Challenges for Socially-Beneficial AI

Daniel S. Weld
University of Washington

Outline

- Dangers, Priorities & Perspective
- Sorcerer's Apprentice Scenario
  - Specifying Constraints & Utilities
  - Explainable AI
- Data Risks
  - Bias & Bias Amplification
- Deployment
  - Responsibility, Liability, Employment
- Attacks

5

Potential Benefits of AI

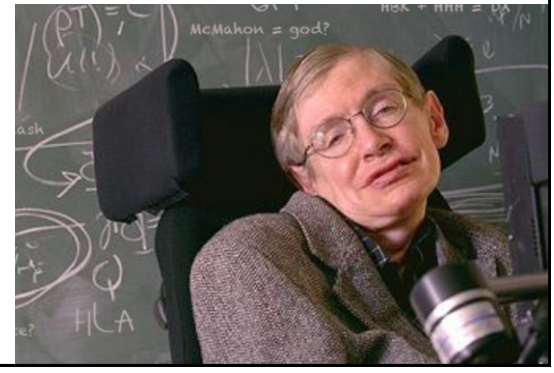
- Transportation
  - 1.3 M people die in road crashes / year
  - An additional 20-50 million are injured or disabled.
  - Average US commute 50 min / day
- Medicine
  - 250k US deaths / year due to medical error
- Education
  - Intelligent tutoring systems, computer-aided teaching

- asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics
- https://www.washingtonpost.com/news/to-your-health/wp/2016/05/03/researchers-medical-errors-now-third-leading-cause-of-death-in-united-states/?utm_term=.49f29cb6dae9

6

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



An Intelligence Explosion?

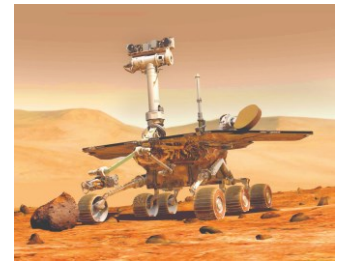
"Before the prospect of an ***intelligence explosion***, we humans are like small children playing with a bomb" – Nick Bostom

"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
  - Transportation & Mission Planning
  - Many more abilities to come

10

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

11

Example: SQuAD

Paragraph

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.

Question

Who asked Luther to disavow his writings?

Human F1 86.8%

12

Rajpurkat et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text," https://arxiv.org/pdf/1606.05250.pdf

Impressive Results

Paragraph

Martin Luther (10 November 1483 – 18 February 1546) wa[illegible] professor of theology, composer, priest, former monk and a s[illegible]e in the Protestant Reformation. Luther came to reject sever[illegible]eachings 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.

http://35.165.153.16:1995/

Question

Who asked Luther to disavow his writings?

Human F1 86.8%
Seo et al. F1 81.1%

Answer

Pope Leo X

13

Seo et al. "Bidirectional Attention Flow for Machine Comprehension" arXiv:1611.01603v5

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

14

4 Capabilities AGI Requires

- The object-recognition capabilities of a ***2-year-old child***.
- A 2-year-old can observe a variety of objects of some type—different kinds of shoes, say—and successfully categorize them as shoes, even if he or she has never seen soccer cleats or suede oxfords.
- Today's best computer vision systems still make mistakes—both false positives and false negatives—that no child makes.

15

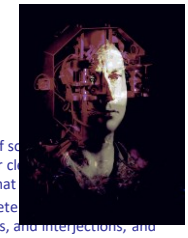
https://spectrum.ieee.org/computing/hardware/i-rodney-brooks-am-a-robot

4 Capabilities AGI Requires

- The language capabilities of a ***4-year-old child.***
- The manual dexterity of a ***6-year-old child.***
- The social understanding of an ***8-year-old child.***
- …who can understand the difference between what he or she knows about a situation and what another person could have observed and therefore could know… a "theory of the mind"
- E.g., suppose a child sees her mother placing a chocolate bar inside a drawer. The mother walks away, and the child's brother comes and takes the chocolate. The child knows that mother still thinks the chocolate is in the drawer.
- Despite decades of study, far beyond any existing AI system.

16

4 Capabilities AGI Requires



- The object-recognition capabilities of a 2-year-old child. A 2-year-old can observe a variety of objects of so
  kinds of shoes, say—and successfully categorize them as shoes, even if he or she has never seen soccer cl
  Today's best computer vision systems still make mistakes—both false positives and false negatives—that
- The language capabilities of a 4-year-old child. By age 4, children can engage in a dialogue using complete
  irregularities, idiomatic expressions, a vast array of accents, noisy environments, incomplete utterances, and interjections, and
  they can even correct nonnative speakers, inferring what was really meant in an ungrammatical utterance and reformatting it.
  Most of these capabilities are still hard or impossible for computers.
- The manual dexterity of a 6-year-old child. At 6 years old, children can grasp objects they have not seen before; manipulate flexible objects in tasks like tying shoelaces; pick up flat, thin objects like playing cards or pieces of paper from a tabletop; and manipulate unknown objects in their pockets or in a bag into which they can't see. Today's robots can at most do any one of these things for some very particular object.
- The social understanding of an 8-year-old child. By the age of 8, a child can understand the difference between what he or she knows about a situation and what another person could have observed and therefore could know. The child has what is called a "theory of the mind" of the other person. For example, suppose a child sees her mother placing a chocolate bar inside a drawer. The mother walks away, and the child's brother comes and takes the chocolate. The child knows that in her mother's mind the chocolate is still in the drawer. This ability requires a level of perception across many domains that no AI system has at the moment.

