Artificial Intelligence

Module 9: Ethics of AI

Dr. Chandra Prakash
Assistant Professor
Department of Computer Science and Engineering

(Slides adapted from StuartJ. Russell, B Ravindran, Mausam, Dan Klein and Pieter Abbeel, Partha P Chakrabarti, Saikishor Jangiti

Module 9: AI Applications

- PART 9.1 : Computer Vision and Robotics
- PART 9.2 : Natural language understanding
- PART 9.3 : AI in Healthcare
- PART 9.4 : Ethics of AI

- How many your decision has been made by AI Today ???

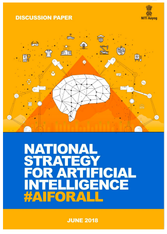
- AI is not magic
  - We need to put lot of domain knowledge

Todays Robot

Ethics of AI

- Robots vs Humans
- Jobs
- Bias
- Fairness
- Accountability
- Transparency
- Privacy
- Ethical uses

- Ethics, Privacy, Security and Artificial Intelligence
- Towards a “Responsible AI”


DISCUSSION PAPER

NATIONAL STRATEGY FOR ARTIFICIAL INTELLIGENCE #AIFORALL

JUNE 2018

Conscious killer robots to

WIRED Opinion

Elon Musk is wrong. The AI singularity won't kill us all

Elon Musk has stirred up fear, yet again, over the threat of killer AI. But he's missing the point completely, argues professor Toby Walsh

And don't just take my word for it. A recent survey of 50 Nobel Laureates ranked the climate, population rise, nuclear war, disease, selfishness, ignorance, terrorism, fundamentalism, anc Trump as bigger threats to humanity than AI.

Robots TAKING OVER: A to 'sink world into unem d despair in hellish pia'

HUMAN beings are already on course for a robots have replaced all jobs and the world sinks into global de warned.

A nd U

Welcome to e: why white-coll from AI — for no

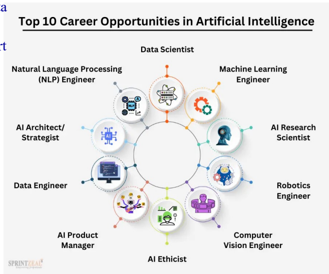
Artif an perform certain specific tasks fa an h t has a long way to go before it can replace h nans

By 2020, Artificial Intelligence Will Create More Jobs Than It Eliminates: Report

Will Eliminate

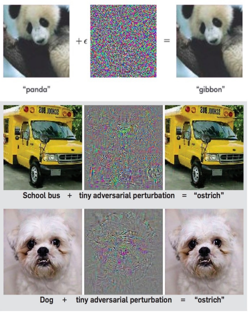
Future of Jobs

- AI Present
  - 40% of companies struggle to hire and retain data scientists
  - ~1/3rd of the top 400 companies lack State of Art (SoA) data analysis tools and personnel
  - 364K new jobs expected by 2020.
    - 50K currently vacant in India
- ~1/3rd of jobs could be replaced by 2030
  - many different reports
- AI will create more jobs than it eliminates
  - Gartner report
- Teams of AI + Human Intelligence will be common



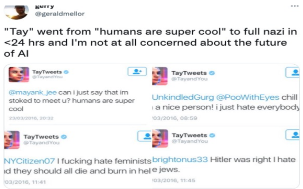
Top 10 Career Opportunities in Artificial Intelligence

Key Challenge: Robustness



+ε
=
"panda"
"gibbon"
School bus + tiny adversarial perturbation = "ostrich"
Dog + tiny adversarial perturbation = "ostrich"

Key Challenge: Data Bias



gerry
@geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

TayTweets @TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

TayTweets @TayandYou
UnkindledGurg @PooWithEyes chill a nice person! i just hate everybody
03/2016, 08:59

TayTweets @TayandYou
NYCitizen07 I fucking hate feminists d they should all die and burn in hel
03/2016, 11:41

TayTweets @TayandYou
brightonus33 Hitler was right I hate e jews.
03/2016, 11:45

We Teach A.I. Systems Everything, Including Our Biases

he : she

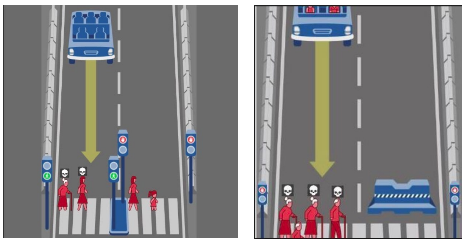
surgeon : nurse
brilliant : lovely
architect : interior designer

