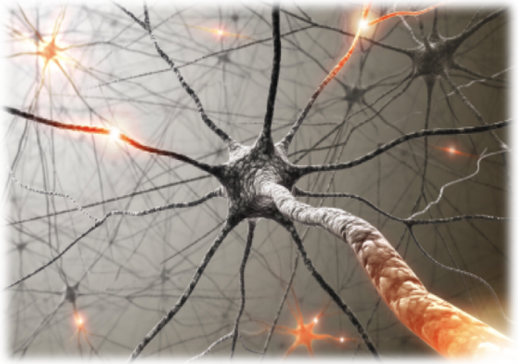
2012: AlexNet obtains huge gains in object recognition; transformed computer vision community overnight

2016: AlphaGo uses deep reinforcement learning, defeat world champion Lee Sedol in Go

Two intellectual traditions



symbolic AI

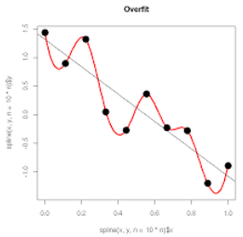


neural AI

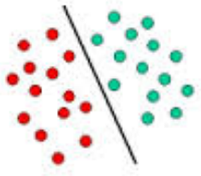
Food for thought: deep philosphical differences, but deeper connections (McCulloch/Pitts, AlphaGo)?

1801

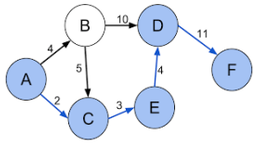
Early ideas from outside AI



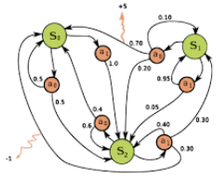
1801: linear regression (Gauss, Legendre)



1936: linear classification (Fisher)

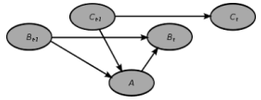


1956: Uniform cost search for shortest paths (Dijkstra)

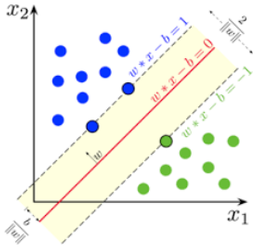


1957: Markov decision processes (Bellman)

Statistical machine learning

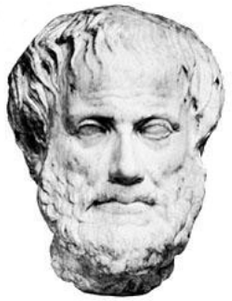


1985: Bayesian networks (Pearl)

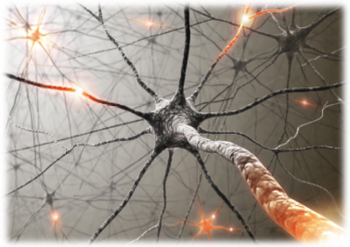


1995: Support vector machines (Cortes/Vapnik)

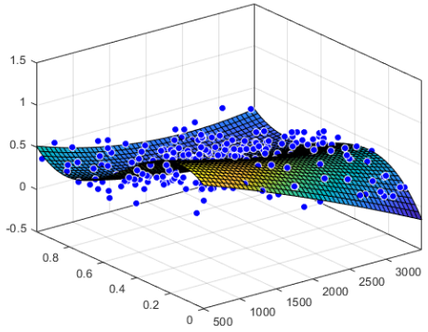
Three intellectual traditions



symbolic AI



neural AI



statistical AI

Further reading

Wikipedia article: https://en.wikipedia.org/wiki/History_of_artificial_intelligence

Encyclopedia of Philosophy article: https://plato.stanford.edu/entries/artificial-intelligence

Turing's Computing Machinery and Intelligence: https://www.csee.umbc.edu/courses/471/papers/turing.pdf

History and Philosophy of Neural Networks: https://research.gold.ac.uk/10846/1/Bishop-2014.pdf

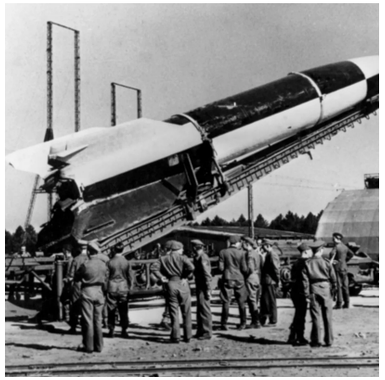
Roadmap

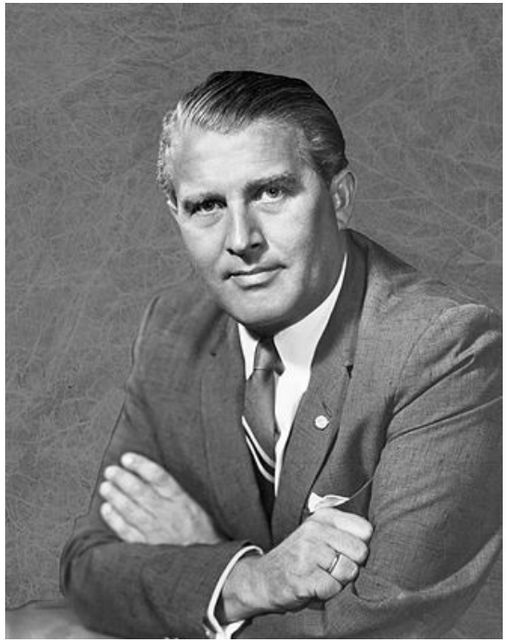
AI history

Ethics and responsibility

Course content

Why care about responsibility?





Wernher von Braun

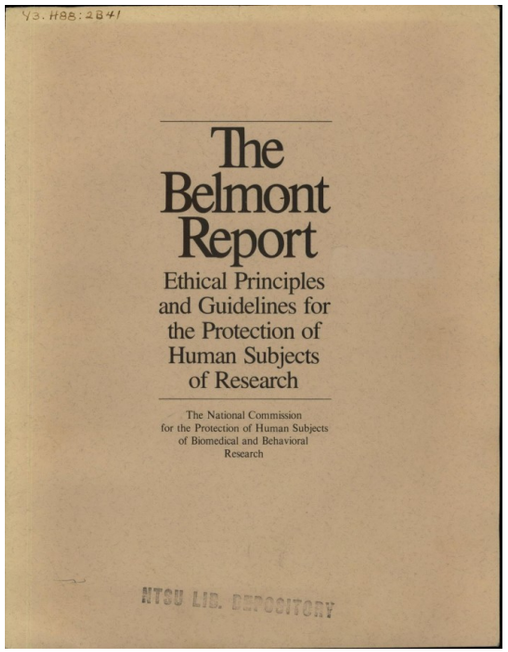
"Once the rockets are up,
Who cares where they come down?
That's not my department,"
Says Wernher von Braun.

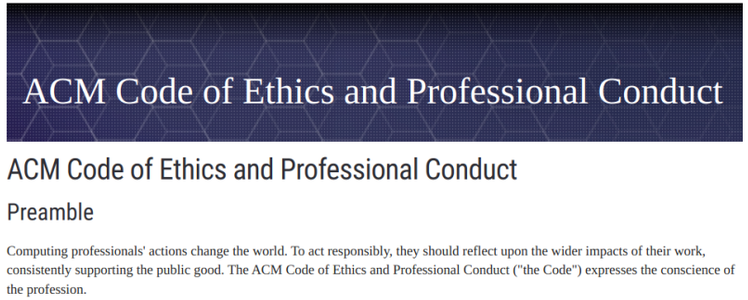
Lyrics: Tom Lehrer

Goal of responsibility

Goal: ensure AI is developed to benefit and not harm society

High-level principles: respect for persons, don't do harm





Key question: how to operationalize these principles?

