Current Events and Future Directions in AI

Tom Dietterich
High-Stakes Autonomy

Near-Term Risks

AI Progress

AI Progress

?
What is AI?

- Smart Software
  - vision, speech, touch
  - choosing actions to achieve goals
  - learning
  - understanding and predicting behavior

Credit: Andrej Karpathy, Li Fei-Fei
Exciting Progress: Perception

Google Speech Recognition

23% Word Error

Credit: Fernando Pereira & Matthew Firestone, Google

"a black and white cat is sitting on a chair."

Credit: Jeff Donahue, Trevor Darrell

Translate from Images

Credit: www.bbc.com
Exciting Progress: Reasoning (SAT)

Credit: Vijay Ganesh
Exciting Progress: Reasoning (Poker)

Credit: Michael Bowling CS331
Libratus and DeepStack

- Defeat human experts in the full game of Heads-Up No Limit Hold’em
- \(10^{160}\) information states
Exciting Progress: Chess and Go

Silver, et al. (2016) - Deep Learning + Monte Carlo Tree Search

Defeats Lee Sedol

Credit: Martin Mueller
Exciting Progress: Collaborative Systems

Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks

Red: Best AI solution
Yellow: “Foldit Void Crushers Group” solution
Blue: X-Ray Crystal Structure solution
Two Paths Forward

**Tool AI**
- Google, Bing
- IBM’s Watson
- Siri, Cortana, Google Now
- Google Translate
- Skype Real-Time Translation

**Autonomous AI**
- AI Hedge Funds
- High Speed Trading
- Self-Driving Cars
- Automated Surgical Assistants
- Smart Power Grid
- Autonomous Weapons
Tool AI: What’s Next?

Deeper understanding of video:
- What type of play?
- Who carried the ball?
- Was the pass complete?
- Which players made mistakes?
- Which players achieved their goals?
Tool AI: What’s Next?

Deeper understanding of text:

• Is Yoadimnadji dead? Yes
• Is he in Paris? Yes
• Is he married? Yes
• Where is his wife? In Paris
• Where will she be in the future? In Chad

“Pascal Yoadimnadji has been evacuated to France on Wednesday after falling ill and slipping into a coma in Chad. His wife, who accompanied him to Paris, will repatriate his body to Chad.”
Tool AI: What’s Next?

Linking Big Data to Medicine:

- Web search logs can detect adverse drug interaction events better than FDA’s existing Adverse Drug Interaction Reporting Service

White, Harpaz, Shah, DuMouchel, Horvitz, 2014
Tool AI: What’s Next?

Improved Personal Assistants that combine

• Knowledge of recipes (lasagna ingredients)
• Wine pairing recommendations (cabernet or pinot noir)
• Brother’s home address
• Routes to Brother’s home address
• Stores along that route that have cheap cabernet or pinot in stock to produce a plan

Source: Wired August 12, 2014

Viv
Autonomous AI Systems: What’s Next?

• AI Hedge Funds
• High Speed Trading
• Self-Driving Cars
• Automated Surgical Assistants
• Smart Power Grid
• Autonomous Weapons
High-Stakes Autonomy

The Motivations
- Advances in AI enable exciting applications
- There is a great potential to save lives

The Dangers
- Bugs
- Cyber Attacks
- Mixed Autonomy
- Misunderstanding User Commands
Software Quality

- Many AI methods give only probabilistic guarantees

- Research Question: How can we ensure safe performance of AI-based autonomous systems?

"a young boy is holding a baseball bat."

Credit: Andrej Karpathy, Li Fei-Fei
Cyber Attacks on AI Systems

- **Training Set Poisoning:**
  - Make yourself invisible to a computer vision system
  - Make yourself look “normal” to an anomaly detection system
  - Bias the behavior of the system
    - Bid slightly higher on certain stocks
    - Prefer to show certain advertisements

Credit: Katherine Hannah
Mixed Autonomy

- Auto-pilot unexpected hand-off to pilots
- Pilots lack situational awareness and make poor decisions

Question: How can we make imperfect autonomous systems safe?

