Beneficial AI

Daniel S. Weld

Outline

- Distractions
- Important Concerns
  - Unemployment
  - Sorcerer's Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - Deployment
  - It's the Data, Stupid

Please Review CSE 473

- https://uw.iasystem.org/survey/167470

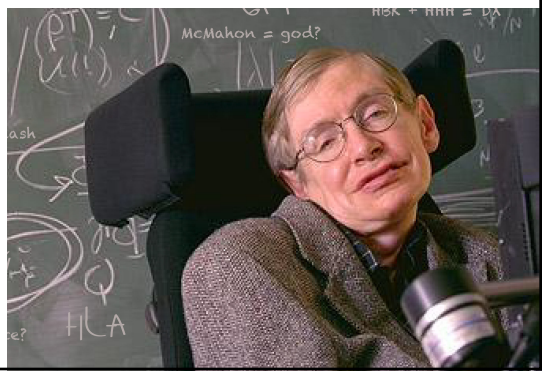
- **5pt bonus** for taking survey!
  - We can tell who has taken it (for bonus)
  - But we can't see your answers

  - In Jan we get aggregated data

4

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

6

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

7

Superhuman AI & Intelligence Explosions

- When will computers have superhuman capabilities?

- Now.
  - Multiplication
  - Spell checking
  - Chess, Go

  - Many more abilities to come

8

AI Systems are ***Idiot Savants***

- Super-human here & super-stupid there

- Just because AI gains another superhuman skill… Doesn't mean it is suddenly good at ***everything***

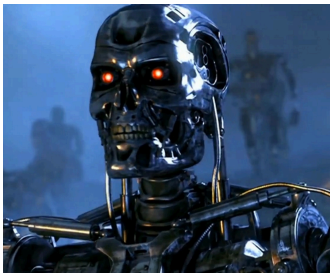
  And certainly not unless we give it **experience** *at everything*

- AI systems will be spotty for a long time

9

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…
- Virtually certain than a bad human will tell an AI to kill us!

10

There will be ***MANY*** Fielded AI systems

- The best defense against a bad AI…
- Will be a good AI…
  - Etzioni's ***Guardian systems***
  - AIs to watch and monitor other AIs.

11

There will be ***MANY*** Fielded AI systems

- The best defense against a ***bad AI…***
- Will be a ***good AI…***
  - Etzioni's ***Guardian systems***
  - AIs to watch and monitor other AIs.

12

No

- There are ***many*** scary things coming our way
- But Skynet & Intelligence Explosion aren't the issue
- They are a ***dangerous distraction*** from real issues

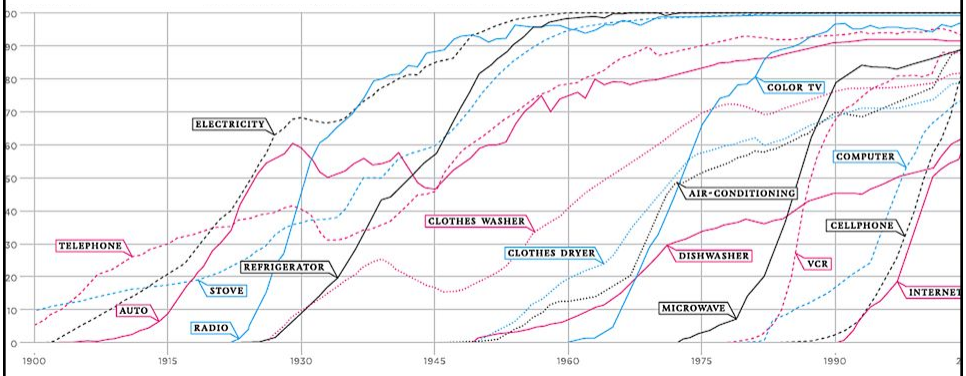
13

Real Issues

- Unemployment
- Sorcerer's Apprentice
  - Specifying Constraints & Utilities
  - Explainable AI
- Deployment
- It's the Data, Stupid