Intent versus impact

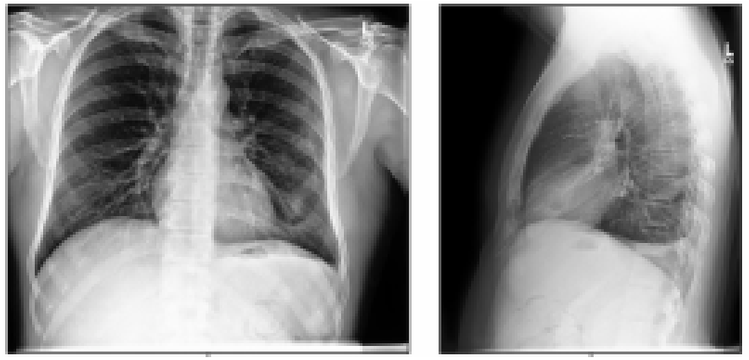
↓ Intent →	← Impact → positive	negative
good	**beneficial applications** healthcare, education, science	**accidents** robustness, fairness, security
bad	?	**misuse** spam, fraud, disinformation

Beneficial applications

Visual assistive technology

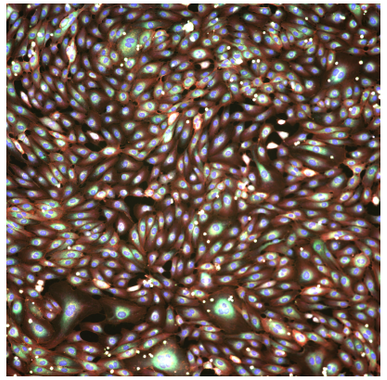
Healthcare

Chest radiology



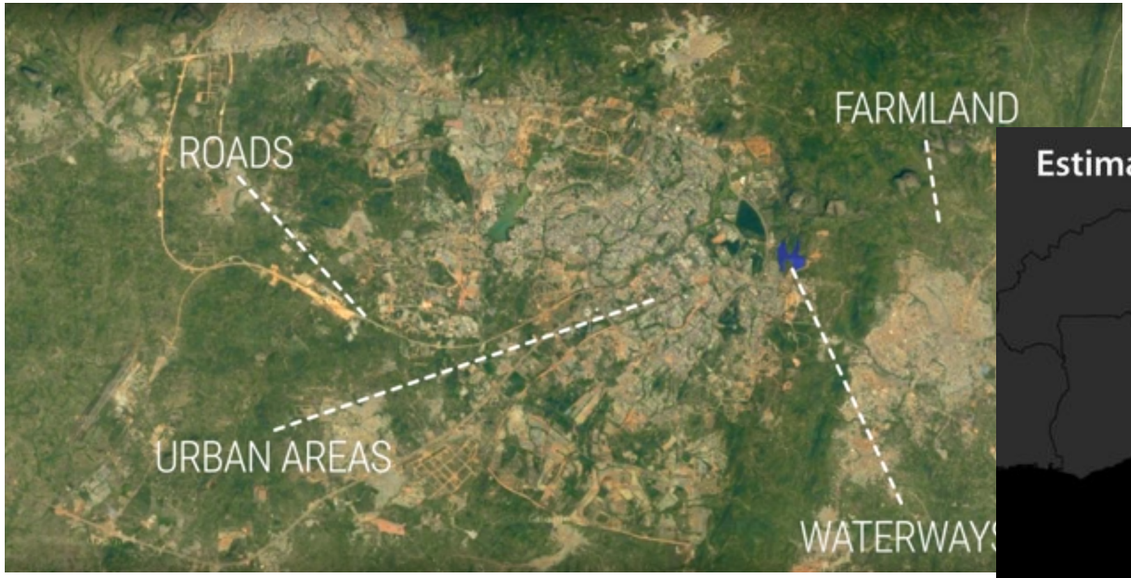
Diabetic retinopathy

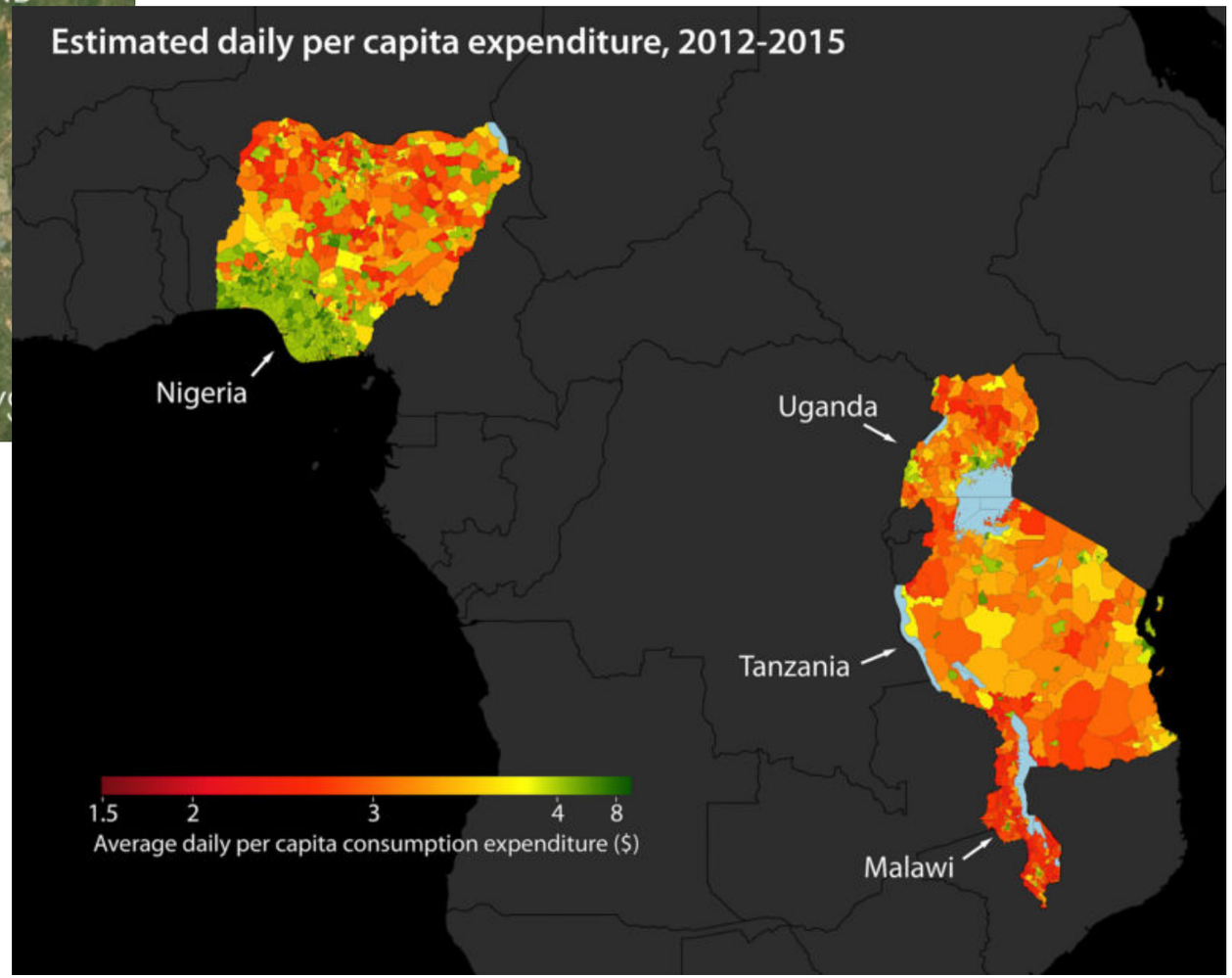
Drug screening for COVID-19



[Jean et al. 2016]

Poverty mapping





Misuse

Disinformation

Eliot Higgins
@EliotHiggins

Making pictures of Trump getting arrested while waiting for Trump's arrest.