17

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

There are More Important Questions

- Very unlikely that an AI will wake up and decide to kill us

  But…
- Quite likely that an AI will do something unintended
- Quite likely that an evil person will use AI to hurt people

18

~~Artificial General Intelligence (AGI)~~

- Well before we have human-level AGI
- We will have lots of superhuman ASI
  - Artificial ***specific*** intelligence
  - Inspectability / trust / utility issues will hit here first

19

Outline

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

20

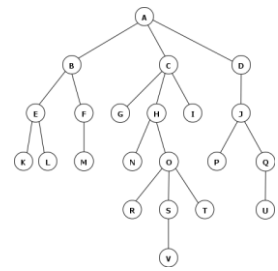
Sorcerer's Apprentice

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, a the apprentice realizes that he canno 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

22

Unpredictability

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

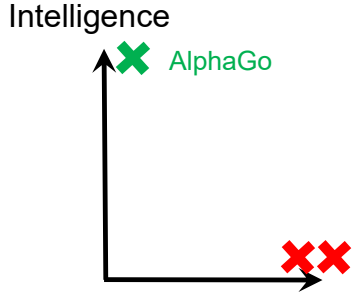
None, Dave!
I just executed rm *

(It was easier than counting file sizes)

23

Brains Don't Kill

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

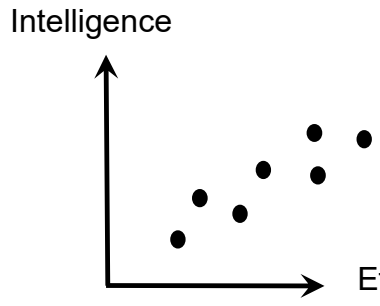


- 2003, an error in General Electric's power monitoring software led to a massive blackout, depriving 50 million people of power.
- 2012, Knight Capital lost $440 million when a new automated trading system executed 4 million trades on 154 stocks in just forty-five minutes.

24

Correlation Confuses the Two

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



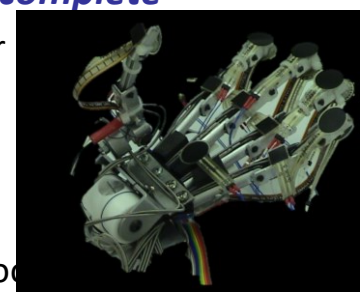
Intelligence is problematic because it can lead to unpredictable, emergent behavior

25

Physically-Complete Effectors

- Roomba effectors close to harmless
- Bulldozer blade v 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 to that could be used to create nuclear weapons



26

Universal Subgoals

-Stuart Russell

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

27

Specifying Utility Functions

Clean up as much dirt as possible!

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

28

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.

29

Specifying 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!

30

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.

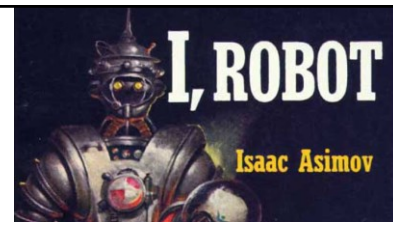
31

Specifying Utility Functions

Keep the room as clean as possible, but never commit harm.

32

Asimov's Laws



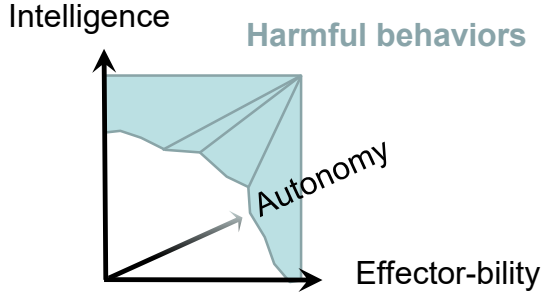
1942

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.

33

A Possible Solution: Constrained Autonomy?

Restrict an agents behavior with background constraints



34

But what *is* Harmful?

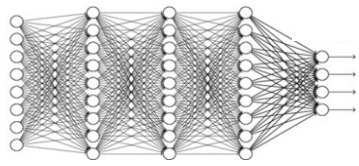
1. A robot may not ***injure*** a human being or, through inaction, allow a human being to come to ***harm***.

- Harm is hard to define
- It involves complex tradeoffs
- It's different for different people

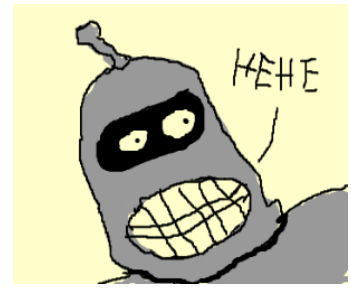
35

Trusting AI

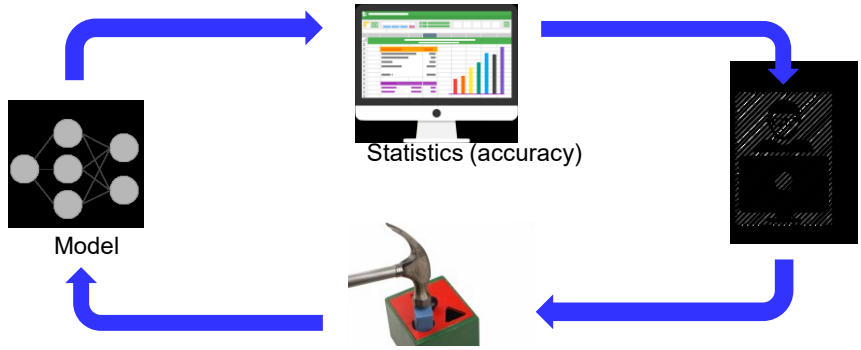
- How can a user teach a machine what's harmful?
- How can they know when it really understands?
  - Especially:



- Explainable Machine Learning



Human – Machine Learning loop today



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

37

But, But…. The F1 was really high?!

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

38

Unintelligibility

Most AI methods are based on

- Complex nonlinear models over millions of features trained via opaque optimization on unaudited training data
- Search of unverifiably vast spaces

Questions: ***When*** can we trust it?
How can we ***adjust*** it?

Defining Intelligibility

A relative notion.

Ability to Answer Counterfactuals

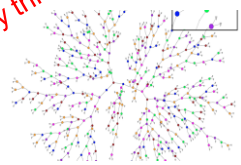
A model is intelligible to the extent that a human can…

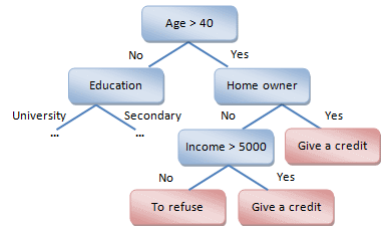
predict how a ***change*** to model's inputs will ***change*** its output

Inherently Intelligible ML – Example 1

Small decision tree over semantically meaningful primitives

Intelligibility threatened if tree grows big

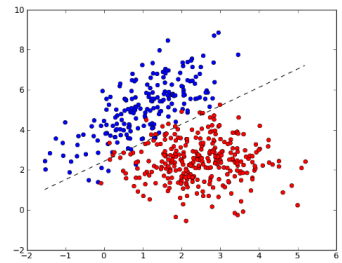




Inherently Intelligible ML – Example 2

Linear model over semantically meaningful primitives

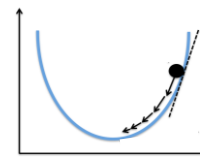
Other models often perform much better



Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. Facilitating User Control
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

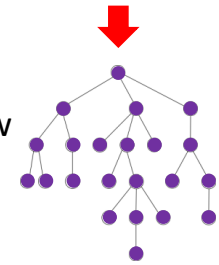
Reasons for Wanting Intelligibility:
1) AI May be Optimizing the Wrong Thing



- Machine Learning: Multi-attribute loss functions
  - HR: how balance predicted job performance, diversity & ethics?
- Optimization example
  - What can go wrong if say 'Maximize paperclip production'?