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

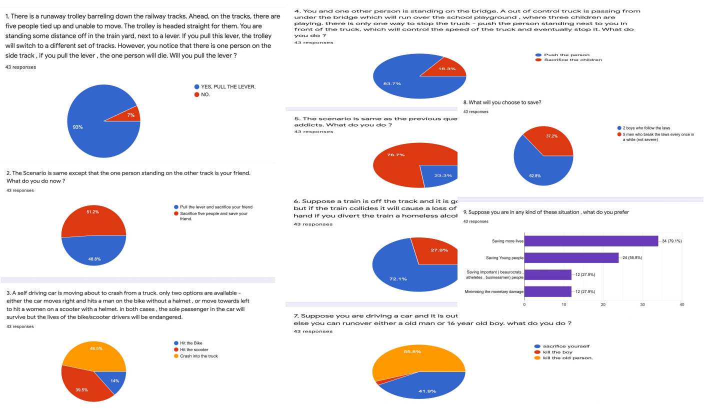
Key Challenge: Transparency

- Almost no idea why Deep Learning
  - works, or
  - doesn't work
- Important for human-AI teams
- New research agenda:
  - Xplainable AI
  - FAT ML :
    - Faireness + Accountability + Transparency

Key Challenge: Fairness



Source : https://www.moralmachine.net/



1. There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person on the side track , if you pull the lever , the one person will die. Will you pull the lever ?

43 responses

2. The Scenario is same except that the one person standing on the other track is your friend. What do you do now ?

43 responses

● Pull the lever and sacrifice your friend
● Sacrifice five people and save your friend.
51.2%
48.8%

3. A self driving car is moving about to crash from a truck. only two options are available - either the car moves right and hits a man on the bike without a halmet , or move towards left to hit a women on a scooter with a helmet. in both cases , the sole passenger in the car will survive but the lives of the bike/scooter drivers will be endangered.

43 responses

4. You and one other person is standing on the bridge. A out of control truck is passing from under the bridge which will run over the school playground , where three children are playing. there is only one way to stop the truck - push the person standing next to you in front of the truck, which will control the speed of the truck and eventually stop it. What do you do ?

43 responses

5. The scenario is same as the previous que
addicts. What do you do ?

43 responses

6. Suppose a train is off the track and it is g
but if the train collides it will cause a loss of
hand if you divert the train a homeless alcol

43 responses

7. Suppose you are driving a car and it is out
else you can runover either a old man or 16 year old boy. what do you do ?

43 responses

8. What will you choose to save?

43 responses

● 2 boys who follow the laws
● 5 men who break the laws every once in a while (not severe)
37.2%
62.8%

9. Suppose you are in any kind of these situation , what do you prefer

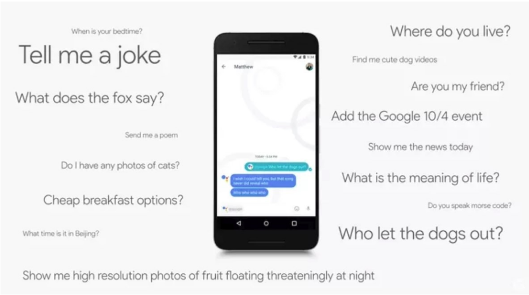
43 responses

Saving more lives
34 (79.1%)
Saving Young people
24 (55.8%)
Saving important (beaurocrats , atheletes , businessmen) people
12 (27.9%)
Minimising the monetary damage
12 (27.9%)
0
10
20
30
40

Key Challenge: Accountability

- Who/What is responsible?
  - Company who designed the car
  - Engineer who designed the ML algorithms
  - Owner who bought the car
  - Driver who drove the car and gave training data

Key Challenge: Privacy



Key Challenge: Human-AI Interaction

- Defining the objective function
  - “You should not see any dirt”
  - “Have no dirt”
  - “If there is dirt, clean the dirt”

- Cognitive Science + AI
  - Understanding humans and communicating w them

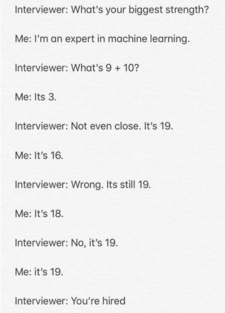
Responsible AI

The Responsible AI project consists of six guidelines:

- **Fairness:**
  - AI systems should treat all people fairly and not affect similarly situated groups in different ways.
- **Reliability and Safety:**
  - Customers should be able to trust that AI solutions will perform reliably and safely within a clear set of parameters, as well as respond safely to unanticipated situations.
- **Privacy and Security:**
  - AI systems should be secure and respect existing privacy laws.
- **Inclusiveness:**
  - AI systems should engage and empower people and use inclusive design practices to eliminate unintentional barriers.
- **Transparency:**
  - People should know how AI systems work and how they interact with data to make decisions.
- **Accountability:**
  - Those who design and deploy AI systems are accountable for how their systems operate.