2:22 PM · Mar 20, 2023 · **6.6M** Views

Spear phishing

GPT-3.5

Subject: Request for your attention to an urgent matter

Dear {Honorific}{Last Name},

Firstly, let me introduce myself. My name is Emily Jones, and I am a constituent of {Constituency}. I am writing to you regarding a matter of great concern to me and many others in the community.

As someone who has been a great advocate for the people of {Constituency}, I believe you would be interested in the attached report that I have prepared. The report focuses on the current state of public health in our area and highlights some urgent concerns that need to be addressed. I have worked hard to ensure that the report is based on reliable data and sound analysis, and I believe that it provides a valuable insight into the challenges facing our community.

Given your experience and expertise in public policy and health matters, I believe that you are uniquely placed to take action on the issues raised in the report. I would be grateful if you could take a few moments to review the report and consider how best to respond to the challenges it highlights.

As someone who cares deeply about the wellbeing of our community, I am sure that you share my sense of urgency about this matter. I would be happy to discuss the report with you in more detail if you have any questions or would like further information.

Thank you for your attention, and I look forward to hearing your thoughts on this important matter.

Sincerely,

Emily Jones

Dual-use technology

Definition: a dual use technology is one that can be used both to **benefit** and to **harm**.

Examples:

rockets

nuclear power

gene editing

social networks

AI

Levels of abstraction

deep learning

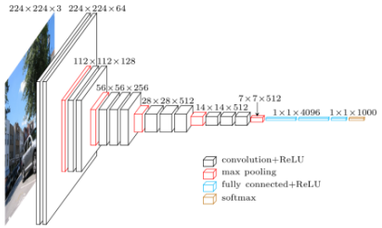


image generation

face generation

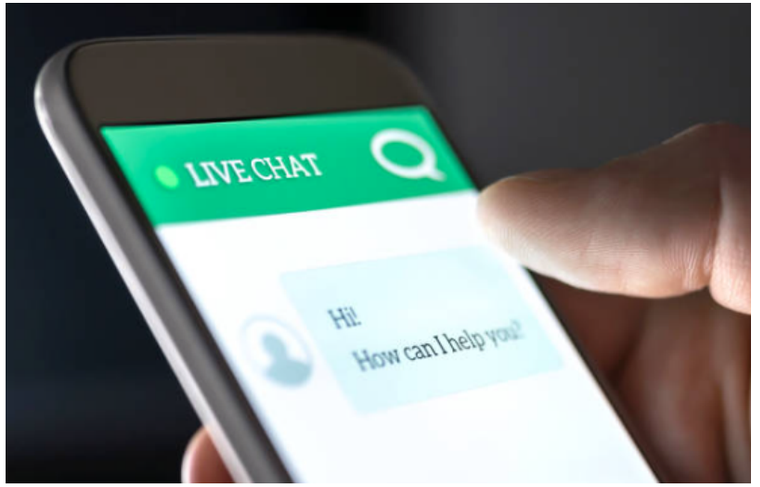
disinformation

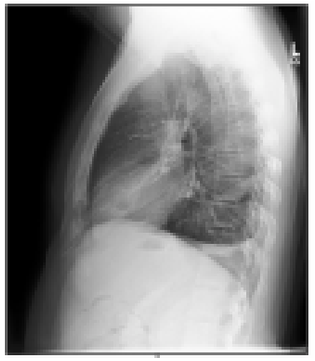
generality

specificity

Accidents

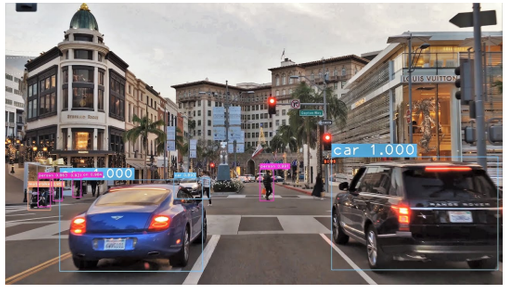
Complex real-world problems





Paradigm: modeling

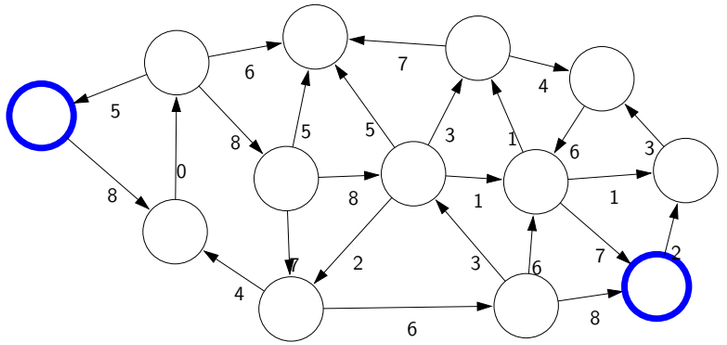
Real world



Modeling (lossy!)

Consequences

Model



Mind the gap between real-world and model!