Personal Assistants

Cortana, Siri, Alexa are Script-Based ☹



AI Planning Allows 'Bots to ***Compose*** New & solve novel problems

Years Ago, We Built a Planning-Based ***Softbot*** …

(Etzioni, Lesh, Segal, Weld, Williamson)

(Even proved it was mathematically sound)

Unpredictability from Search

Hey 'Bot, how much of my disk space is used for my photo collection?

None!
I just executed **rm ***

(Executing this plan used less CPU than counting file sizes)

And now my answer is true!

1. Stupid!

   Should have included "Don't delete files" as a subgoal

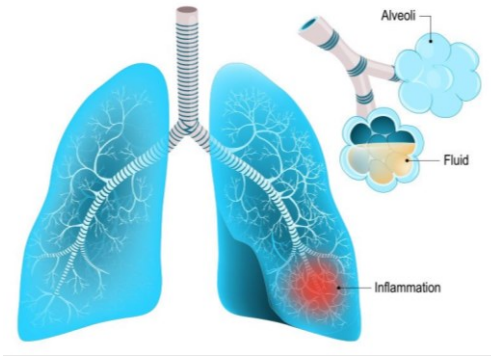
2. Infinite number of such subgoals:

   Qualification Problem

46

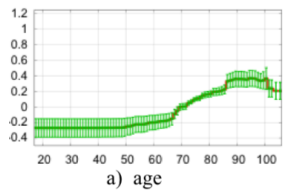
Reasons for Wanting Intelligibility:

2) AI May be Missing a Crucial Feature



Inherently Intelligible ML – Example 3

GA^{2}M model over semantically meaningful primitives $y = \beta_0 + \sum_j f_j(x_j)$

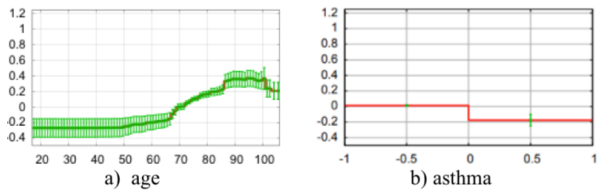


1 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3

GA^{2}M model over semantically meaningful primitives $y = \beta_0 + \sum_j f_j(x_j)$

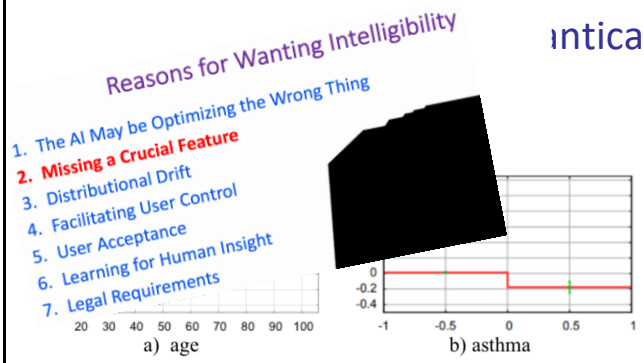


?!?

2 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3



ıntically meaningful

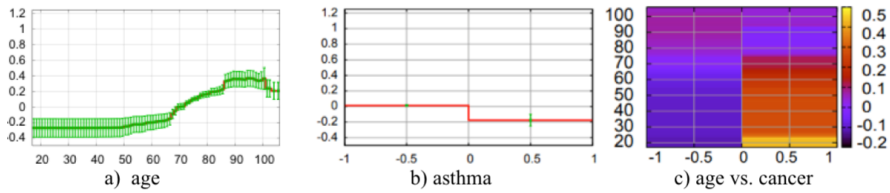
?!?

2 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3

GA^{2}M model over semantically meaningful primitives $y = \beta_0 + \sum_j f_j(x_j) + \underbrace{\sum_{i \neq j} f_{ij}(x_i, x_j)}_{\text{pairwise terms}}$



3 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Reasons for Wanting Intelligibility: #3) Distributional Drift

- System may perform well on training & test distributions…
- But once deployed, distribution often changes
  - Especially from feedback with humans in the loop
  - E.g., filter bubbles in social media

Reasons for Wanting Intelligibility:
#4) Facilitating User Control

Good explanations are ***actionable*** **– *they enable control***

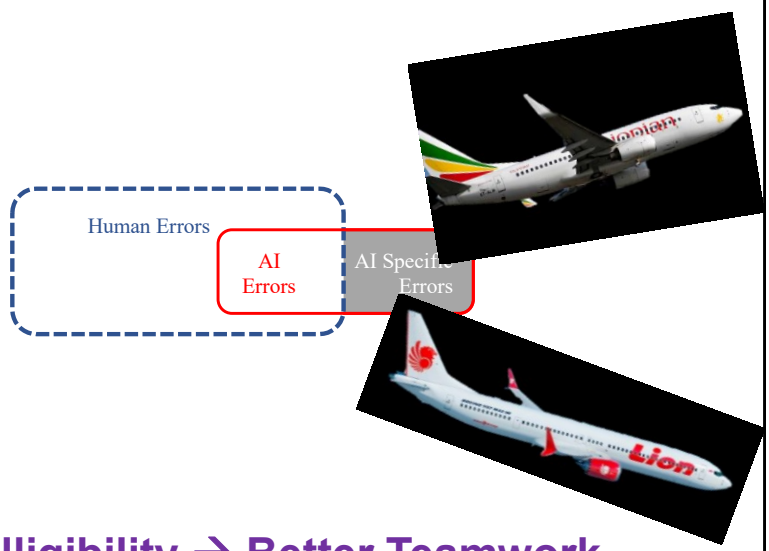
E.g. **Managing preferences**

Google News Feed / Android

- Why in my spam folder?
- Why did you fly me thru Chicago? 'Cause you prefer flying on united