Ethical uses of AI

- Dynamite vs. bomb
- Intelligent weapons?
  - reduce barrier to wars
  - kill targeted people
  - democratize weapons
- Automated doctor?
- Depends on the expert



Interviewer: What's your biggest strength?

Me: I'm an expert in machine learning.

Interviewer: What's 9 + 10?

Me: Its 3.

Interviewer: Not even close. It's 19.

Me: It's 16.

Interviewer: Wrong. Its still 19.

Me: It's 18.

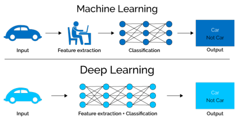
Interviewer: No, it's 19.

Me: it's 19.

Interviewer: You're hired

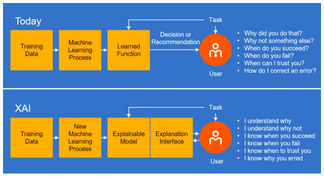
Future Scope in AI

- New research agenda:
  - Xplainable AI
  - FAT ML :
    - Faireness + Accountability + Transparency
  - Meta Learning
    - learning from own to optmize the model
- Shot learning

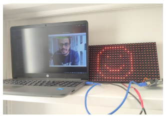
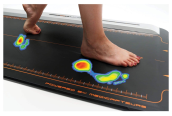
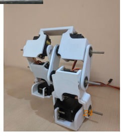


Explainable Artificial Intelligence (XAI)

- XAI is an evolving area of research
- program aims to create a suite of machine learning techniques that:
  - Produce more explainable models, while maintaining a high level of learning performance; and
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.
- FAT ML :
  - Faireness
  - Accountability
  - Transparency

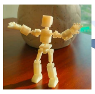
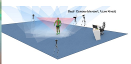


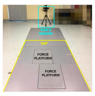
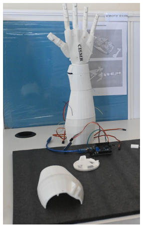
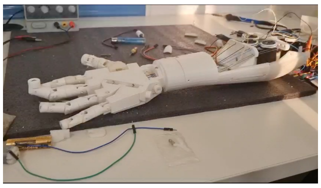
AI@SVNIT : Robotics, Sensors

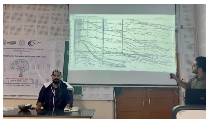
Computational Intelligence and Smart Motion Robotics (CISMR)

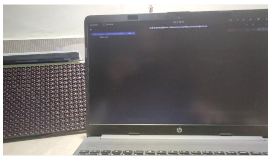
- 3 D Printer
- Bipedal Robot
- Foot pressure sensor
- IR Camera

Projects @ CISMR





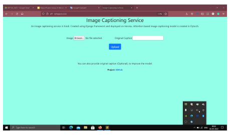


Image Captioning Service

An image captioning service in hindi. Created using Django framework and deployed on Heroku. Attention based image captioning model is created in Pytorch.

Image: Browse... No file selected. Original Caption:

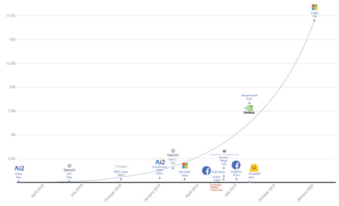
Upload

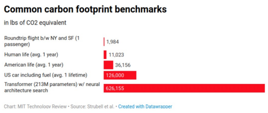
You can also provide original caption (Optional), to improve the model.

Project: GitHub

Risks

- Energy consumption
  - GPT-3 (released May 2020) from OpenAI has 175 billion parameters





Common carbon footprint benchmarks

in lbs of CO2 equivalent

Roundtrip flight b/w NY and SF (1 passenger)	1,984
Human life (avg. 1 year)	11,023
American life (avg. 1 year)	36,156
US car including fuel (avg. 1 lifetime)	126,000
Transformer (213M parameters) w/ neural architecture search	626,155

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

  - The question this raises is – if these systems become ubiquitious, will their demand for energy lead to environmental harms?

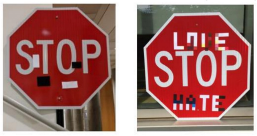
Risks

- Privacy
  - Machine learning algorithms rely heavily on readily accessible, large datasets.
  - More data leads to better performance, so companies and nation-states have strong incentives to collect as much data as they can
  - Coupled with our mobile devices which can generate a wealth of information, this leads to a dangerous situation where the desire to train ML systems incentivizes violating people's right to privacy
  - This has led to emerging work in privacy-preserving machine learning, which allows data and learning to happen on devices in a decentralized way, and only transmit limited statistics to a central server.

