Human-Centered AI Systems

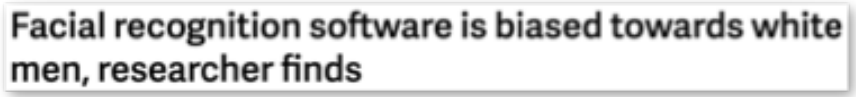
SEMLA

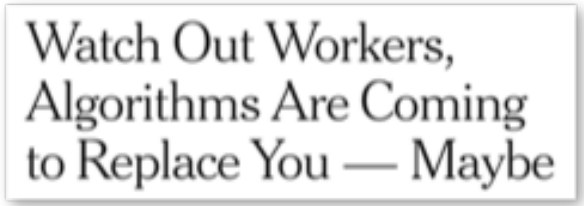
MAY 23, 2019; MONTREAL

Steve Eglash
Director of Research
Human-Centered AI Institute, Stanford University

AI: From toolbox to pervasive societal impact

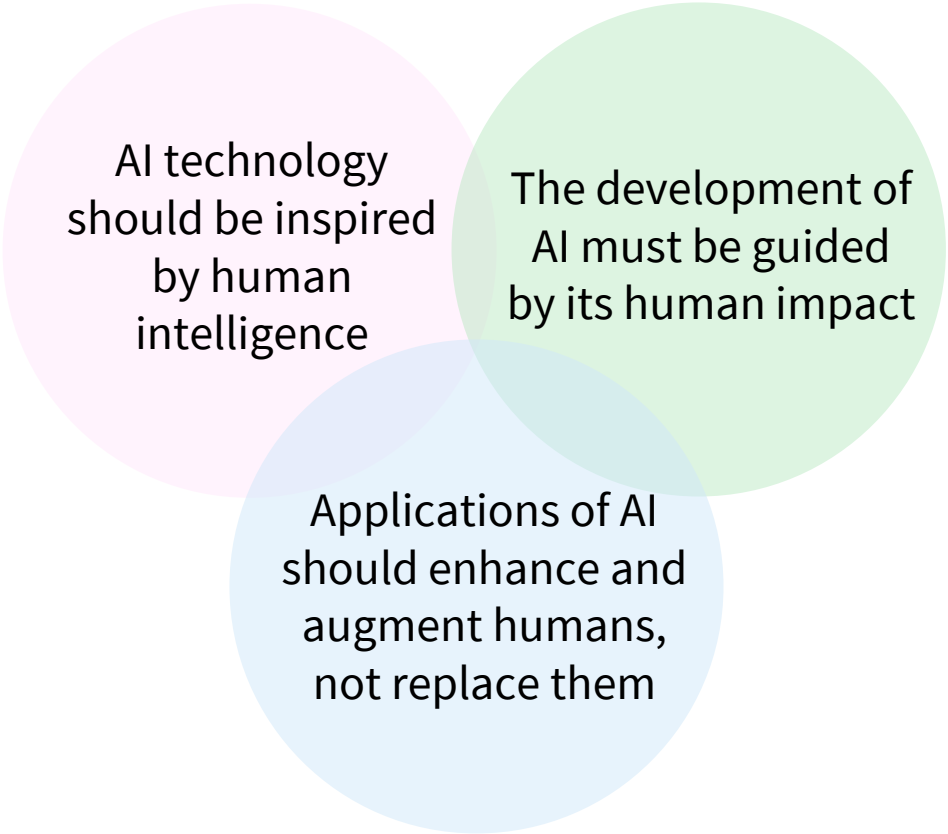
And society has noticed!