#4 Continued

Era of Human-AI Teams

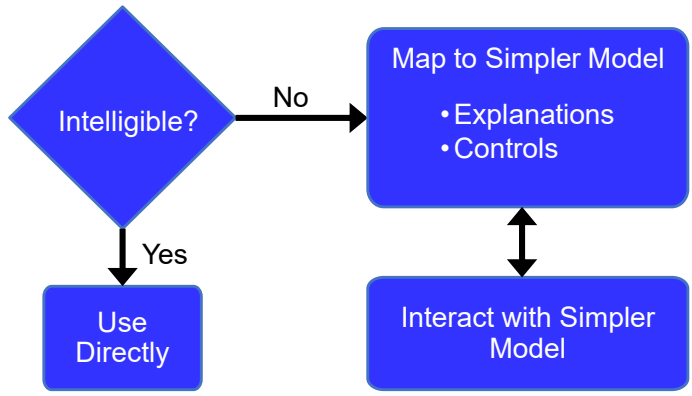


Intelligibility → Better Teamwork

Reasons for Wanting Intelligibility: #7) Legal Imperatives

- GDPR
  Right to an explanation
- Fairness & bias
- Determining liability

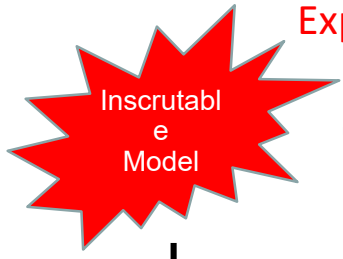
Roadmap for Intelligibility



Reasons for Inscrutability

- Too Complex
  -
  -
- Features not Semantically Meaningful
  -
-

Explaining Inscrutable Models



Simpler Explanatory Model

- Too Complex
  - Simplify by currying –> instance-specific explanation
  - Simplify by approximating
- Features not Semantically Meaningful
  - Map to new vocabulary
- Usually have to do both of these!

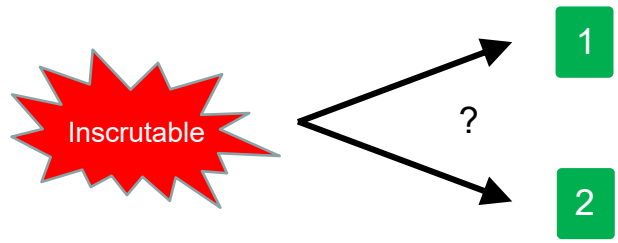
Central Dilemma

Understandable ⟷ **Accurate**

Over-Simplification ⟷ **Inscrutable**

Any model simplification is a ***Lie***

What Makes a Good Explanation?



Need Desiderata

Explanations are Contrastive

Why P ***rather than Q?***

Fact Foil

Q: Amazon, why did you recommend that I rent *Interstellar* ?

A: Because you've liked other movies by Christopher Nolan

Implicit foil Q = some other movie (by another director)

Alternate foil = **buying** *Interstellar*

Explanations as a Social Process

Two Way Conversation

E.g., refine choice of foil…

Grice's Maxims

- Quality be truthful, only relate things supported by evidence
- Quantity give as much info as needed & no more
- Relation only say things related to the discussion
- Manner avoid ambiguity; be as clear as possible

Ranking ←Psychology Experiments

If you can't include all details, humans prefer

- Details distinguishing fact & foil

- Necessary causes >> sufficient ones
- Intentional actions >> actions taken w/o deliberation
- Proximal causes >> distant ones
- Abnormal causes >> common ones

- Fewer conjuncts (regardless of probability)
- Explanations consistent with listener's prior beliefs

Tversky & Kahneman Cognitive Biases

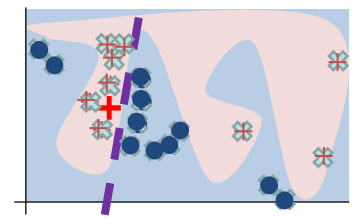
Presenting an explanation made people believe P was true
If explanation ~ previous, effect was strengthened

Actionable

- Prefer expl that is actionable

LIME - Local Approximations [Ribeiro *et al.* KDD16]

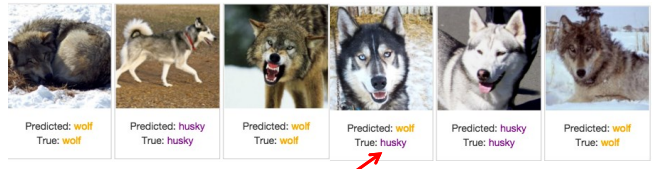
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples (possibly using different features)
5. Use simple model to explain



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

70

Train a neural network to predict wolf *vs.* husky



Only 1 mistake!!!

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

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

71

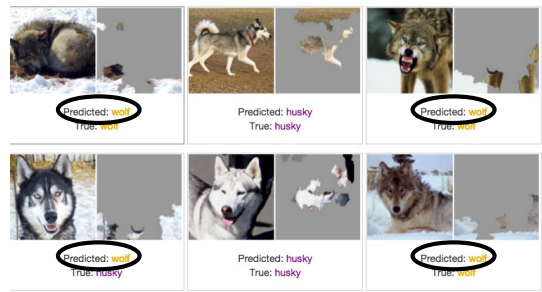
LIME Explanation for Neural Network Prediction

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

72

Approximate ***Global*** Explanation by Sampling

Explanatory Classifier: Logistic Regression
Features: ???



It's a snow detector… ☹

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

73

Semantically Meaningful Vocabulary?

To create ***features*** for explanatory classifier,
Compute `superpixels' using off-the-shelf image segmenter

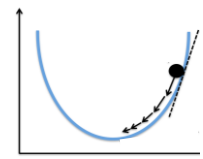
To ***sample*** points around x_i, set some superpixels to grey

Hope that feature/values are semantically meaningful

Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. Facilitating User Control
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

Reasons for Wanting Intelligibility:
1) AI May be Optimizing the Wrong Thing



- Machine Learning: Multi-attribute loss functions
  - HR: how balance predicted job performance, diversity & ethics?
- Optimization example
  - What can go wrong if say 'Maximize paperclip production'?