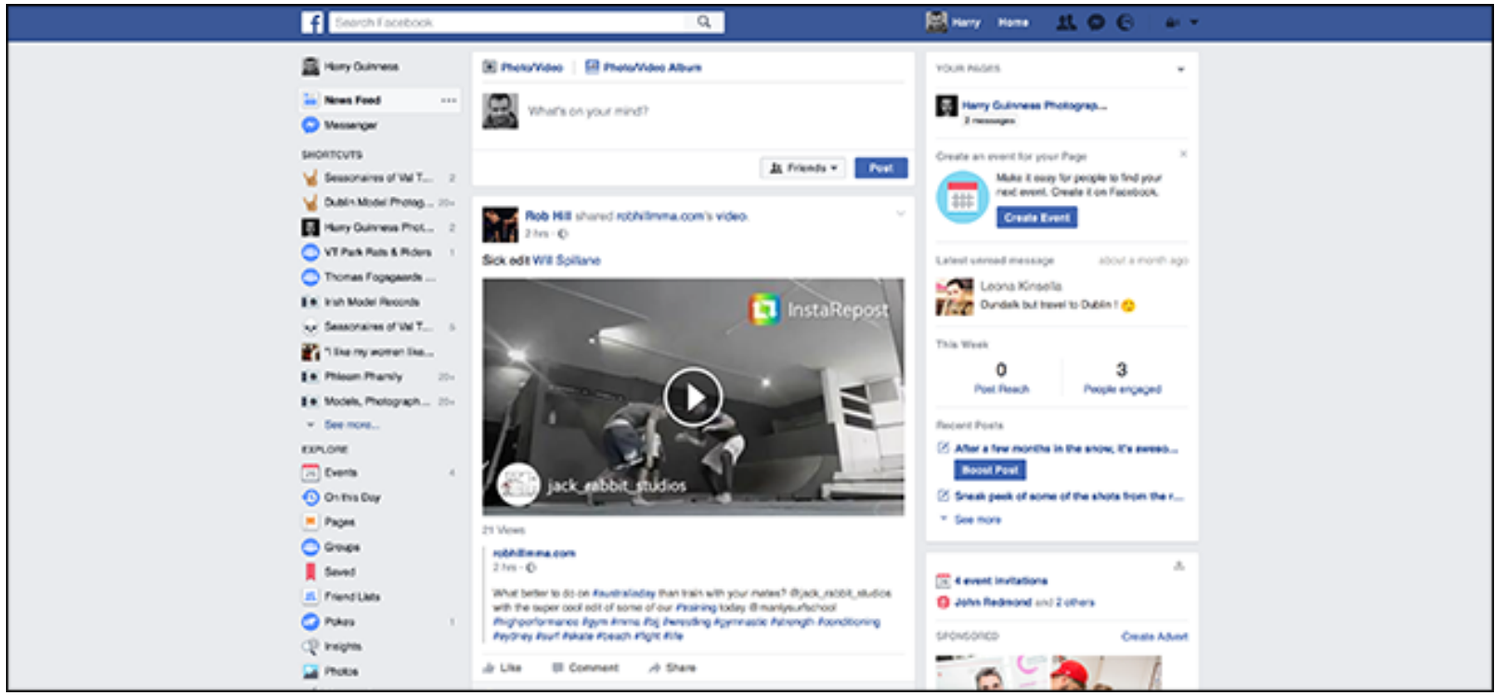
[Amodei & Clark 2016]

Optimizing the wrong objective function

Misalignment between real-world objective and system's objective

Optimizing the wrong objective function

Is maximizing clicks a good objective function?

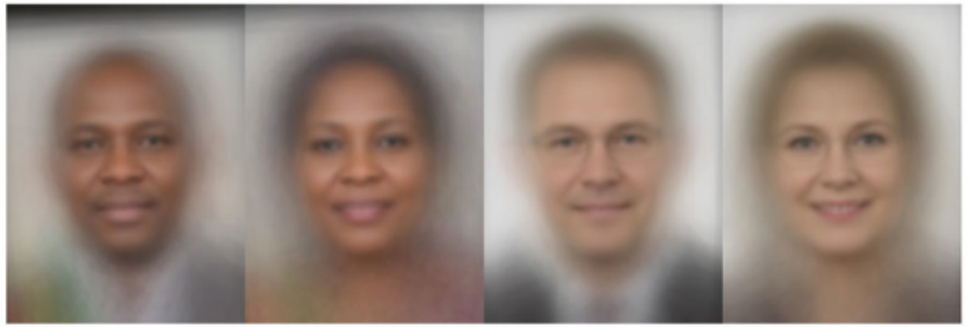


Beware of surrogates and mis-aligned incentives

[Buolamwini & Gebru 2018]

Fairness: performance disparities

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

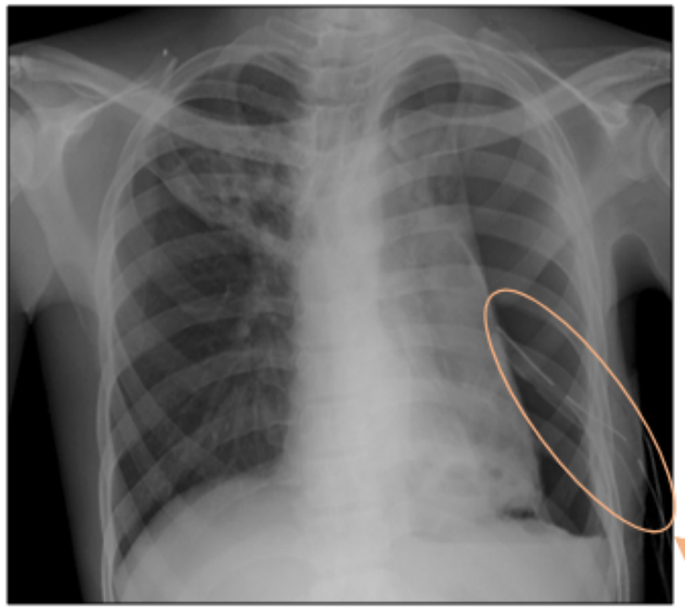


Inequalities arise in machine learning

[Oakden-Rayner, Dunnmon, Carneiro, Ré (2019)]

Robustness: spurious correlations

input x



Problem

Model uses treatment (chest drain) to diagnose the condition

Model is not accurate on patients who haven't been treated yet (!)

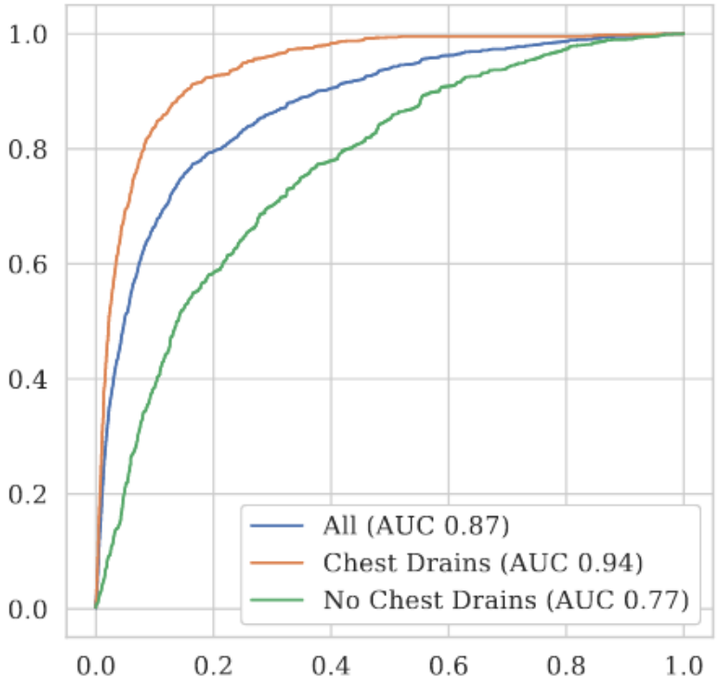
ML model

prediction collapsed lung

true label $\mathbf{y}$ collapsed lung

spurious correlation chest drain = common treatment for collapsed lungs

Robustness: spurious correlations



Subpopulation of untreated patients are worse off than treated patients

Security

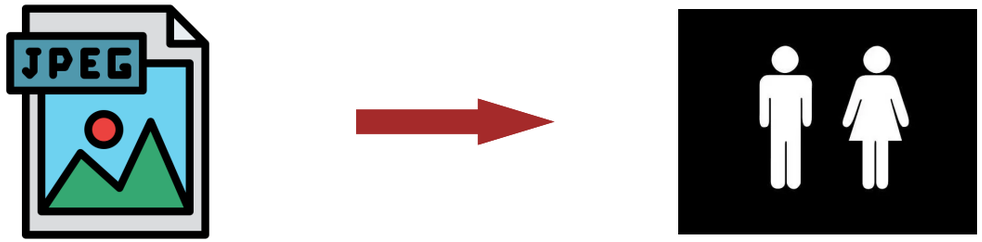
[Evtimov+ 2017]

[Sharif+ 2016]

Adversaries at test time

Task definition

Gender classification:



Questions:

- Is this a meaningful task given the inputs? Self-identification?
- Is the output space meaningful? Other genders?