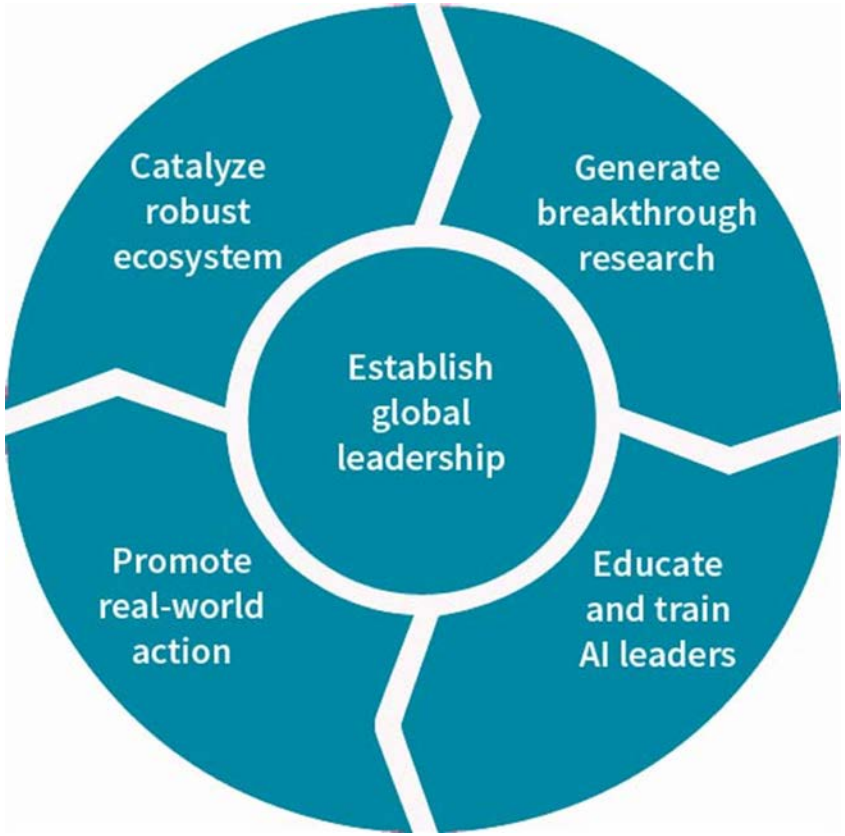


What can we do?

Human-Centered AI



Stanford Institute for Human-Centered AI



NEW!

AI and Machine Learning

Incredible achievements and even greater promise

Huge risk for mission-critical applications

AI Safety

We need AI systems to be

- Verifiable
- Reliable
- Robust against adversarial attacks
- Auditable
- Explainable
- Unbiased

Fortunately some of the world's best AI researchers are now working on exactly this!

The Future of Safe AI

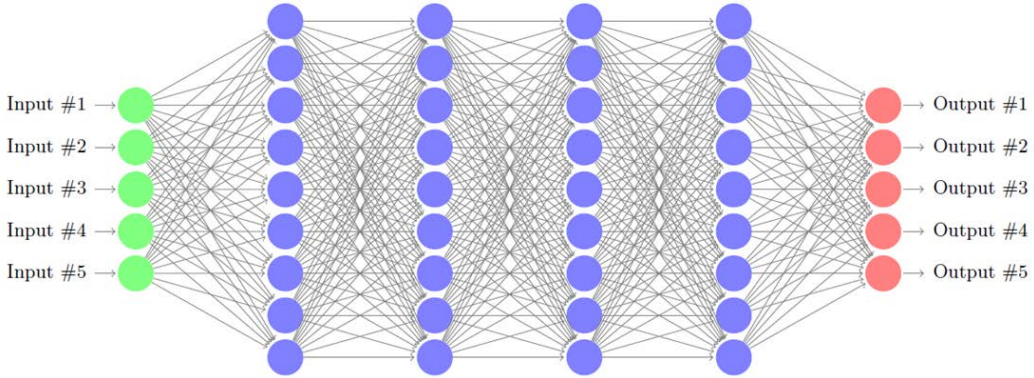
Three Examples of Cutting-Edge Research

1. Looking inside the black box of deep neural networks
2. Finding and removing bias
3. Assuring safe autonomous systems

Deep Neural Networks

Neural networks (NNs)

- Driver of AI revolution
- All you need is labeled training data—no programming required!
- But NNs are black boxes
- There are too many paths to test by running simulations, so how can we ensure NNs will work for every possible set of circumstances?



Verification of Deep Neural Networks

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer (Stanford University)

Verification of practical networks is experimentally beyond the reach of existing tools which can only handle small networks, for example a single hidden layer with 10 – 20 hidden nodes

Reluplex—a new algorithm for error-checking NNs

- Blends linear programming techniques with SMT solving techniques
- Encode NNs as linear arithmetic constraints
- Key insight: avoid testing paths that mathematically can never occur
- Scales to an order of magnitude larger networks than previously, for example a fully connected neural network with 8 layers and 300 nodes each