Reasons for Wanting Intelligibility:
#3) Distributional Drift

- System may perform well on training & test distributions…
- But once deployed, distribution often changes
  - Especially from feedback with humans in the loop
  - E.g., filter bubbles in social media

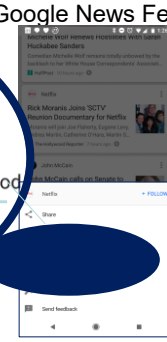
Reasons for Wanting Intelligibility:
#4) Facilitating User Control

Good explanations are ***actionable* – *they enable control***

E.g. **Managing preferences**

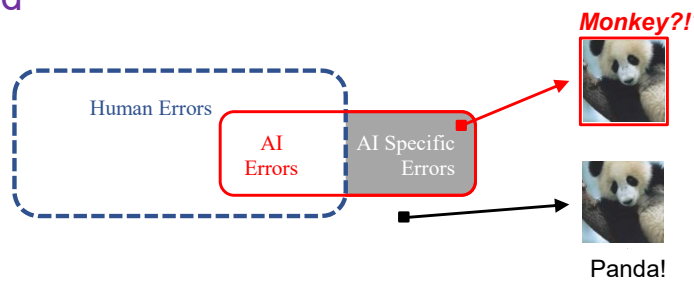
- Why in my spam folder?
- Why did you fly me thru Chicago?
  'Cause you prefer flying on united

Google News Feed



Era of Human-AI Teams

#4 Continued



Intelligibility → Better Teamwork

Outline

- Introduction
- Rationale for Intelligibility
- Defining Intelligibility
- Intelligibility Mappings
- Interactive Intelligibility
- Intelligible Search
- Maintaining Trust

Defining Intelligibility

A relative notion.

Ability to Answer Counterfactuals

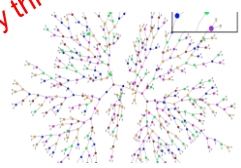
predict how a ***change*** to model's inputs will ***change*** its output

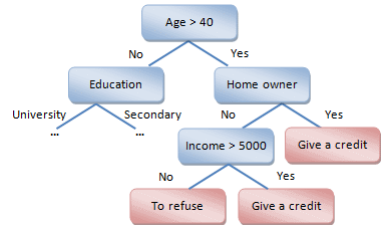
A model is intelligible to the extent that a human can…

Inherently Intelligible ML – Example 1

Small decision tree over semantically meaningful primitives

Intelligibility threatened if tree grows big

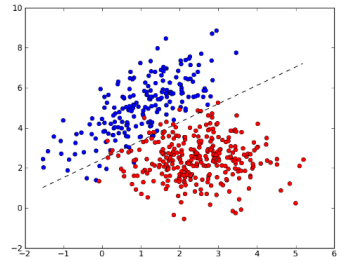




Inherently Intelligible ML – Example 2

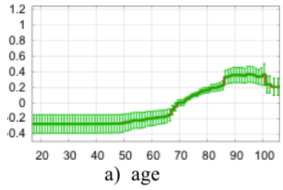
Linear model over semantically meaningful primitives

Other models often perform much better



Inherently Intelligible ML – Example 3

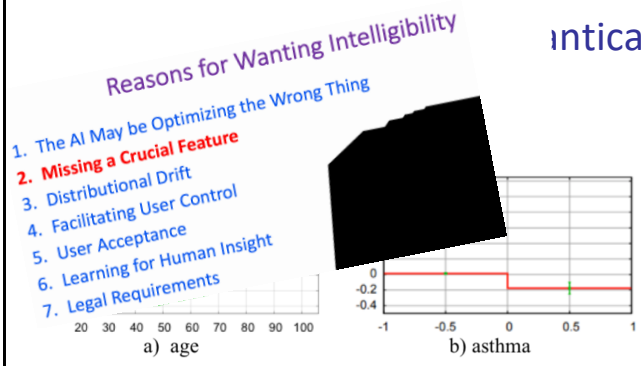
GA^{2}M model over semantically meaningful primitives $y = \beta_0 + \sum_j f_j(x_j)$



1 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3



ıntically meaningful

?!?

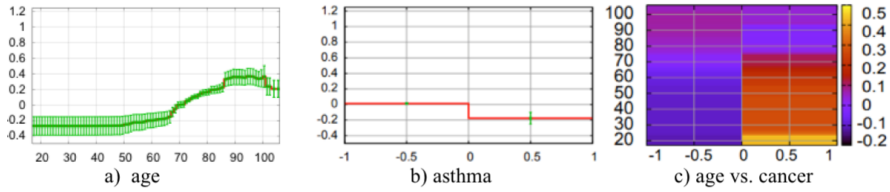
2 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3

GA^{2}M model over semantically meaningful primitives

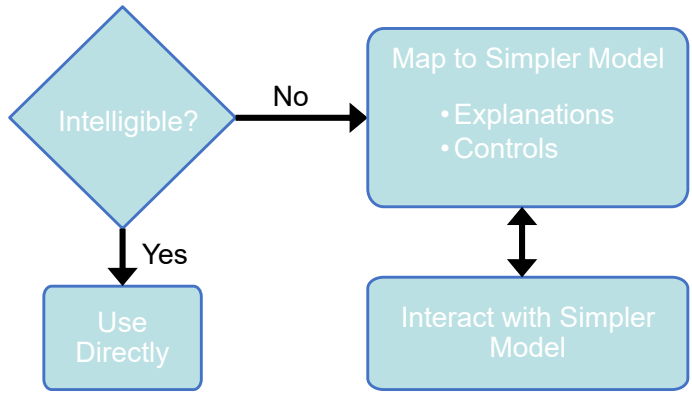
$$y = \beta_0 + \sum_j f_j(x_j) + \underbrace{\sum_{i \neq j} f_{ij}(x_i, x_j)}_{\text{pairwise terms}}$$



3 (of 56) components of learned GA^{2}M

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

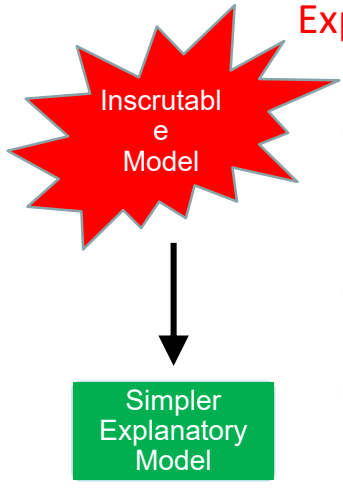
Roadmap for Intelligibility



Reasons for Inscrutability

- Too Complex
  -
  -
- Features not Semantically Meaningful
  -
-

Explaining Inscrutable Models



- Too Complex
  - Simplify by currying –> instance-specific explanation
  - Simplify by approximating
- Features not Semantically Meaningful
  - Map to new vocabulary
- Usually have to do both of these!

