AI, People, and the Open World

Eric Horvitz
Technical Fellow and Director
Microsoft Research

August 16, 2017
Artificial Intelligence

“...to find how to make machines...solve kinds of problems now reserved for humans...” (1955)
Artificial Intelligence

- Perception
- Learning
- Reasoning
- Natural language
Artificial Intelligence

- Reinforcement Learning
- Deep Learning
- Probabilistic Graphical Models
- Computer Vision
- Robotics
- Knowledge Representation

Perception - Learning - Reasoning - Natural language

- Speech and dialog
- Decisions and plans
- Algorithmic Game Theory

Multiple subdisciplines and research communities
Speech recognition

5.8% human error rate
Switchboard challenge

Speech-recognition word-error rate, %

Human-level

Sources: Microsoft: research papers
Vision

152 layers

RESNET
Reading comprehension
Answer questions about information on Wikipedia

SQuAD challenge
New capabilities

I was wondering what you are going to do later?
Me preguntaba lo que vas a hacer después?

Going to the pub, do you want to join us? I think you met some of the team at the party in April.
Al pub, ¿quieres unirte a nosotros? Creo que conociste a algunos miembros del equipo en la fiesta en abril.

Can I join you guys after my meeting?
¿Puedo unirme a ustedes después de mi reunión?
Lab environment → Open world

Competence

AI & People

Ethical, legal, societal influences
Qualification problem
All preconditions?

Ramification problem
All effects of action?
Lab environment

Open world

Framing
Data
Fidelity
Representation
Inference
Knowing that you do not know is the best.

Not knowing that you do not know is an illness.

- Laozi, 500-600 BCE
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*
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

real-world concepts

wrong label
high confidence

Identifying classifier blindspots

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

Identifying classifier blindspots

Partition space by attributes

White Dogs  |  White Cats  |  Brown Cats  |  Brown Dogs

training data

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

\( M \)

Wrong label with high confidence

Data scarcity

Directions

Transfer learning

Learn from rich simulations

Learn generative models
Transfer learning opportunity

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

Predict of risk of infection

A: Community hosp: 10k pts/yr
B: Acute care & teaching: 15k/yr
C: Major teaching & research: 40k/yr

Transfer learning opportunity

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

Hospital A
Hospital B
Hospital C

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

Jenna Wiens,¹ John Guttag,¹ Eric Horvitz²
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

M. Gabel, R. Caruana, M. Philipose, O. Dekel
Embedded deep transfer learning

Less data with better features

ImageNet 1000, 1M photos

Cut off top layer

M. Gabel, R. Caruana, M. Philipose, O. Dekel
Trillions of sessions in complex scenarios
Learn & evaluate core competencies
Learn to optimize action plans
Leveraging rich simulations

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

CNN

Depth Image

Mapping

Planning

Next actions

Map

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
Inject physics to disentangle & generalize

Kulkarni, Whitney, Kohli & Tenenbaum, 2015
Concerns

- AI attack surfaces
- Adversarial machine learning
- Self-modification
Attacks on AI Systems

“Adverserial machine learning”

Goodfellow, et al.
Papernot, et al.
Adversarial Attacks & Self-Modification

Environment

AI system

State → Perception

Reward → Reinforcement

Action

e.g., see: Amodei, Olah, et al., 2016
Adversarial Attacks & Self-Modification

e.g., see: Amodei, Olah, et al., 2016
Adversarial Attacks & Self-Modification

Adversary → Environment → AI system → Reward → State → Perception → Reinforcement → Action

e.g., see: Amodei, Olah, et al., 2016
Adversarial Attacks & Self-Modification

Run-time verification
Static analysis

Ensure isolation * identify meddling * ensure operational faithfulness

Amodei, Olah, et al., 2016
H. 2016
AI & People

Directions

Models of people & tasks

Models of complementarity

Coordination of initiative
Models of people & tasks

Actions, services

Predictions about needs, goals
Models of world & people

Predictions about world

Predictions about user beliefs

Actions

H. Barry, 1995
H. , Apacible, Sarin, Liao, 2005
Models of world & people

H. Barry, 1995
Complementarity

Leverage and extend results from cognitive psychology
Complementarity

Leverage and extend results from cognitive psychology
Complementarity

Leverage and extend results from cognitive psychology
Complementarity

Identifying metastatic breast cancer
(Camelyon Grand Challenge 2016)

Human is superior
Error: 3.4%

AI + Expert: 0.5%

85% reduction in errors.
Complementarity

Label galaxies in Sloan Digital Sky Survey

(Galaxy Zoo)

Kamar, Hacker, H., AAMAS 2012
Complementarity

Label galaxies in Sloan Digital Sky Survey
(Galaxy Zoo)

~453 features

Machine perception

Human perception

Machine learning & inference

Kamar, Hacker, H., AAMAS 2012
Complementarity

Full accuracy: 47% of human effort
95% accuracy: 23% of human effort

Machine perception

~453 features

Ideal fusion, stopping

Human perception

Machine learning & inference

Kamar, Hacker, H., AAMAS 2012
Designs for mix of initiatives

Human cognition

Machine intelligence

Machine learning & inference

Design, learning, optimization
Recognizing intention

- Reach needle #22
- Position #165
- Insert #162
- Left transfer #119
- Right transfer #37
- Pull #160
- Orient #48
- Tighten suture #23
- Loosening #4
- Dropping #39

C.E. Reiley, et al.
Coordination of initiative

- Research da Vinci application - not for human use -

Move tool to start

HMC demo - Nicolas Padoy, CIRL, JHU

Padoy & Hager. ICRA 2011
van den Berg, et al, ICRA, 2010
Multiple influences and concerns

- Trustworthiness and safety
- Fairness, accuracy, transparency
- Ethical and legal aspects of autonomy
- Jobs and economy
Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
Bernard Parker: rated high risk

Dylan Fugett: rated low risk.
Detection Result:
5 faces detected

JSON:
[
  {
    "faceRectangle": {
      "left": 488,
      "top": 263,
      "width": 148,
      "height": 148
    },
    "scores": {
      "anger": 9.075572e-13,
      "contempt": 7.048959e-9,
      "disgust": 1.02152783e-11,
      "fear": 1.778957e-14,
      "happiness": 0.99999999,
      "neutral": 1.31694478e-7,
      "sadness": 6.04054263e-12,
      "surprise": 3.92249462e-11
    }
  }
]
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

A. Howard, C. Zhang, H., 2017
Corporate & community responsibility

Aether Advisory Panel
AI and Ethics in Engineering and Research

Partnership on AI to benefit people and society
Science & engineering

Human-AI collaboration

AI, people, and society

*Much to do*