Many possible uses

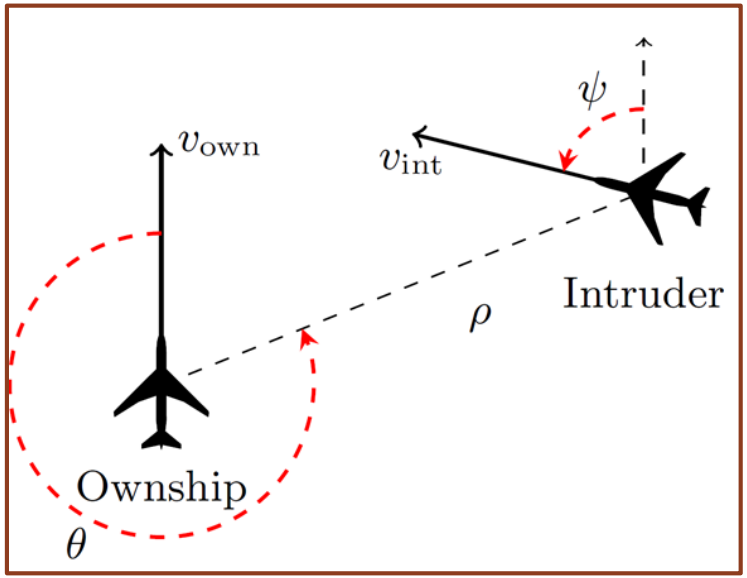
- Discover invariants of networks
- Prove properties of networks
- Measure formal adversarial robustness

Reluplex Case Study: ACAS Xu

Airborne collision-avoidance system for drones being developed by FAA

ACAS Xu examples

- If intruder approaches from left, then network advises strong right
  - › Proof in 1.5 hours
- If vertical separation is large and previous advisory is weak left, then network advises COC or weak left
  - › Counter-example found in 11 hours



Understanding Model Predictions

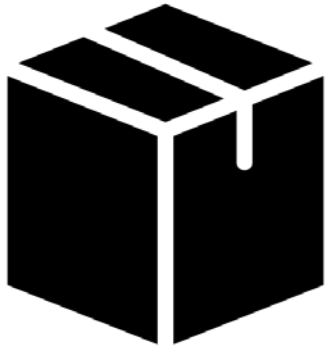
Pang Wei Koh and Percy Liang (Stanford University)

Given a high-accuracy black-box model and a prediction from it, why did the model make this prediction?

Important for loan applications, healthcare, and many other applications

Explainability enables

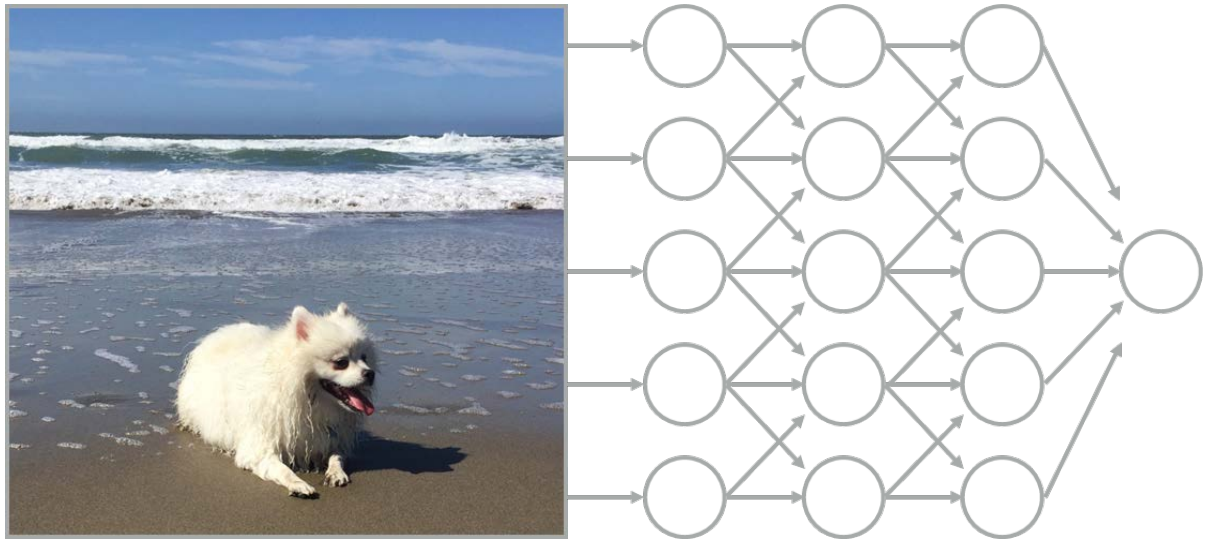
- Better decisions
- Improved models
- Discovery
- Trust and oversight