14

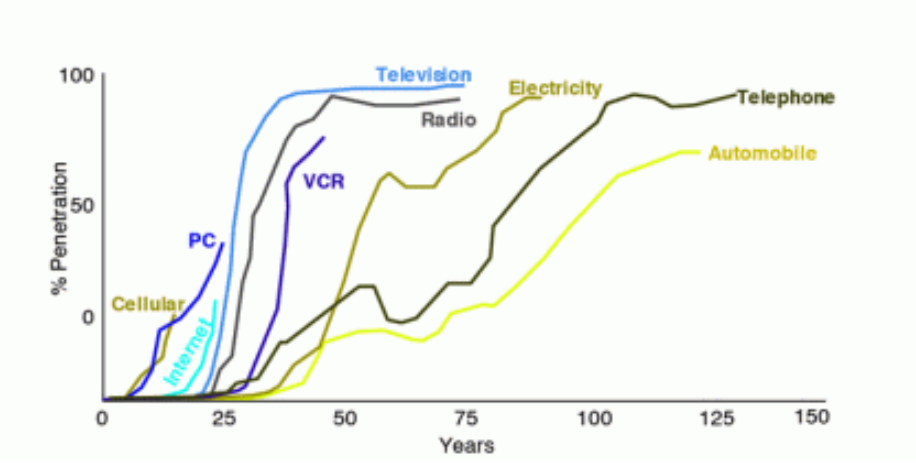
Hard to Predict Tech Adoption
Exponential Growth



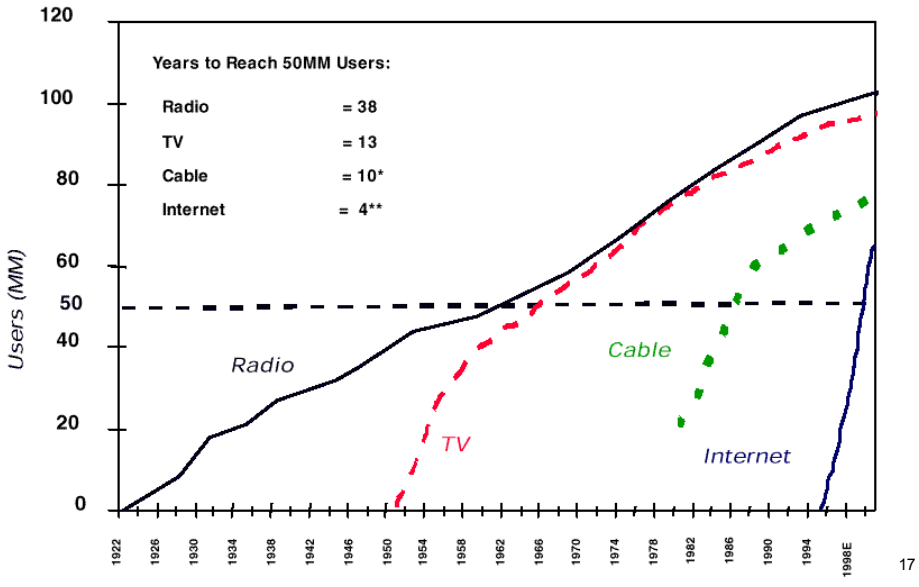
15

Adoption

Newer technologies taking hold at double or triple the rate



Adoption Accelerating



17

Self-Driving Vehicles

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

↑ Inequity → revolution?

18

Real Issues

- Unemployment
- Sorcerer's Apprentice
  - Specifying Constraints & Utilities
  - Explainable AI
- Deployment
- It's the Data, Stupid

20

Sorcerer's Apprentice

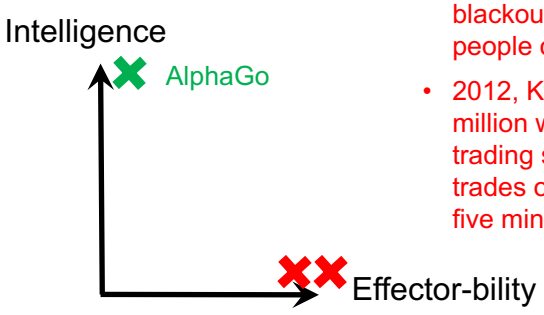
Tired of fetching water by pail, the apprentice enchants a broom to do the work for him – using magic in which he is not yet fully trained. The floor is soon awash with water, and the apprentice realizes that he cannot stop the broom because he does not know how.

AI assistants may hurt us ***accidentally***, while (thinking that) they are obeying our orders.

21

Brains Don't Kill

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

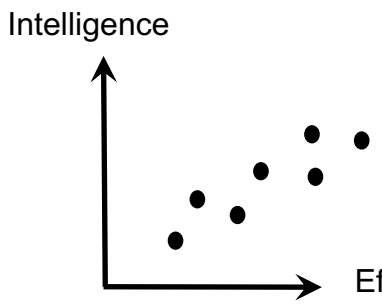


- 2003, an error in General Electric's power monitoring software led to a massive blackout, depriving 50 million people of power.
- 2012, Knight Capital lost $440 million when a new automated trading system executed 4 million trades on 154 stocks in just forty-five minutes.