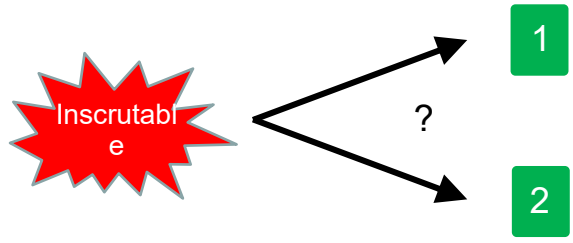
Central Dilemma

Understandable ⟷ **Accurate**

Over-Simplification ⟷ **Inscrutable**

Any model simplification is a ***Lie***

What Makes a Good Explanation?



Need Desiderata

Explanations are Contrastive

Why P ***rather than Q?***

Fact Foil

Q: Amazon, why did you recommend that I rent *Interstellar* ?

A: Because you've liked other movies by Christopher Nolan

Implicit foil Q = some other movie (by another director)

Alternate foil = **buying** *Interstellar*

Explanations as a Social Process

Two Way Conversation

E.g., refine choice of foil…

Ranking ←Psychology Experiments

If you can't include all details, humans prefer

- Details distinguishing fact & foil
- Necessary causes >> sufficient ones
- Intentional actions >> actions taken w/o deliberation
- Proximal causes >> distant ones
- Abnormal causes >> common ones
- Fewer conjuncts (regardless of probability)
- Explanations consistent with listener's prior beliefs

Tversky & Kahneman Cognitive Biases

Presenting an explanation made people believe P was true

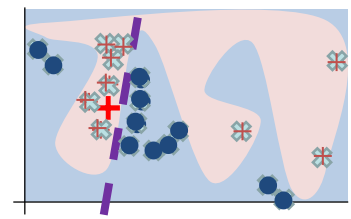
If explanation ~ previous, effect was strengthened

Outline

- Introduction
- Rationale for Intelligibility
- Defining Intelligibility
- Intelligibility Mappings
- Interactive Intelligibility
- Intelligible Search
- Maintaining Trust

LIME - Local Approximations [Ribeiro *et al.* KDD16]

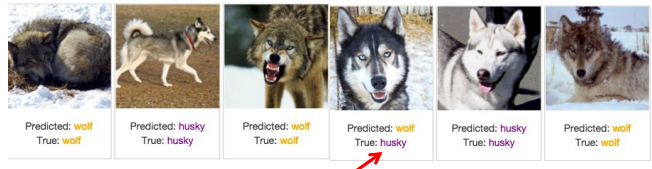
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples (possibly using different features)
5. Use simple model to explain



Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

10
5

Train a neural network to predict wolf *vs.* husky



Only 1 mistake!!!

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

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

106

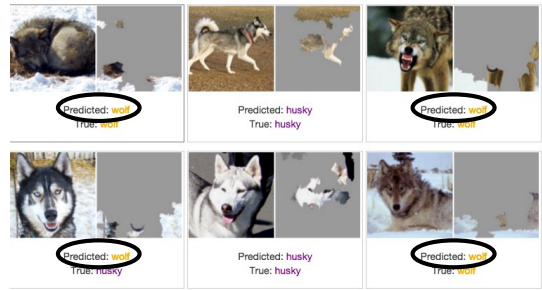
LIME Explanation for Neural Network Prediction

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

107

Approximate ***Global*** Explanation by Sampling

Explanatory Classifier: Logistic Regression
Features: ???



It's a snow detector… ☹

Slide adapted from Marco Ribeiro – see "Why Should I Trust You?: Explaining the Predictions of Any Classifier," M. Ribeiro, S. Singh, C. Guestrin, SIGKDD 2016

108

Semantically Meaningful Vocabulary?

To create ***features*** for explanatory classifier,
Compute `superpixels' using off-the-shelf image segmenter

To ***sample*** points around x_i, set some superpixels to grey

Hope that feature/values are semantically meaningful

Explanations as a Social Process

Gagan Bansal

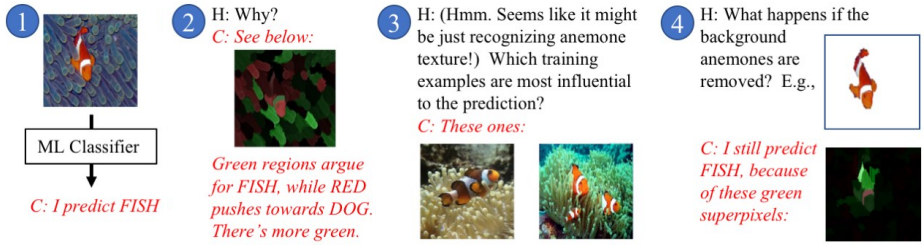
Two Way Conversation

[Weld & Bansal, CACM 201

Dialog Actions

- Redirection by changing the foil
  "Sure, but why didn't you predict class C?"
- Restricting explanation to a sub-region of feature space:
  "Let's focus on short-term, municipal bonds."
- Asking for a decision's rationale: "What made you believe this?"
  System could display the most influential training examples
- Changing explanatory vocabulary by adding (or removing) a feature
  Either from a predefined set, defining with TCAV, or using machine teaching methods
- Perturbing the input example to effect on both prediction & explanation.
  Aids understanding; also useful if affected user wants to contest initial prediction:
  "But officer, one of those prior DUIs was overturned...?"
- Repairing the prediction model
  Use affordances from interactive ML & explanatory debugging

Example



Note: natural-language text is illustrative – system has a GUI (no NLP)

Data Risk

- Quality of ML Output Depends on Data…
- Three Dangers:
  - Training Data Attacks
  - Adversarial Examples
  - Bias Amplification

Attacks to Training Data

Adversarial Examples

\+ 0.007 ×

=

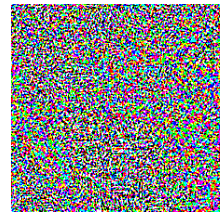
57% Panda

Access to NN parameters

"Explaining and harnessing adversarial examples," I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

11
9

Adversarial Examples


+
0.007 ×

=


57% Panda

99.3% Gibbon

Access to NN parameters

"Explaining and harnessing adversarial examples," I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

