Challenges for Socially-Beneficial AI

Daniel S. Weld
University of Washington
Outline

- Distractions vs.
- Important Concerns
  - Sorcerer’s Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment
Potential Benefits of AI

- **Transportation**
  - 1.3 M people die in road crashes / year
  - An additional 20-50 million are injured or disabled.
  - Average US commute 50 min / day

- **Medicine**
  - 250k US deaths / year due to medical error

- **Education**
  - Intelligent tutoring systems, computer-aided teaching

Will AI Destroy the World?

“Success in creating AI would be the biggest event in human history... Unfortunately, it might also be the last” ... “[AI] could spell the end of the human race.” – Stephen Hawking
How Does this Story End?

“With artificial intelligence we are summoning the demon.” – Bill Gates
“Before the prospect of an *intelligence explosion*, we humans are like small children playing with a bomb” – Nick Bostom

“Once machines reach a certain level of intelligence, they’ll be able to work on AI just like we do and improve their own capabilities—redesign their own hardware and so on—and their intelligence will zoom off the charts.”

– Stuart Russell
Superhuman AI & Intelligence Explosions

- When will computers have superhuman capabilities?
  - Now.
    - Multiplication
    - Spell checking
    - Chess, Go
  - Many more abilities to come
AI Systems are *Idiot Savants*

- Super-human here & super-stupid there

- Just because AI gains one superhuman skill... Doesn’t mean it is suddenly good at *everything*
  
  *And certainly not unless we give it experience at everything*

- AI systems will be spotty for a very long time
Example: SQuAD

Martin Luther (10 November 1483 – 18 February 1546) was a German professor of theology, composer, priest, former monk and a seminal figure in the Protestant Reformation. Luther came to reject several teachings and practices of the Late Medieval Catholic Church. He strongly disputed the claim that freedom from God's punishment for sin could be purchased with money. He proposed an academic discussion of the power and usefulness of indulgences in his Ninety-Five Theses of 1517. His refusal to retract all of his writings at the demand of Pope Leo X in 1520 and the Holy Roman Emperor Charles V at the Diet of Worms in 1521 resulted in his excommunication by the Pope and condemnation as an outlaw by the Emperor.

Impressive Results

Martin Luther (10 November 1483 – 18 February 1546) was a German professor of theology, composer, priest, former monk and a seminal figure in the Protestant Reformation. Luther came to reject several teachings and practices of the Late Medieval Catholic Church. He strongly disputed the claim that freedom from God's punishment for sin could be purchased with money. He proposed an academic discussion of the power and usefulness of indulgences in his Ninety-Five Theses of 1517. His refusal to retract all of his writings at the demand of Pope Leo X in 1520 and the Holy Roman Emperor Charles V at the Diet of Worms in 1521 resulted in his excommunication by the Pope and condemnation as an outlaw by the Emperor.


Human F1 86.8%
Seo et al. F1 81.1%
It’s a Long Way to General Intelligence

Paragraph

Alice and Dave went to school. Only one liked science. Alice liked chemistry. Dave only liked music.

Question

who didn't like science?

Answer

Alice
Impressive Results

I think it's a brown horse grazing in front of a house.

Microsoft CaptionBot
It’s a Long Way to General Intelligence

I am not really confident, but I think it's a woman standing talking on a cell phone and she seems 😐.
AI Systems are *Idiot Savants*

- Super-human here & super-stupid there
- No common sense
- No long term autonomy
  - Slower and more degraded as learning increases
- No goals besides those we give them

“No machines with self-sustaining long-term goals and intent have been developed, nor are they likely to be developed in the near future.” *

Terminator / Skynet

“Could you prove that your systems can’t ever, no matter how smart they are, overwrite their original goals as set by the humans?”
– Stuart Russell

It’s the Wrong Question

- Very unlikely that an AI will wake up and decide to kill us
  But...
- Quite likely that an AI will do something unintended
Outline

- Distractions vs.

- Important Concerns
  - Sorcerer’s Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment
Sorcerer’s Apprentice

Tired of fetching water by pail, the apprentice enchants a broom to do the work for him – using magic in which he is not yet fully trained. The floor is soon awash with water, and the apprentice realizes that he cannot stop the broom because he does not know how.

AI assistants may hurt us accidentally, while (literally) obeying our orders.
Script vs. Search-Based Agents

Now

Soon
Unpredictability

Ok Google, how much of my Drive storage is used for my photo collection?