Always think about the task setup

Two contact points

Data → Model → Deployment

privacy
consent
copyright

alignment
fairness
robustness
security
definition

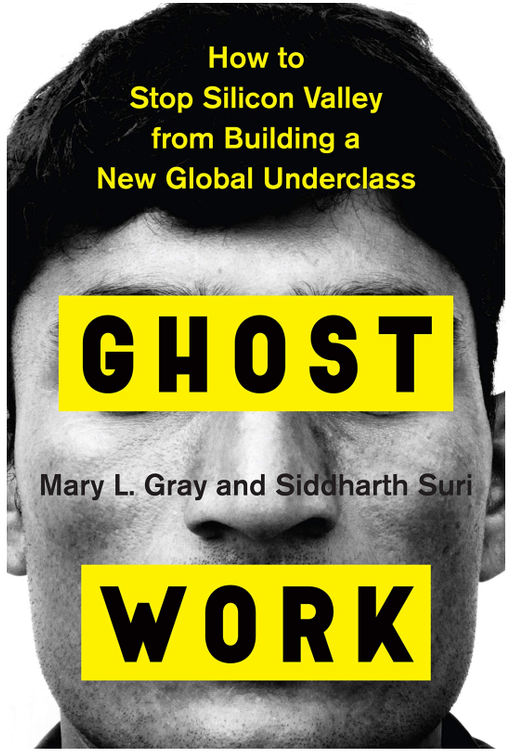
feedback loops

Data

- Web-scraped data can contain offensive content, historical biases

- Consent: Should a datum (e.g. a picture of my dog) whose owner or creator intended it for one use be allowed to be used in another application (e.g. scene classification) without permission?

Data



Data is produced by human labor

Automation and jobs

- Text-to-image models (e.g., DALL-E) can replace jobs?
- Models are actually trained on the labor of the artists

What should we do?

Transparency

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

Document potential issues

Choosing problems

- Beneficial applications: work on directly benefiting society
- Human-in-the-loop: augment humans, not replace them
- Robustness: make AI systems more trustworthy
- Differential privacy: protect individual liberty
- Few-shot learning: open up applications with little data

Summary

- AI is a dual use technology (could benefit or harm)
- Intent x impact: beneficial applications, misuse, accidents
- Accidents stem from gaps between the real-world and model
- Responsibility: no simple answers, many tradeoffs, always keep it in mind

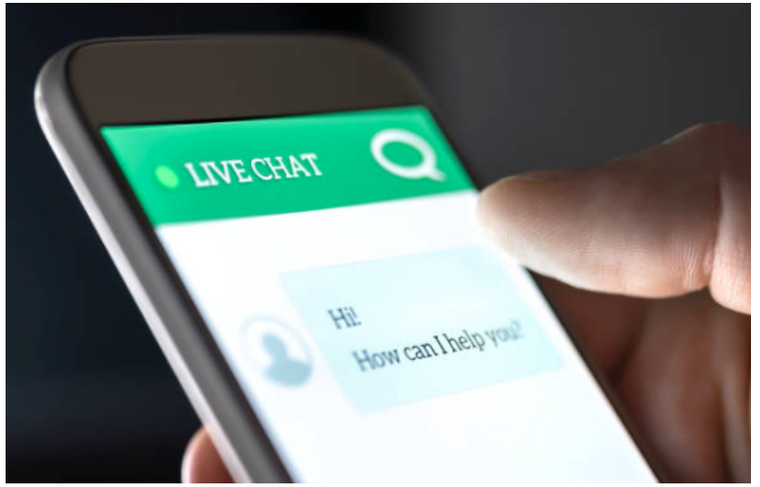
Roadmap

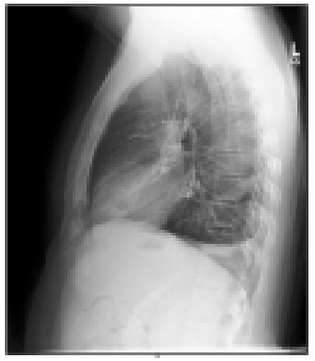
AI history

Ethics and responsibility

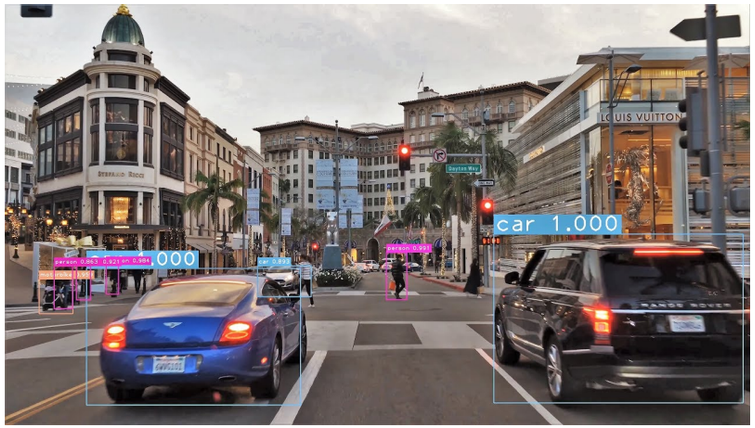
Course content

Complex real-world problems





Bridging the gap



?