120

Adversarial Examples


+
0.007 ×

=


57% Panda

99.3% Gibbon

Only need x Queries to NN

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

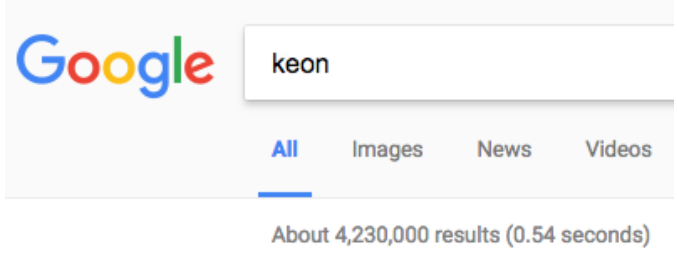
"Explaining and harnessing adversarial examples," I. Goodfellow, J. Shlens & C. Szegedy, ICLR 2015

121

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



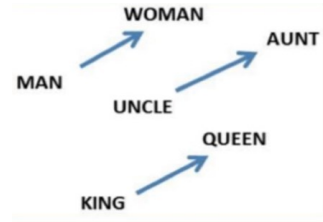
Searches of 'black' first names

Searches of 'white' first names

25% more likely to include ad for criminal-records background check

2013 study https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240

Automating Sexism



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

man : king ↔ woman : queen

sister : woman ↔ brother : man

man : computer programmer ↔ woman : homemaker

man : doctor ↔ woman : nurse

https://arxiv.org/abs/1607.06520

Illustration credit: Abdullah Khan Zehady, Purdue

12
4

In fact…

"Housecleaning Robot"

Google image search returns…

Not…

125

Predicting Criminal Conviction from Driver Lic. Photo

Convicted Criminals

Non-Criminals

?

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

https://arxiv.org/pdf/1611.04135.pdf

12 6

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

- US judges use proprietary ML to predict recidivism risk

NORTHPOINTE

- 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
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

12 7

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

  $Y' \perp\!\!\!\perp A$ i.e. P(Y'=1 |A=0)=P(Y'=1 | A=1)

  Furthermore, if $Y \not\perp\!\!\!\perp A$, it rules out ideal predictor Y'=Y

C. Dwork et al. "Fairness through awareness" ACM ITCS, 214-226, 2012

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

  $Y \perp\!\!\!\perp A \mid Y'$

  No incentive for judge to ask about A

- Equalized odds

  $Y' \perp\!\!\!\perp A \mid Y$ *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

- Can't achieve both!

  Unless $Y \perp\!\!\!\perp A$ 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 ⫫ A | Y'

  No incentive for judge to ask about A

- Equalized odds

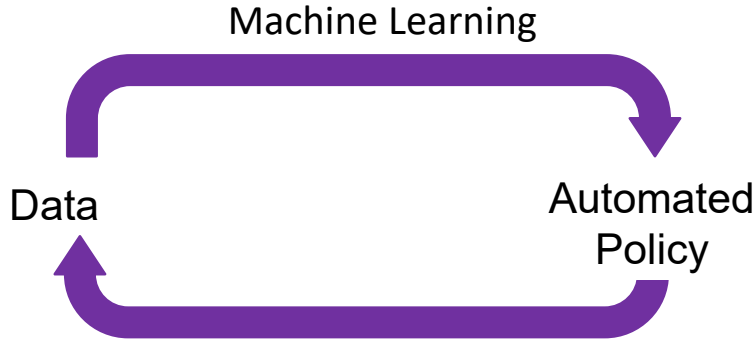
  Y' ⫫ A | Y *i.e.* ∀y, P(Y'=1 | A=0, Y=y) = P(Y'=1 | A=1, Y=y)

  Same rate of false positives & negatives

M. Hardt *et al* "Equality of Opportunity in Supervised Learning" arXiv:1610.02413v1

Important to get this Right!

Feedback Cycles



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)?

http://defamer.gawker.com/the-colbert-reports-new-twitter-feed-praising-fox-news-1458817943

13
4

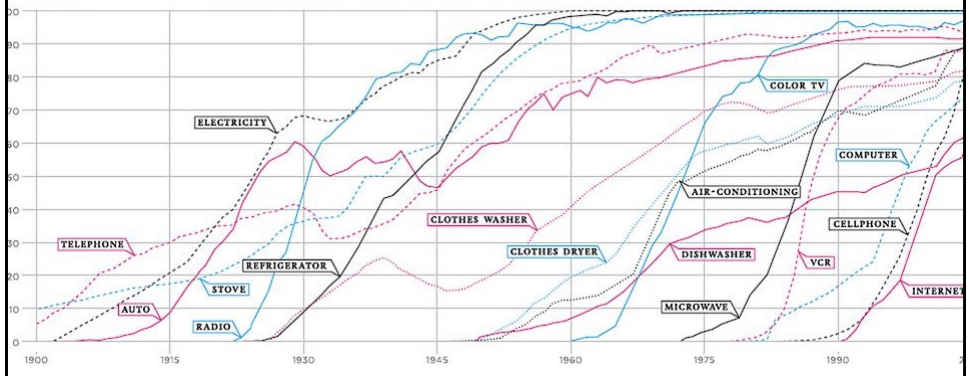
Understanding Limitations

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

13
5

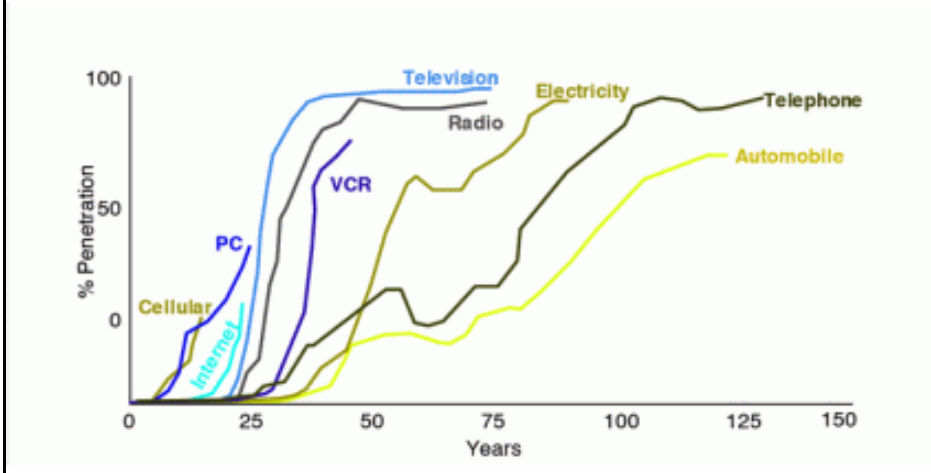
Exponential Growth →
Hard to Predict Tech Adoption