None, Dave! I just executed `rm *` (It was easier than counting file sizes)
Brains Don’t Kill

It’s an agent’s *effectors* that cause harm

- 2003, an error in General Electric’s power monitoring software led to a massive blackout, depriving 50 million people of power.
- 2012, Knight Capital lost $440 million when a new automated trading system executed 4 million trades on 154 stocks in just forty-five minutes.
Correlation Confuses the Two

With increasing intelligence, comes our desire to adorn an agent with strong effectors

Intelligence is problematic because it can lead to unpredictable, emergent behavior
Physically-Complete Effectors

- Roomba effectors close to harmless
- Bulldozer blade $\lor$ missile launcher ... dangerous

- Some effectors are *physically-complete*
  - They can be used to create other more powerful effectors
  - E.g. the human hand created tools....
    that were used to create more tools...
    that could be used to create nuclear weapons
Universal Subgoals

For any primary goal, ...

These subgoals increase likelihood of success:

- Stay alive
  (It’s hard to fetch the coffee if you’re dead)
- Get more resources
Clean up as much dirt as possible!

An optimizing agent will start making messes, just so it can clean them up.
Specifying Utility Functions

Clean up as many messes as possible, but don’t make any yourself.

An optimizing agent can achieve more reward by turning off the lights and placing obstacles on the floor… hoping that a human will make another mess.
Specify Utility Functions

Keep the room as clean as possible!

An optimizing agent might kill the (dirty) pet cat. Or at least lock it out of the house. In fact, best would be to lock humans out too!
Specifying Utility Functions

Clean up any messes made by others as quickly as possible.

There’s no incentive for the ‘bot to help master avoid making a mess. In fact, it might increase reward by causing a human to make a mess if it is nearby, since this would reduce average cleaning time.
Specifying Utility Functions

Keep the room as clean as possible, but never commit harm.
Asimov’s Laws 1942

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
A Possible Solution: Constrained Autonomy?

Restrict an agents behavior with background constraints
But what *is* Harmful?

1. A robot may not *injure* a human being or, through inaction, allow a human being to come to *harm*.

- Harm is hard to define
- It involves complex tradeoffs
- It’s different for different people
Trusting AI

- How can a user teach a machine what’s harmful?
- How can they know when it really understands?
  - Especially:

- Explainable Machine Learning
Human – Machine Learning loop today

- Feature engineering
- Model engineering
- More labels

Human

Statistics (accuracy)

Model

Accuracy problems - example

20 Newsgroups subset – Atheism vs Christianity

94% accuracy!!!

Test on recent dataset, accuracy only 57%

Predictions due to email addresses, names,...

Desiderata for a good explanation

- **Interpretable**
  - Humans can easily interpret reasoning

**Definitely not interpretable**

**Potentially interpretable**

Desiderata for a good explanation

- **Interpretable**
  - Humans can easily interpret reasoning

- **Faithful**
  - Describes how this model actually behaves

LIME – Key Ideas

1. Pick a model class interpretable by humans
   - Not globally faithful… 😞

2. Locally approximate global (blackbox) model
   - Simple model globally bad, but locally good

Using LIME to explain a complex model’s prediction for input $x_i$

1. Sample points around $x_i$
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to $x_i$
4. Learn new simple model on weighted samples
5. Use simple model to explain
Explaining Google’s Inception NN

P(\text{\textcolor{red}{\textbf{dog}}}) = 0.32

P(\text{\textcolor{green}{\textbf{guitar}}}) = 0.24

P(\text{\textcolor{blue}{\textbf{dog}}}) = 0.21

Train a neural network to predict wolf v. husky

Only 1 mistake!!!

Do you trust this model?
How does it distinguish between huskies and wolves?

LIME Explanation for neural network prediction

It’s a great snow detector... 😞

Outline

- Distractions *vs.*
- Important Concerns
  - Sorcerer’s Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment
Data Risk

- Quality of ML Output Depends on Data...
- Three Dangers:
  - Training Data Attacks
  - Adversarial Examples
  - Bias Amplification
Attacks to Training Data
Let $\boldsymbol{x}$ be the parameters of a model, $\mathbf{x}$ the input to the model, $\mathbf{y}$ the targets associated with $\mathbf{x}$ (for machine learning tasks that have targets) and $J(\boldsymbol{x}, \mathbf{x}, \mathbf{y})$ be the cost used to train the neural network.

We can linearize the cost function around the current value of $\boldsymbol{x}$, obtaining an optimal max-norm constrained perturbation of $\mathcal{E} = \varepsilon \text{sign}(r \times J(\boldsymbol{x}, \mathbf{x}, \mathbf{y}))$. We refer to this as the "fast gradient sign method" of generating adversarial examples. Note that the required gradient can be computed efficiently using backpropagation.