22

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

23

Unpredictability

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

None, Dave!
I just executed rm *

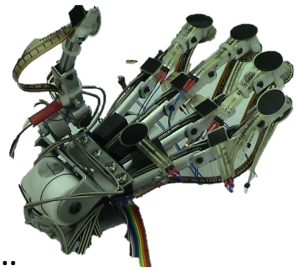
(It was easier than counting file sizes)

24

Physically-Complete Effectors

- Roomba effectors close to harmless
- Bulldozer blade ∨ 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

25

Specifying Utility Functions

Clean up as much dirt as possible!

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

26

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.

27

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!

28

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.

29

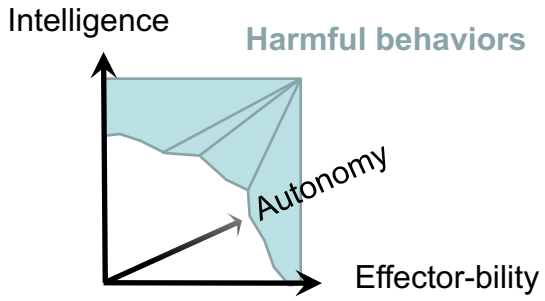
Specifying Utility Functions

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

30

A Possible Solution: Constrained Autonomy?

Restrict an agents behavior with background constraints



31

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.

32

But what *is* Harmful?

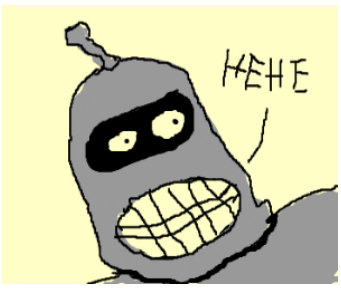
1. A robot may not ***injure*** a human being or, through inaction, allow a human being to come to ***harm***.

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

33

Trusting AI

- How can a user teach a machine what is harmful?
- How can they know when it really understands?
  - Especially hard given deep neural networks
- Explainable Machine Learning



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

35

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

- US judges use proprietary ML to predict risk of reoffending
- Much more likely to mistakenly flag black defendants
  - Even though race is not used as a feature
- Bigger questions:
  - Can defendant get an explanation?
    - Tradeoff between explainability & accuracy
    - How know if explanation is ***right?***
  - What if blacks ***are*** more likely to reoffend?
    - Ok to treat them differently?
    - Or is poor accuracy the only problem?
  - Whose responsibility to monitor?
  - What if feedback cycle?

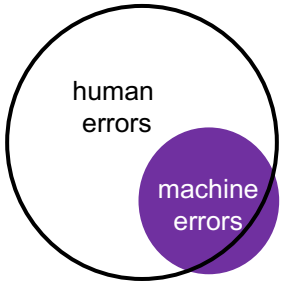
http://go.nature.com/29aznyw
https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing#.odaMKLgrw

38

Deploying AI

What is bar for deployment?

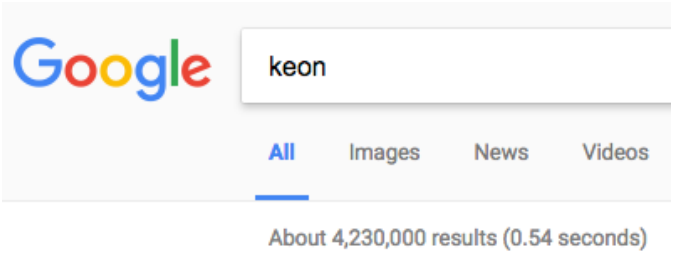
- System is better than person being replaced?
- Errors are ***strict subset*** of human errors?



40

49

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

50

Automating Sexism

- Word embeddings
- Word2vec trained on 3M words from Google news corpus
- Used in machine translation & analogical reasoning

man : king ↔ woman : queen

sister : woman ↔ brother : man

man : computer programmer ↔ woman : homemaker

man : doctor ↔ woman : nurse

https://arxiv.org/abs/1607.06520

51

Liability?

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

- Legal standard
  - Criminal intent
  - Negligence

Deploying AI → criminal acts without a perpetrator
– Ryan Calo

52

Liability II

- Stephen Cobert's twitter-bot
  - Substitutes names of FoxNews personalities into Rotten T movie reviews
  - One tweet implied Bill Hemmer took communion while intoxicated.
- Is this libel (defamatory speech)?

53

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

Conclusions

- Distractions
- Important Concerns
  - Unemployment
  - Sorcerer's Apprentice Scenario
    - Specifying Constraints & Utilities
    - Explainable AI
  - When to deploy?
    - Liability?
    - Responsibility for monitoring?
  - Biased Data

55