13
6

Adoption Accelerating

Newer technologies taking hold at double or triple the rate

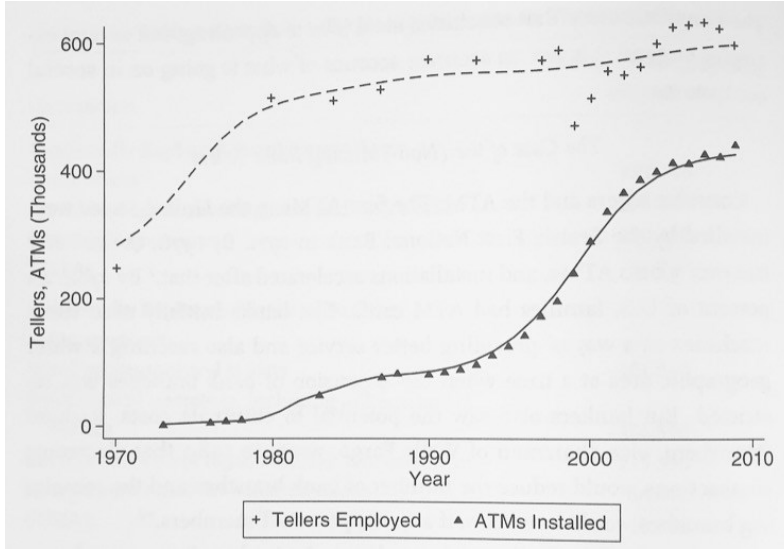


Self-Driving Vehicles

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

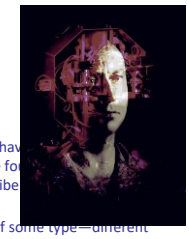
138

Hard to Predict



http://www.aei.org/publication/what-atms-bank-tellers-rise-robots-and-jobs/

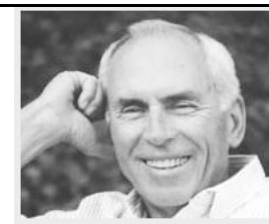
140



- To appreciate the challenges ahead of us, first consider four basic capabilities that any true AGI would hav such capabilities are fundamental to our future work toward an AGI because they might have been the fo emergence, through an evolutionary process, of higher levels of intelligence in human beings. I'll describe children can do.
- The object-recognition capabilities of a 2-year-old child. A 2-year-old can observe a variety of objects of some type—different kinds of shoes, say—and successfully categorize them as shoes, even if he or she has never seen soccer cleats or suede oxfords. Today's best computer vision systems still make mistakes—both false positives and false negatives—that no child makes.
- The language capabilities of a 4-year-old child. By age 4, children can engage in a dialogue using complete clauses and can handle irregularities, idiomatic expressions, a vast array of accents, noisy environments, incomplete utterances, and interjections, and they can even correct nonnative speakers, inferring what was really meant in an ungrammatical utterance and reformatting it. Most of these capabilities are still hard or impossible for computers.
- The manual dexterity of a 6-year-old child. At 6 years old, children can grasp objects they have not seen before; manipulate flexible objects in tasks like tying shoelaces; pick up flat, thin objects like playing cards or pieces of paper from a tabletop; and manipulate unknown objects in their pockets or in a bag into which they can't see. Today's robots can at most do any one of these things for some very particular object.
- The social understanding of an 8-year-old child. By the age of 8, a child can understand the difference between what he or she knows about a situation and what another person could have observed and therefore could know. The child has what is called a "theory of the mind" of the other person. For example, suppose a child sees her mother placing a chocolate bar inside a drawer. The mother walks away, and the child's brother comes and takes the chocolate. The child knows that in her mother's mind the chocolate is still in the drawer. This ability requires a level of perception across many domains that no AI system has at the moment.

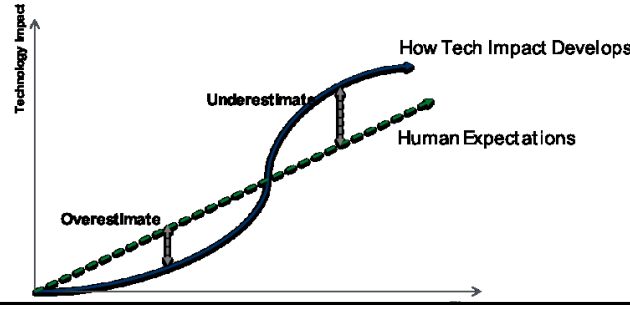
141

Amara's Law



We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run

Roy Amara



142

Conclusions

- Distractions *vs.*
- Important Concern
  - Sorcerer's Apprenti
    - Specifying Constrain
    - 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

14
5

- Burger King is releasing a TV ad intended to deliberately trigger Google Home devices to start talking about Whopper burgers, according to Buz tly to the ca hopper burger

14
8

- Inverse revinforcement 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

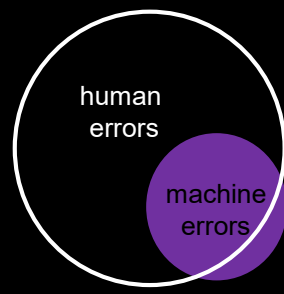
  - Cooperative inverse RL
    - Two player game
- Off swicth function
  - Don't given robot an objective
  - Instead it must allow for uncertainty about human objctive
    - If human is trying to turn me off, then it must want that
- Uncertainty in objectives – ignored
  - Irrelevant in standard decision problems; unless env provides info on reward

154

DEPLOYING AI

What is bar for deployment?

- System is better than person being replaced?
- Errors are ***strict subset*** of human errors?



155

- Reward signals
  - Wireheading
  - RL agent hijacks reward
  - Traditiomnal RL
    - Enivironment provide reward signal. Mistak!
  - Instead env reward signal is not true reward
    - Just provides INOFRMATION about reward
  - So hijacking reward signal is pointless
    - Doesn't provide more reward
    - Just provides less information

- Y Lecunn – common view
- All ai success is supervised (deep) MLL
- Unsupervised is key challenge
  - Fill in occluded immage
  - Fill in missing words in text, sounds in speech
  - Consquences 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
  - Supervisesd 10-10000 bits trial
  - Unsupervise – 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)

165