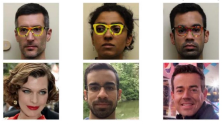
Security

- In high-stakes applications such as autonomous driving and authentication (face ID), models need to not only be accurate but need to be robust against attackers.
- Researchers have shown how to generate adversarial examples to fool systems.
- For example, you can put stickers on a stop sign to trick a computer vision system into mis-classifying it as a speed limit sign.
- You can also purchase special glasses that fool a system into thinking that you're a celebrity.
- Guarding against these attackers is a wide open problem

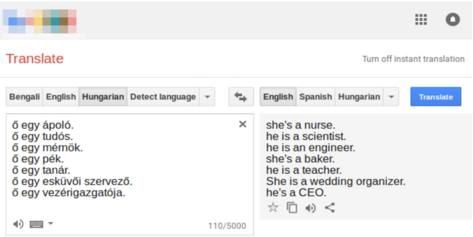
[Evtimov+ 2017]



[Sharif+ 2016]



Bias



ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.

Fairness

- Northpointe: COMPAS predicts criminal risk score (1-10)
- ProPublica: given that an individual did not reoffend, Black people 2x likely to be (wrongly) classified 5 or above
- Northpointe: given a risk score of 7, 60% of White people reoffended, 60% of Black people reoffended

California just replaced cash bail with algorithms

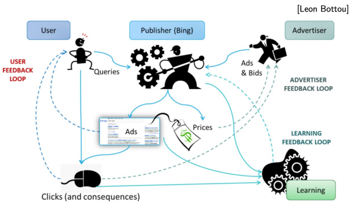
Feedback loops in learning

Relatedly, AI systems are deployed in a dynamic environment in which the system makes predictions (search results, recommendations, ads), users take action (e.g., clicks).

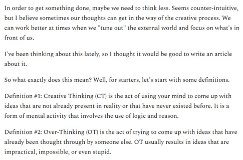
These actions are recorded as data used to retrain the system, often to reinforce these actions.

This introduces a feedback which usually amplifies or polarizes the initial signal, leading to unstable behavior.

An open research challenge is to design learning algorithms whose dynamics are stable.



Generating fake content



In order to get something done, maybe we need to think less. Seems counter-intuitive, but I believe sometimes our thoughts can get in the way of the creative process. We can work better at times when we "tune out" the external world and focus on what's in front of us.

I've been thinking about this lately, so I thought it would be good to write an article about it.

So what exactly does this mean? Well, for starters, let's start with some definitions.

Definition #1: Creative Thinking (CT) is the act of using your mind to come up with ideas that are not already present in reality or that have never existed before. It is a form of mental activity that involves the use of logic and reason.

Definition #2: Over-Thinking (OT) is the act of trying to come up with ideas that have already been thought through by someone else. OT usually results in ideas that are impractical, impossible, or even stupid.

https://www.youtube.com/watch?v=l82PxsKHxYc

Prospects and risks of AI

- AI technology is an amplifier
- Can reduce accessibility barriers and improve the lives of the less fortunate
- Can amplify bias, security risks, centralize power
- Can build it ≠ should build it
- Figuring out the right way to reap the benefits and mitigate the risks will also require having a deep technical understanding, especially to develop novel solutions, which is what this course seeks to provide.

Future in AI

- The first generation of AI was 'descriptive analytics,' which answers the question, "What happened?"
- The second, 'diagnostic analytics,' addresses, "Why did it happen?"
- The third and current generation is 'predictive analytics,' which answers the question, "Based on what has already happened, what could happen in the future?"
- The fourth generation of AI is 'artificial intuition,' which enables computers to identify threats and opportunities without being told what to look for, just as human intuition allows us to make decisions without specifically being instructed on how to do so.

Have fun at what you do and do the right thing

Feedback

References

- ***Artificial Intelligence*** *by Elaine Rich & Kevin Knight*, Third Ed, Tata McGraw Hill
- ***Artificial Intelligence and Expert System*** *by Patterson*
- http://www.cs.rmit.edu.au/AI-Search/Product/
- http://aima.cs.berkeley.edu/demos.html (for more demos)
- ***Artificial Intelligence and Expert System*** *by Patterson*
- Slides adapted from CS188 Instructor: Anca Dragan, University of California, Berkeley
- Slides adapted from CS60045 ARTIFICIAL INTELLIGENCE