Eric Horvitz
Microsoft

*AI and the Open World*

Chair: Christopher Bishop, Microsoft
INTRODUCING

MSR AI
Pool talent from multiple areas of AI
Key aspirations

Attain more general intelligence

Master human-AI collaboration

Lead with insights on AI, people, and society
Key aspirations

- Attain more general intelligence
- Master human-AI collaboration
- Lead with insights on AI, people, and society
Attain more general intelligence

- Wedges of competency ➔ Integrative intelligence
- Combine competencies into symphonies of intelligence
- Principles of more general artificial intelligence
Key aspirations

- Attain more general intelligence
- Master human-AI collaboration
- Lead with insights on AI, people, and society
Augment and extend human capabilities

Leverage and extend results from cognitive psychology
Augment and extend human capabilities

Designs for mix of Initiatives

Human cognition

Machine intelligence

Machine learning & inference

Design, learning, optimization
Key aspirations

- Attain more general intelligence
- Master human-AI collaboration
- Lead with insights on AI, people, and society
AI, people, and society

Directions

- Trustworthiness & safety
- Fairness & accuracy
- Transparency & explanation
- Explorations of human-AI relationship
AI and the Open World
Lab environment
Open world
Lab environment

Open world

Competence

Interaction

Ethical, legal, social, societal influences
Lab environment

Open world

Competence
Interaction

Ethical, legal, social, societal influences
Qualification problem

All preconditions?

Ramification problem

All effects of action?
Lab environment

Framing

Data

Fidelity

Open world

Representational

Inferential
Knowing that you do not know is the best.

Not knowing that you do not know is an illness.

- Laozi, 500-600 BCE
Models of inference reliability

Toyama & H. 2000
Inference reliability $\Rightarrow$ Robust portfolios

Unmodeled situations

- facing away
- jolted camera
- distraction
- lights out

Toyama & H. 2000
Learn about abilities & failures

Successes & failures

Performance

Confidence

$p(\text{fail} | E, t)$

Deep learning about deep learning performance

Quality score $[0,1]$

$$s = \frac{e^{W \cdot f}}{1 + e^{W \cdot f}}$$

Caption:

*a man holding a tennis racquet on a tennis court*

Fang, et al., 2015
Grappling with Open-World Complexity

Reliable predictions of performance: *Known unknowns*
Grappling with Open-World Complexity

Reliable predictions of performance: *Known unknowns*

Challenge of *unknown unknowns*
Unknown unknowns

Directions

Expanded real-world testing

Algorithmic portfolios

Failsafe designs

People + machines
Identifying classifier blindspots

Conceptual incompleteness

$x = (f_1, \ldots, f_k)$

How to define & search regions of data space?
How to trade exploration and exploitation?

Partition space by attributes

White Dogs | White Cats | Brown Cats | Brown Dogs

Identifying classifier blindspots

\[ x = (f_1, \ldots, f_K) \]

Wrong label high confidence

Directions

Transfer learning

Learn from rich simulations

Learn generative models
Transfer learning opportunity

Site-specific data
- Observations, definitions
- Patients, prevalencies
- Covariate dependencies

A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions

Jenna Wiens, John Guttag, Eric Horvitz
Embedded deep transfer learning

Less data with better features

ImageNet 1000, 1M photos

Cut off top layer
Embedded deep transfer learning

Less data with better features

ImageNet 1000, 1M photos

Cut off top layer
Simulated Environments

Trillions of sessions in complex scenarios
Learn & evaluate core competencies
Learn to optimize action plans
Leveraging rich simulations

Hybrid Deep Stereo

Depth Image

Mapping

Next actions

Plans

D. Dey, S. Sinha, S. Shah, A. Kapoor
Leveraging rich simulations

Hybrid Deep Stereo

Depth Image

Mapping

Planning

Next actions

Plans

D. Dey, S. Sinha, S. Shah, A. Kapoor
Learn expressive generative models

Generalize from minimal training sets

Harness physics
Learning generative models

**Multilevel variational autoencoder**
Learn disentangled representations
Groups of observations $\rightarrow$ latent models

Vary style  Vary ID

Smooth control over learned latent space

D. Buchacourt, R. Tomioka, S. Nowozin, 2017
Inject physics to disentangle & generalize

Kulkarni, Whitney, Kohli & Tenenbaum, 2015
Inject physics to disentangle & generalize

Kulkarni, Whitney, Kohli & Tenenbaum, 2015
Directions

Potential biases in data, performance

Best practices & processes

Industry and beyond
Bernard Parker: rated high risk

Dylan Fugett: rated low risk.

There's some sort of biased rating system here.
Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems

Ayanna Howard1, Cha Zhang2, Eric Horvitz2
March 2017

Machine learning “contact lens” for children
Aether Advisory Panel
AI and Ethics in Engineering and Research

Partnership on AI
to benefit people and society
Scientific foundations

Engineering practices

AI, people, and society

Much to do
Thank you