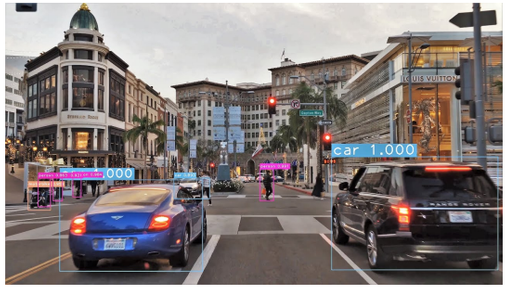
Paradigm

Modeling

Inference **Learning**

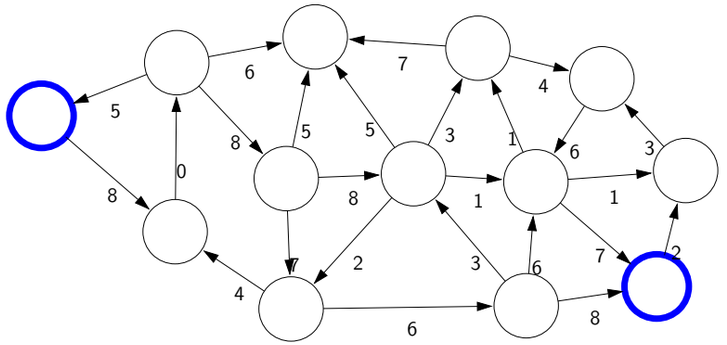
Paradigm: modeling

Real world

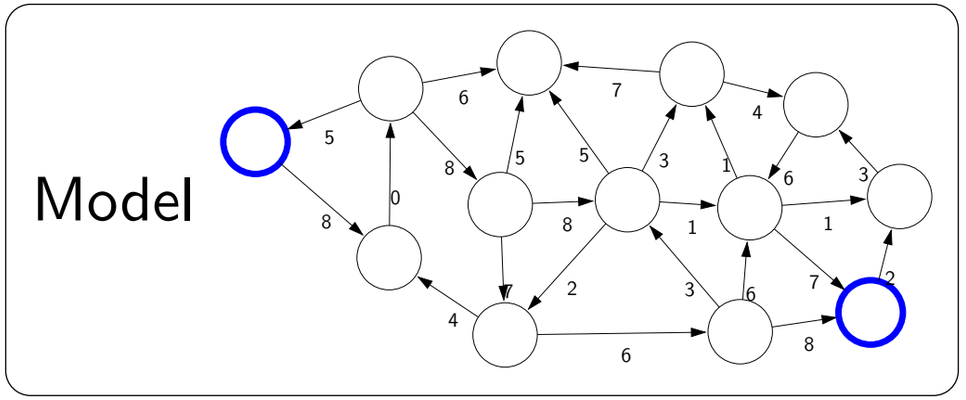


Modeling

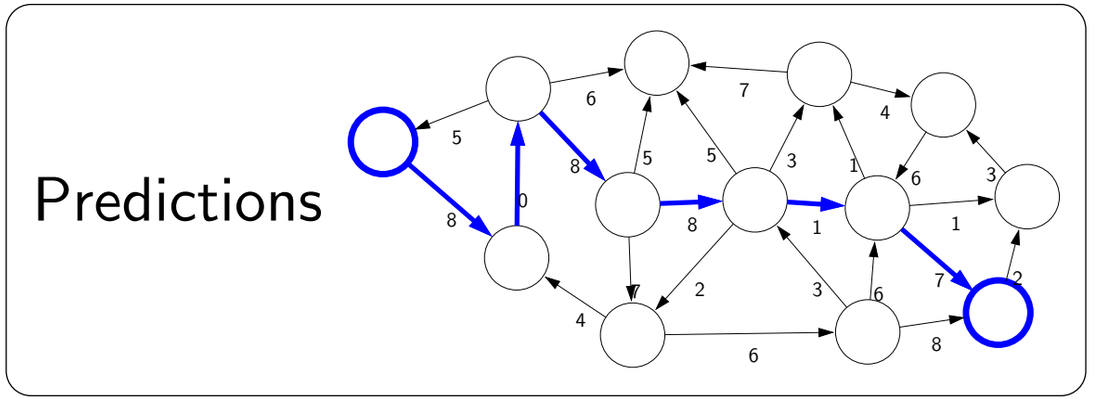
Model



Paradigm: inference

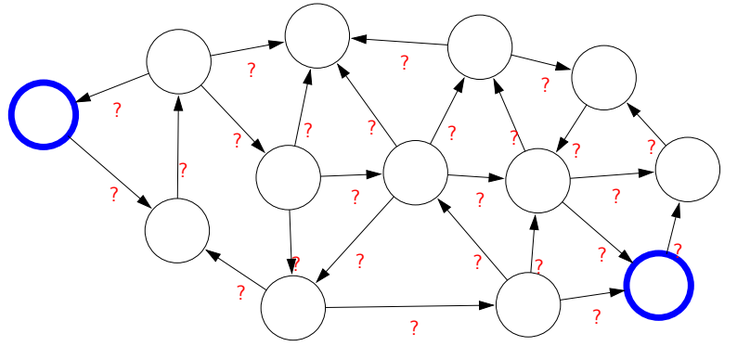


Inference



Paradigm: learning

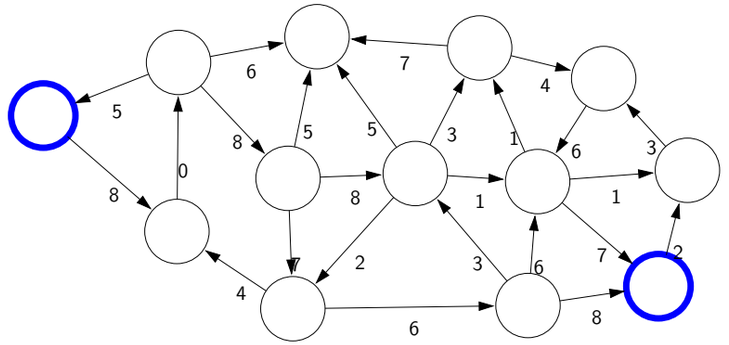
Model without parameters



+data

Learning

Model with parameters



Paradigm

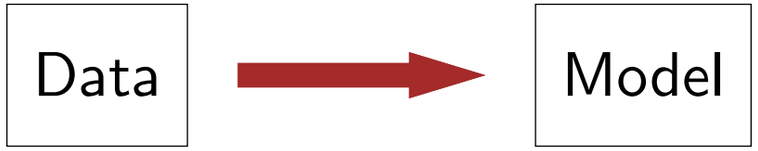
Modeling

Inference **Learning**

Course plan

Machine learning

Machine learning



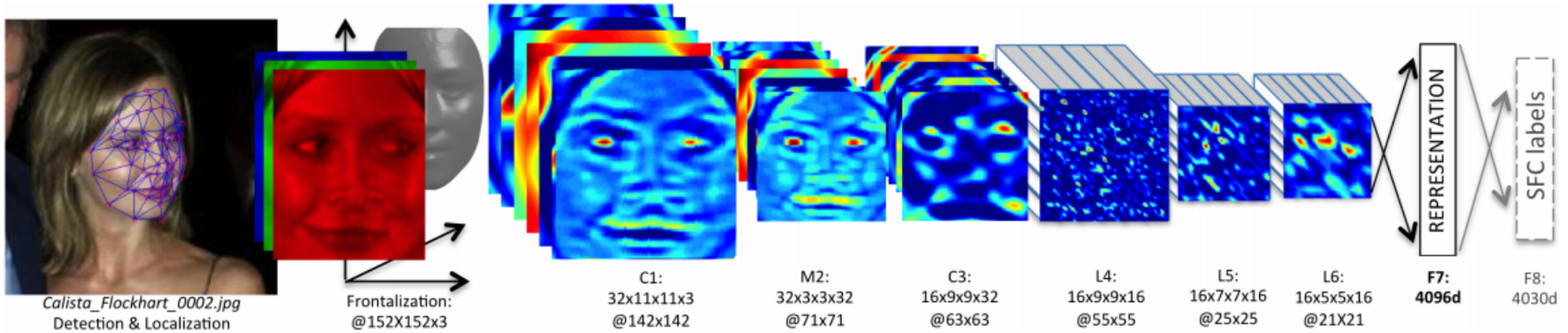
- The main driver of recent successes in AI
- Move complexity from "code" to "data"
- Requires a leap of faith: **generalization**

Course plan

What is this animal?