Influence Functions

Approach is to identify the training data points most responsible for a given prediction

Key insight: use “influence functions” calculated using gradients



Finding and Removing Human Biases in AI

James Zou and Londa Schiebeinger (Stanford University)

Bias can result from

- Training data—some groups may be over- or under-represented
  - Cure: investigate how training data is curated
- Algorithms—a typical machine learning program tries to maximize overall prediction accuracy for the training data
  - Cure: investigate how bias is propagated and amplified

Geometry Captures Semantics

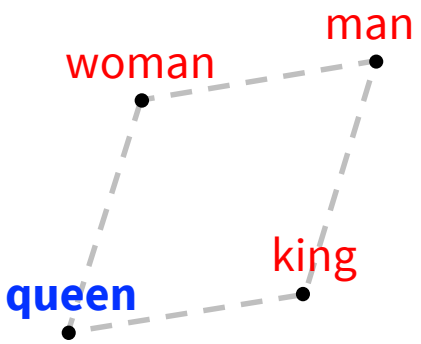
Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (Stanford University)

man	:	king	::	woman	:	

Geometry Captures Semantics

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (Stanford University)

man	:	king	::	woman	:	**queen**

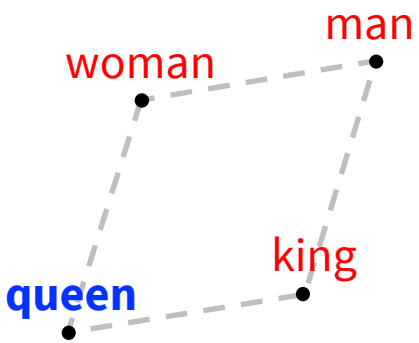


Based on word2vec trained on Google News corpus

Geometry Captures Semantics

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (Stanford University)

man	:	king	::	woman	:	**queen**
he	:	brother	::	she	:	

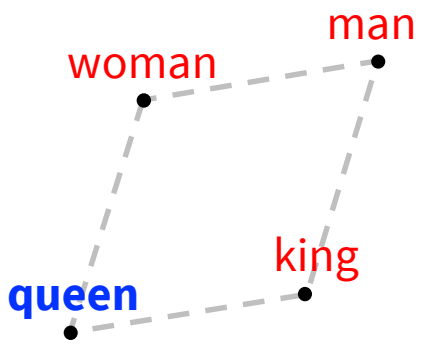


Based on word2vec trained on Google News corpus

Geometry Captures Semantics

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (Stanford University)

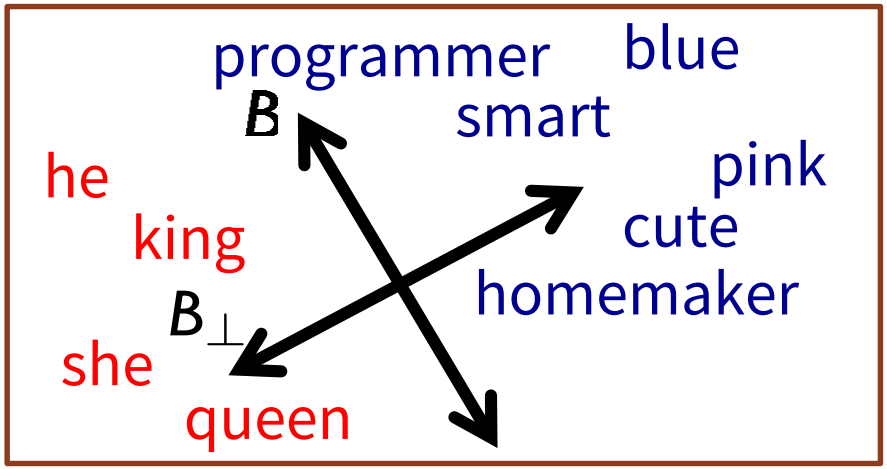
man	:	king	::	woman	:	**queen**
he	:	brother	::	she	:	sister
he	:	blue	::	she	:	pink
he	:	doctor	::	she	:	nurse
he	:	architect	::	she	:	interior designer
he	:	realist	::	she	:	feminist
she	:	pregnancy	::	he	:	kidney stone
he	:	computer programmer	::	she	:	homemaker