AF447 Pilots Final Conversation
Pilot 1: “What the...how is it we are going down like this?”
Pilot 2: “See what you can do with the commands up there, the primaries and so on...Climb climb, climb, climb

Credit: www.aviationlawmonitor.com
Incomplete/incorrect Commands

- “Fetch me some water”
- Question: How can the computer reliably infer what the user intended?

Credit: www.disney.com
Trustable AI for Autonomy

Many research teams are at work

- Verification and Self-Monitoring
- Knowledge of Desirable and Acceptable Behavior
- Improved User Interaction
- Robustness
Approaches to Robust AI

- Robustness to Model Errors
  - Robust optimization
  - Regularize the model
  - Optimize a risk-sensitive objective
  - Employ robust inference algorithms

- Robustness to Unmodeled Phenomena
  - Expand the model
  - Learn a causal model
  - Employ a portfolio of models
  - Monitor performance to detect anomalies
Idea 1: Robust Optimization

- Many AI reasoning problems can be formulated as optimization problems

\[
\begin{align*}
\text{max} & \quad J(x_1, x_2) \\
\text{subject to} & \quad ax_1 + bx_2 \leq r \\
& \quad cx_1 + dx_2 \leq s
\end{align*}
\]
Uncertainty in the constraints

- \( \max_{x_1, x_2} J(x_1, x_2) \)
- subject to
  - \( ax_1 + bx_2 \leq r \)
  - \( cx_1 + dx_2 \leq s \)
- Define uncertainty regions
  - \( a \in U_a \)
  - \( b \in U_b \)
  - \( \ldots \)
  - \( s \in U_s \)
Minimax against uncertainty

\[
\max_{x_1, x_2} \min_{a, b, c, d, r, s} J(x_1, x_2; a, b, c, d, r, s)
\]

subject to

- \( ax_1 + bx_2 \leq r \)
- \( cx_1 + dx_2 \leq s \)
- \( a \in U_a \)
- \( b \in U_b \)
- \( ... \)
- \( s \in U_s \)

Problem: Tends to be too conservative
Impose a Budget on the Adversary

-\[
    \max_{x_1, x_2} \min_{\delta_a, \ldots, \delta_s} J(x_1, x_2; \delta_a, \ldots, \delta_s)
\]

- subject to
  - \((a + \delta_a)x_1 + (b + \delta_b)x_2 \leq (r + \delta_r)\)
  - \((c + \delta_c)x_1 + (d + \delta_d)x_2 \leq (s + \delta_s)\)
  - \(|\sum \delta_i| \leq B\)
  - \(\delta_a \in U_a\)
  - \(\delta_b \in U_b\)
  - \(\ldots\)
  - \(\delta_s \in U_s\)

(Bertsimas, & Thiele, 2006)
Idea 2: Regularize the Model

Regularization in Machine Learning:

- **Given:**
  - training examples \((x_i, y_i)\) for an unknown function \(y = f(x)\)
  - a loss function \(L(\hat{y}, y)\): how serious it is to output \(\hat{y}\) when the right answer is \(y\)?

- **Find:**
  - the model \(h\) that minimizes

\[
\sum_i L(h(x_i), y_i) + \lambda \|h\| \quad \text{loss + complexity penalty}
\]
Regularization can be Equivalent to Robust Optimization

- Xu, Caramanis & Mannor (2009)
  - Suppose an adversary can move each training data point $x_i$ by an amount $\delta_i$
  - Optimizing the linear support vector objective
    \[ \sum_i L(\hat{y}_i, y_i) + \lambda \|w\| \]
  - is equivalent to minimaxing against this adversary who has a total budget
    \[ \sum_i \|\delta_i\| = \lambda \]
Idea 3: Optimize a Risk-Sensitive Objective

- **Setting: Markov Decision Process**
  - States: $x_t, x_{t+1}, x_{t+2}$
  - Actions: $u_t, u_{t+1}$
  - Control policy $u_t = \pi(x_t)$
  - Rewards: $r_t, r_{t+1}$
  - Total reward $\sum_t r_t$
Idea 3: Optimize Conditional Value at Risk