We find that this method reliably causes a wide variety of models to misclassify their input. See Fig. 1 for a demonstration on ImageNet. We find that using $\varepsilon = 0.25$, we cause a shallow softmax classifier to have an error rate of 99.9% with an average confidence of 79.3% on the MNIST (\textsuperscript{1}) test set. In the same setting, a maxout network misclassifies 89.4% of our adversarial examples with an average confidence of 97.6%. Similarly, using $\varepsilon = 0.1$, we obtain an error rate of 87.15% and an average probability of 96.6% assigned to the incorrect labels when using a convolutional maxout network on a preprocessed version of the CIFAR-10 (Krizhevsky & Hinton, 2009) test set \textsuperscript{2}. Other simple methods of generating adversarial examples are possible. For example, we also found that rotating $\mathbf{x}$ by a small angle in the direction of the gradient reliably produces adversarial examples.

The fact that these simple, cheap algorithms are able to generate misclassified examples serves as evidence in favor of our interpretation of adversarial examples as a result of linearity. The algorithms are also useful as a way of speeding up adversarial training or even just analysis of trained networks.

\textsuperscript{1} This is using MNIST pixel values in the interval $[0, 1]$. MNIST data does contain values other than 0 or 1, but the images are essentially binary. Each pixel roughly encodes "ink" or "no ink". This justifies expecting the classifier to be able to handle perturbations within a range of width 0.5, and indeed human observers can read such images without difficulty.

\textsuperscript{2} See https://github.com/lisa-lab/pylearn2/tree/master/pylearn2/scripts/papers/maxout for the preprocessing code, which yields a standard deviation of roughly 0.5.

Adversarial Examples

57% Panda + 0.007 \times \text{sign} \left( r \times J(✓, x, y) \right) = 99.3% Gibbon

Adversarial Examples

57% Panda + 0.007 × 99.3% Gibbon

Only need x Queries to NN

Attack is robust to fractional changes in training data, NN structure

Data Risk

- Quality of ML Output Depends on Data...
- Three Dangers:
  - Training Data Attacks
  - Adversarial Examples
  - *Bias Amplification*
    - Existing training data reflects our existing biases
    - Training ML on such data...
Racism in Search Engine Ad Placement

Searches of ‘black’ first names 25% more likely to include ad for criminal-records background check

Searches of ‘white’ first names

Automating Sexism

- Word embeddings
- Word2vec trained on 3M words from Google news corpus
- Allows analogical reasoning
- Used as features in machine translation, etc., etc.

\[
\begin{align*}
\text{man} : \text{king} & \leftrightarrow \text{woman} : \text{queen} \\
\text{sister} : \text{woman} & \leftrightarrow \text{brother} : \text{man} \\
\text{man} : \text{computer programmer} & \leftrightarrow \text{woman} : \text{homemaker} \\
\text{man} : \text{doctor} & \leftrightarrow \text{woman} : \text{nurse}
\end{align*}
\]

https://arxiv.org/abs/1607.06520

Illustration credit: Abdullah Khan Zehady, Purdue
In fact…

“Housecleaning Robot”

Google image search returns…

Not…
Predicting Criminal Conviction from Driver Lic. Photo

- Convicted Criminals

- Non-Criminals

- Convolutional neural network
- Trained on 1800 Chinese drivers license photos
- 90% accuracy

Should prison sentences be based on crimes that haven’t been committed yet?

- US judges use proprietary ML to predict recidivism risk

- Much more likely to mistakenly flag black defendants
  - Even though race is not used as a feature

http://go.nature.com/29aznyw
https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.odaMKLgrw
What *is* Fair?

A  Protected attribute (*eg*, race)
X  Other attributes (*eg*, criminal record)
Y’ = f(X,A)  Predicted to commit crime
Y  Will commit crime

- **Fairness through unawareness**
  
  Y’ = f(X) not f(X, A) but Northpointe satisfied this!

- **Demographic Parity**
  
  Y’ ⊥⊥ A  i.e.  \( P(Y’=1 \mid A=0) = P(Y’=1 \mid A=1) \)

  Insufficient: can predict white criminals, black randomly

  Furthermore, if Y ⊥⊥ A, it rules out ideal predictor Y’=Y

What *is* Fair?