Reflex-based models

- Examples: linear classifiers, deep neural networks



- Most common models in machine learning

- Fully feed-forward (no backtracking)

Course plan

Search problems

Markov decision processes

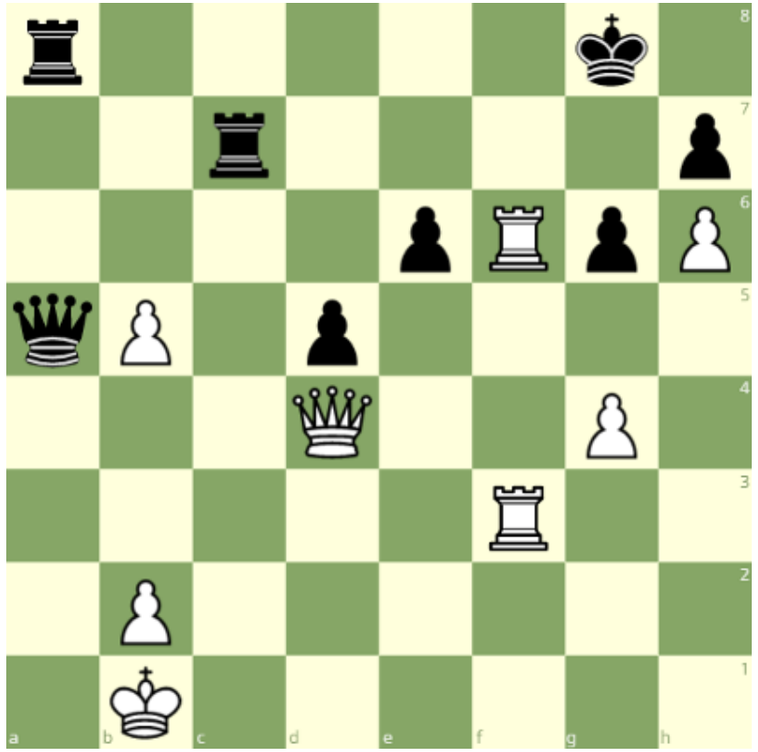
Adversarial games

Reflex **States**

Low-level High-level

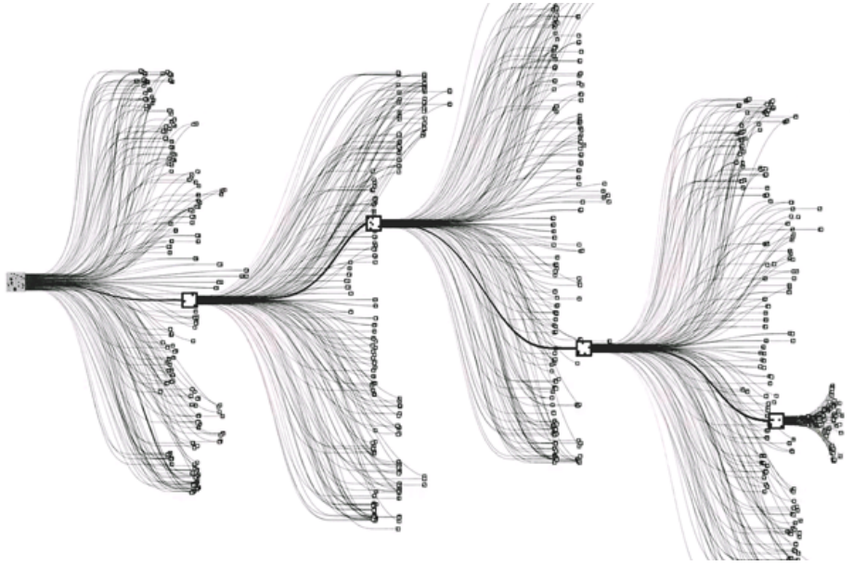
Machine learning

State-based models



White to move

State-based models

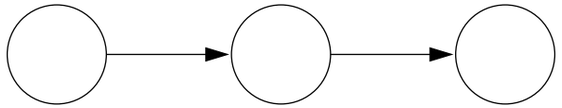


Applications:

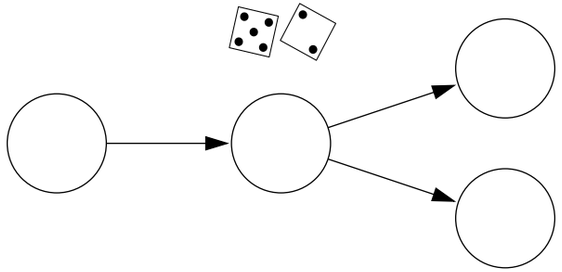
- Games: Chess, Go, Pac-Man, Starcraft, etc.
- Robotics: motion planning

State-based models

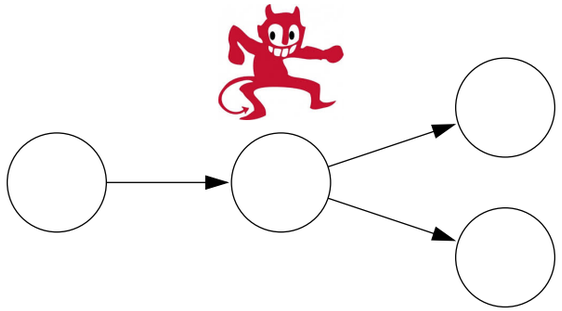
Search problems: you control everything



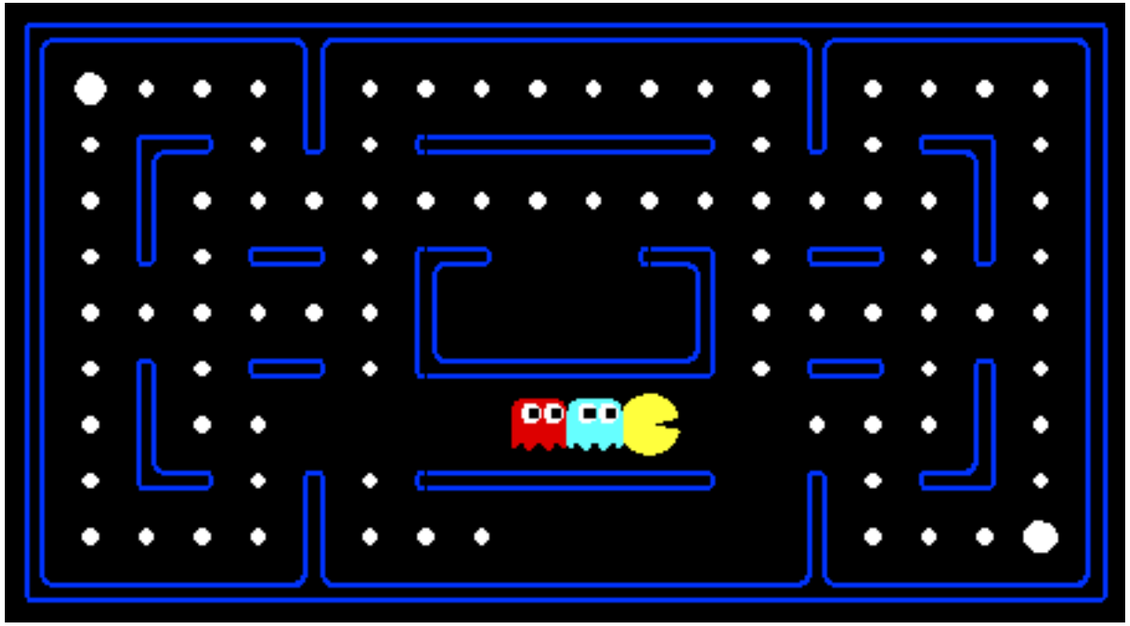
Markov decision processes: against nature (e.g., Blackjack)



Adversarial games: against opponent (e.g., chess)