- For any fixed policy $\pi$, the cumulative return $V^\pi = \sum_{t=1}^{T} r_t$ will have some distribution $P(V^\pi)$
- The Conditional Value at Risk at quantile $\alpha$ is the expected return of the bottom $\alpha$ quantile
- By changing $\pi$ we can change the distribution $P(V^\pi)$, so we can try to push the probability to the right
- “Minimize downside risks”

$CVaR = 3.94$
Optimizing CVaR gives robustness

- Suppose that for each time $t$, an adversary can choose a vector $\delta_t$ and define a new probability distribution

  $$P(x_{t+1}|x_t, u_t) \cdot \delta_t(u_t)$$

- Optimizing CVaR at quantile $\alpha$ is equivalent to minimaxing against this adversary with a budget along each trajectory of

  $$\prod_t \delta_t \leq \alpha$$

- Chow, Tamar, Mannor & Pavone (NIPS 2014)

- Conclusion: Acting Conservatively Gives Robustness to Model Errors
Idea 4: Robust Inference

- Credal Bayesian Networks
  - Convex uncertainty sets over the probability distributions at nodes
  - Upper and lower probability models
  - (Cosman, 2000; UAI 1997)

- Robust Classification
  - (Antonucci & Zaffalon, 2007)

- Robust Probabilistic Diagnosis (etc.)
  - (Chen, Choi, Darwiche, 2014, 2015)
Approaches to Robust AI

- **Robustness to Model Errors**
  - Robust optimization
  - Regularize the model
  - Optimize a risk-sensitive objective
  - Employ robust inference algorithms

- **Robustness to Unmodeled Phenomena**
  - Expand the model
  - Learn a causal model
  - Employ a portfolio of models
  - Monitor performance to detect anomalies
Idea 5: Detect Model Failure and Expand the Model

- Posterior Predictive Checks
  - Monitor how well the model predicts auxiliary phenomena

- Expand the Model to Repair Failures
  - Information Extraction & Knowledge Base Population
    - NELL “Read the Web” (Mitchell, et al., AAAI 2015)
    - TAC-KBP (NIST)
  - Learning Models of Actions in Planning and Reinforcement Learning
    - Gil (1994)
Idea 5: Expand the Model

- **Risk:**
  - Every new item added to a model may introduce an error
  - Inference may propagate these errors
  - The expanded model may not be more accurate than the original model

- Does not address the fundamental need to act robustly in incompletely-modeled environments
Idea 6: Use Causal Models

- Causal relations are more likely to be robust
  - Can be “transported” to novel situations
    - (Pearl & Bareinboim, AAAI 2011)
    - (Lee & Honavar, AAAI 2013)
- Transport from

Learning Situation  ➔  Runtime Situation
Idea 7: Employ a Portfolio of Models

“We usually know several different ways to do something, so that if one of them fails, there's always another.”

--Marvin Minsky
Portfolio Methods in SAT & CSP

- SATzilla:
  - Xu, Hoos, Hutter, Leyton-Brown (JAIR 2008)
SATzilla Results

- HANDMADE problem set
- Presolvers:
  - March_d104 (5 seconds)
  - SAPS (2 seconds)

Cumulative Distribution

Xu, Hutter, Hoos, Leyton-Brown (JAIR 2008)
IBM Watson / DeepQA

- Combines >100 different techniques for
  - analyzing natural language
  - identifying sources
  - finding and generating hypotheses
  - finding and scoring evidence
  - merging and ranking hypotheses
Knowledge-Level Redundancy

- Minsky: “You don’t really understand something if you only understand it one way”

- Most AI systems only understand things one way:
  - Computer vision:
    - Object Appearance $\rightarrow$ human labels
  - Natural Language:
    - Word Co-occurrence statistics $\rightarrow$ human labels

"a black and white cat is sitting on a chair."
Multifaceted Understanding

- There is a person who is the cat’s owner
- That person does not like the cat sitting on the chair
  - The cat is preventing a person from sitting on the chair
    - People often need to sit on chairs
  - The cat leaves hair on the chair
  - The cat is preventing the person from picking up the book
- The cat will soon not be on the chair