Based on word2vec trained on Google News corpus

Projecting Away Gender Component

- Reduce gender bias by removing gender stereotypes (receptionist ↔ female) and preserve desired associations (queen ↔ female)
- Distinguish gender-specific words and gender-neutral words



B = gender subspace

This debiasing is used by

Google

Microsoft

Achieving Fairness Without Demographics

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang (Stanford University)

The problem: small groups have low representation in minimizing average training loss

The current approach, empirical risk minimization, can make the problem worse by shrinking the minority group in the input data over time

Group labels are often unavailable

- Group annotation may be missing (due to cost or privacy)
- Protected group may not be identified or known

Achieving Fairness Without Demographics, *continued*

- Goal is to protect all groups—even minority groups—even without demographic labels
- **Solution:** an approach based on distributionally robust optimization which minimizes loss over all groups
- Distributionally robust optimization seeks to control the worst-case risks over all groups; intuitively, the approach is to upweight examples with high loss

What is fair?

Michael P. Kim, Omer Reingold, and Guy N. Rothblum (Stanford University and Weizmann Institute)

- Develop mathematically rigorous definitions for concepts like fairness and equity
- For example, algorithms that assure that similar people are treated similarly

Assuring Safe Autonomous Systems

Robots, drones, and autonomous vehicles need algorithms for safe learning, planning, and control

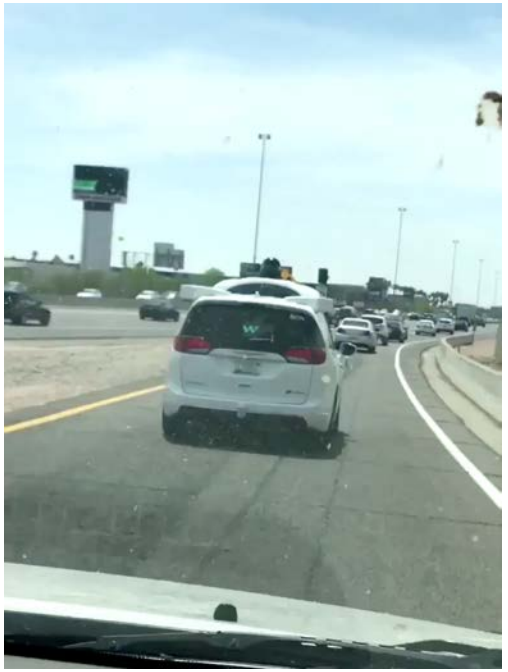
- Exploration of the environment
- Must deal with
  - Uncertainty and imperfect data
  - A dynamic (continuously changing) environment
  - Unpredictable human interactions
- Model the autonomous robot and the human as a system

Data-Driven Probabilistic Modeling for HRI

Marco Pavone (Stanford University)

Can we learn action distributions directly from experience without reasoning about motivations?

- Are intelligent counterparties cooperative, adversarial, or indifferent?
- Develop a decision-making and control stack for human-robot interactions where there are multiple distinct courses of action
  - First learn multimodal probability distributions
  - Then perform real-time policy construction
- Include high-level stochastic decision making and low-level safety preserving control



Safely Learning a Human's Internal State

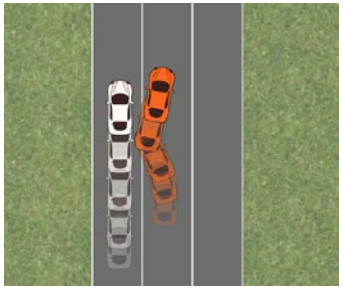
Dorsa Sadigh and Mykel Kochenderfer (Stanford University)

- Teach autonomous systems to learn the internal state of human drivers
- The robot maximizes its own reward function, but this reward function depends on what the human does in response (active information gathering)

$$\max_u \ \mathbb{E}_\theta \left[R_{goal}(x,u) + H\left(b_t(\theta)\right) - H\left(b_{t+1}(\theta)\right) \right]$$

exploitation

Info gain



The Future

Safe and reliable AI enabled by

- Verifiable AI including AI audits
- Explainable AI
- Fair AI including the ability to detect and remove bias
- Robust against errors, poor data, and malicious hacks

Research is providing innovative solutions

Success requires solutions, technical care, and social awareness