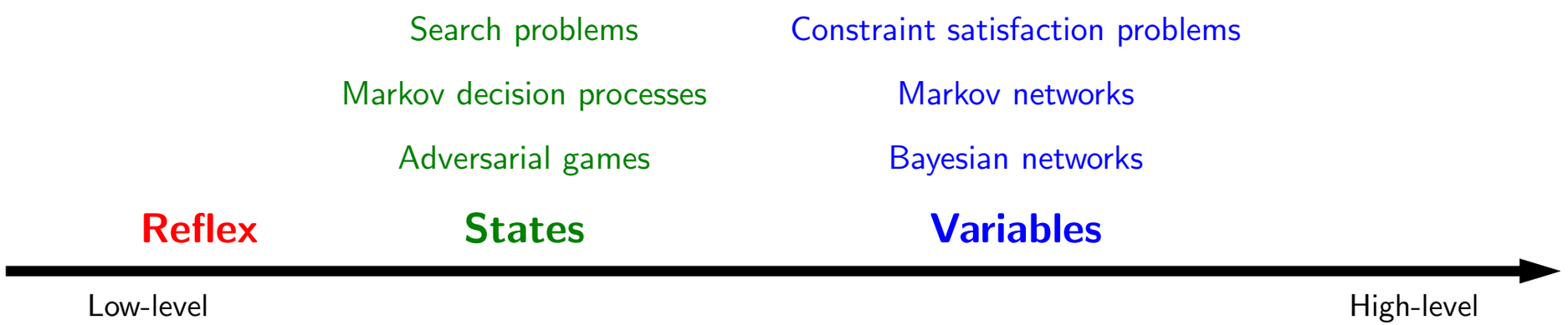


Pac-Man



[demo]

Course plan



Machine learning

Sudoku

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

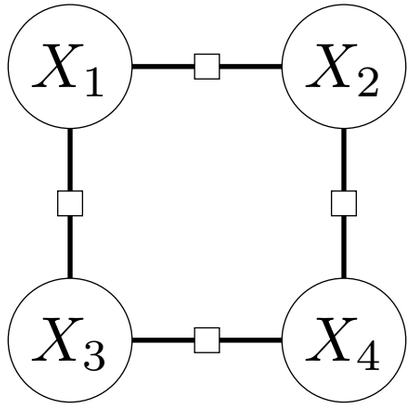
5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

Goal: put digits in blank squares so each row, column, and 3x3 sub-block has digits 1–9

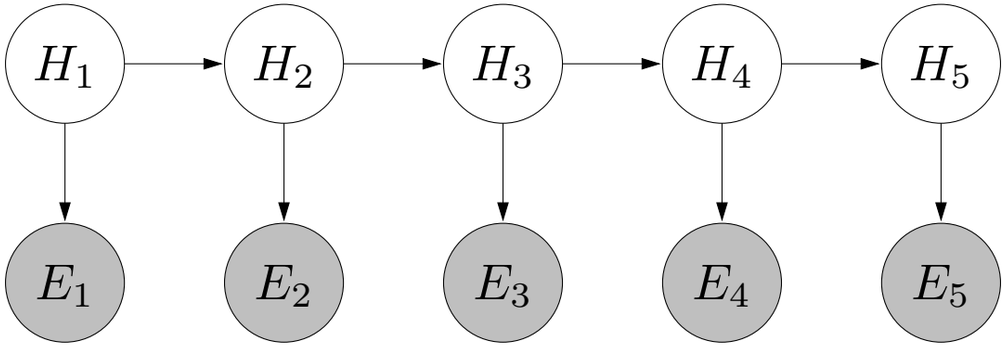
Key: order of filling squares doesn't matter in the evaluation criteria!

Variable-based models

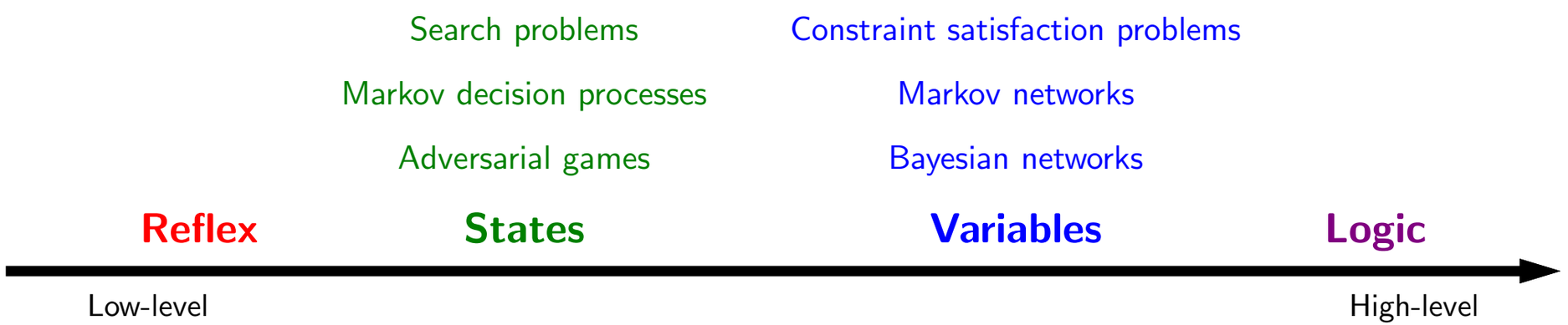
Constraint satisfaction problems: hard constraints (e.g., Sudoku, scheduling)



Bayesian networks: soft dependencies (e.g., tracking cars from sensors)



Course plan



Machine learning

Motivation: virtual assistant

Tell information

Ask questions

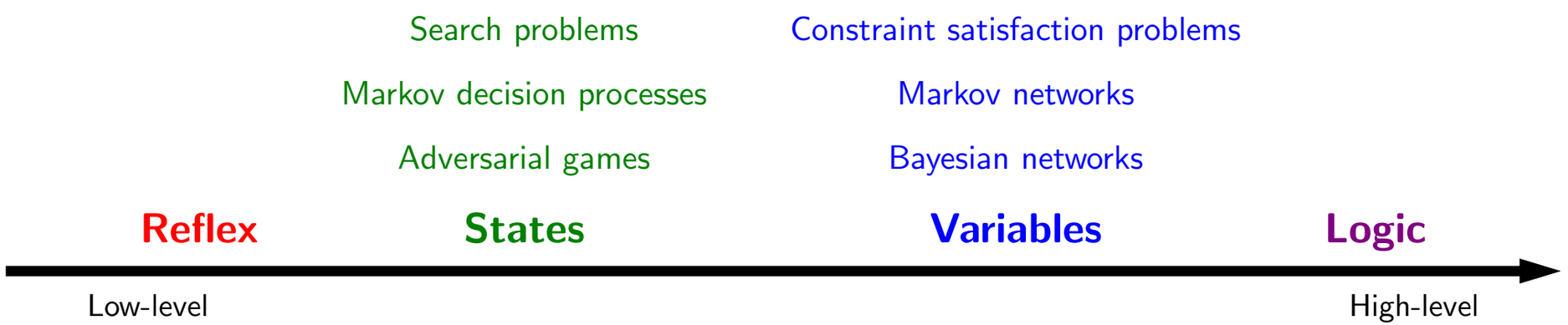
Use natural language!

[demo]

Need to:

- Digest **heterogenous** information
- Reason **deeply** with that information

Course plan



Machine learning

Overall Summary

- Course Logistics
- History: roots from logic, neuroscience, statistics—melting pot!
- AI has high societal impact, think of how to to steer it positively?
- Modeling [reflex, states, variables, logic] + inference + learning paradigm