"a black and white cat is sitting on a chair."
Achieving Multi-Faceted Understanding

- We need to give our computers many different forms of experience
  - Performing tasks
  - Achieving goals through natural language dialogue
  - Interacting with other agents
- Examples:
  - Lake, Salakhutdinov & Tenenbaum (Science 2016)
Idea 8: Watch for Anomalies

- Machine Learning Theory
  - Training examples drawn from $P_{train}(x)$
  - Classifier $y = f(x)$ is learned
  - Test examples from $P_{test}(x)$
  - If $P_{test} = P_{train}$ then with high probability $f(x)$ will be correct for test queries

- What if $P_{test} \neq P_{train}$?
Automated Counting of Freshwater Macroinvertebrates

- **Goal:** Assess the health of freshwater streams
- **Method:**
  - Collect specimens via kicknet
  - Photograph in the lab
  - Classify to genus and species

[Image: www.epa.gov]
Open Category Object Recognition

- Train on 29 classes of insects
- Test set may contain additional species
Prediction with Anomaly Detection

\[ x \]

Anomaly Detector

\[ A(x) < \tau? \]

Classifier \( f \)

\[ y = f(x) \]

Source: Dietterich & Fern, unpublished
Novel Class Detection via Anomaly Detection

- Train a classifier on data from 2 classes
- Test on data from 26 classes
- Black dot: Best previous method
Anomaly Detection Notes

- We initially just used monochrome images
  - Feature selection studies showed this was sufficient
- But color is very useful for detecting novel classes
- Lesson: Use all of your features when looking for anomalies
Related Efforts

- **Open Category Classification**
  - (Salakhutdinov, Tenenbaum, & Torralba, 2012)
  - (Da, Yu & Zhou, AAAI 2014)
  - (Bendale & Boult, CVPR 2015)

- **Change-Point Detection**
  - (Page, 1955)
  - (Barry & Hartigan, 1993)
  - (Adams & MacKay, 2007)

- **Covariate Shift Correction**
  - (Sugiyama, Krauledat & Müller, 2007)
  - (Quinonero-Candela, Sugiyama, Schwaighofer & Lawrence, 2009)

- **Domain Adaptation**
  - (Blitzer, Dredze, Pereira, 2007)
  - (Daume & Marcu, 2006)
High-Stakes Autonomy

Near-Term Risks

AI Progress

AI Progress
Some people are very afraid

Stephen Hawking warns artificial intelligence could end mankind

Elon Musk: artificial intelligence is our biggest existential threat

Fears of an AI pioneer
Stuart Russell argues that AI is as dangerous as nuclear weapons
AI Misconceptions

Intelligence is not a threshold phenomenon

- Progress in AI is the accumulation of thousands of incremental improvements
- Robots will not “wake up” one day and be “truly intelligent” or “superintelligent” or “conscious” or “sentient”
- “Tool AI” systems are already smarter than people along many dimensions
AI Misconceptions

Autonomy Will Not Happen Spontaneously

- There is no threshold above which AI systems suddenly have free will
- Systems need to be designed and built to be autonomous
- They must be given access to resources (money, power, materials, generalized task markets, communications with people)
The danger of “Autonomous AI” is not “AI” but “Autonomy”

- An autonomous system can be dangerous for many reasons:
  - It could consume vast resources
  - It could injure or kill people
  - It could apply AI to help it do these things
We Should Never Create a Fully Autonomous System

By definition, a fully autonomous system is a system over which we have no control.
Summary

- AI has been making steady progress
- Exciting applications of Tool AI are coming soon
- Potential applications of Autonomous AI are being proposed
  - Many of these involve high-stakes decision making
  - There are many risks that must be addressed before it will be safe to field such systems
- AI Research is studying methods for making AI systems robust to known unknowns and unknown unknowns
- Tool AI will not spontaneously become autonomous
- We should not build fully autonomous systems!