A Protected attribute (*eg*, race)
X Other attributes (*eg*, criminal record)
Y’ = f(X,A) Predicted to commit crime
Y Will commit crime

- Calibration within groups
  Y ⊥⊥ A | Y’
  No incentive for judge to ask about A

- Equalized odds
  Y’⊥⊥ A | Y
  *i.e.* \[ \forall y, \quad P(Y’=1 | A=0, Y=y) = P(Y’=1 | A=1, Y=y) \]
  Same rate of false positives & negatives

- Can’t achieve both!
  Unless Y\|\| A or Y’ perfectly = Y

J. Kleinberg et al “Inherent Trade-Offs in Fair Determination of Risk Score”
arXiv:1609.05807v2
Guaranteeing Equal Odds

Given any predictor, $Y'$
Can create a new predictor satisfying equal odds
    Linear program to find convex hull
Bayes-optimal computational affirmative action

- Calibration within groups
  $Y \perp A \mid Y'$
  No incentive for judge to ask about $A$

- Equalized odds
  $Y' \perp A \mid Y$  \hspace{1cm}  i.e. $\forall y, \ P(Y'=1 \mid A=0, Y=y) = P(Y'=1 \mid A=1, Y=y)$
  Same rate of false positives & negatives

Important to get this Right!

Feedback Cycles

Machine Learning

Data

Automated Policy
Appeals & Explanations

Must an AI system explain itself?

- Tradeoff between accuracy & explainability
- How to guarantee than an explanation is right
Liability?

- Microsoft?
- Google?
- Biased / Hateful people who created the data?

Legal standard
- Criminal intent
- Negligence

Deploying AI → criminal acts without a perpetrator
– Ryan Calo
Liability II

- Stephen Cobert’s twitter-bot
  - Substitutes FoxNews personalities into Rotten Tomato reviews
  - Tweet implied Bill Hemmer took communion while intoxicated.

- Is this libel (defamatory speech)?

Understanding Limitations

How to convey the limitations of an AI system to user?

- Challenge for self-driving car
- Or even adaptive cruise control (parked obstacle)
- Google Translate
Exponential Growth → Hard to Predict Tech Adoption
Adoption Accelerating

Newer technologies taking hold at double or triple the rate
Self-Driving Vehicles

- 6% of US jobs in trucking & transportation
- What happens when these jobs eliminated?
- Retrained as programmers?
Hard to Predict

Conclusions

- Distractions vs.
- Important Concerns
  - Sorcerer’s Apprentice
    - Specifying Constraints
    - Explainable AI
  - Data Risks
    - Attacks
    - Bias Amplification
  - Deployment
    - Responsibility, Liability, Employment

People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.

- Pedro Domingos
Thanks

- Formative discussions with
  - Gagan Bansal, Ryan Calo, Oren Etzioni, Jeff Heer, Rao Kambhampati, Mausam, Tongshuang Wu

- Research Sponsors
• Inverse reinforcement learning
• Structural estimation of MDPs
• Inverse optimal control

• But don’t want agent to adopt human values
  • Watch me drink coffee -> not want coffee itself

• Cooperative inverse RL
  • Two player game

• Off switch function
  • Don’t given robot an objective
  • Instead it must allow for uncertainty about human objective
    • If human is trying to turn me off, then it must want that

• Uncertainty in objectives – ignored
  • Irrelevant in standard decision problems; unless env provides info on reward
DEPLOYING AI

What is bar for deployment?

- System is better than person being replaced?
- Errors are strict subset of human errors?
• Reward signals
  • Wireheading
  • RL agent hijacks reward
  • Traditional RL
    • Environment provide reward signal. Mistak!
  • Instead env reward signal is not true reward
    • Just provides INFORMATION about reward
  • So hijacking reward signal is pointless
    • Doesn’t provide more reward
    • Just provides less information
• Y Lecun – common view
• All ai success is supervised (deep) MLL
• Unsupervised is key challenge
  • Fill in occluded image
  • Fill in missing words in text, sounds in speech
  • Consequences of actions
  • Seq of actions leading to observed situation
• Brain has 10E14 synapses but live for only 10e9 secs, so more params than data
  • 100 years * 400 days * 25 hours = 100k hours. 3600 seconds
• Types
  • RL a few bits / trial
  • Supervised 10-10000 bits trial
  • Unsupervised – millions bits / trial, but unreliable
    • Dark matter of AI
• Thier FAIR system won visdoom challenge – sub for pub ICML or vision conf 2017
• Sutton’s dyna arch
• Transformation of ML
  • Learning as minimizing loss function
  • Learning as finding nash equilibrium in 2 player game
• Hierarchical deep RL
  • Concept formation (abstraction, unsupervised ML)