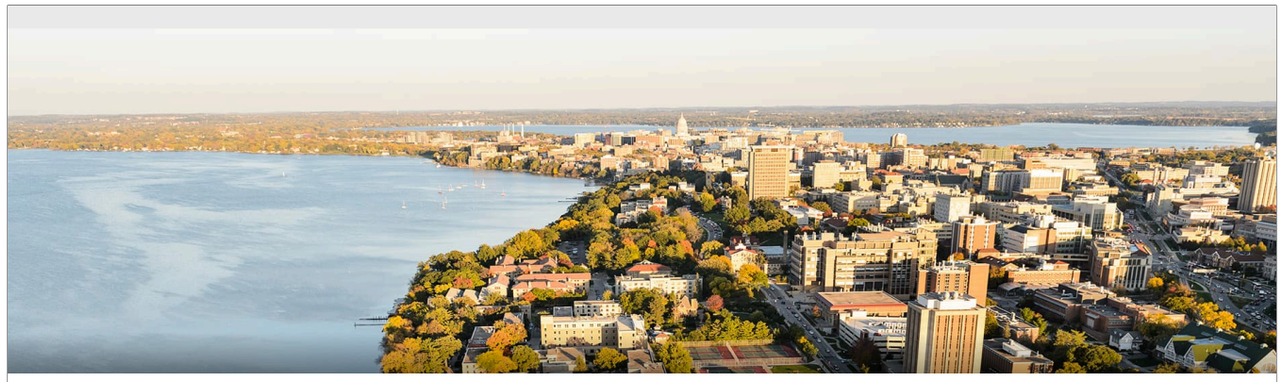


# CS540 Introduction to Artificial Intelligence

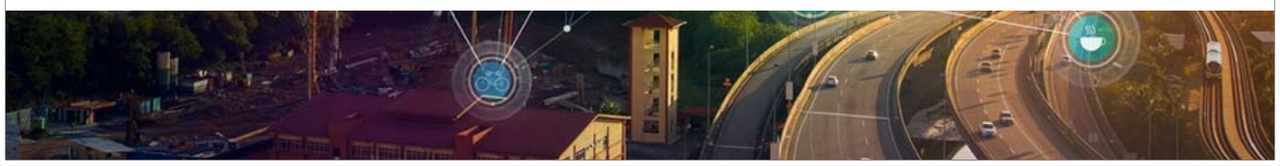
## Ethics and Trust in AI

Yingyu Liang
University of Wisconsin-Madison

**Dec 14, 2021**

Slides created by Sharon Li [modified by Yingyu Liang]

# Artificial Intelligence in Society



We have learned about some AI techniques. But we haven't considered the impact when they're deployed in society.
The goal of AI is to build systems that can solve intelligent tasks. Once deployed, these systems take up work of human beings and can have various consequences on our society and our lives.

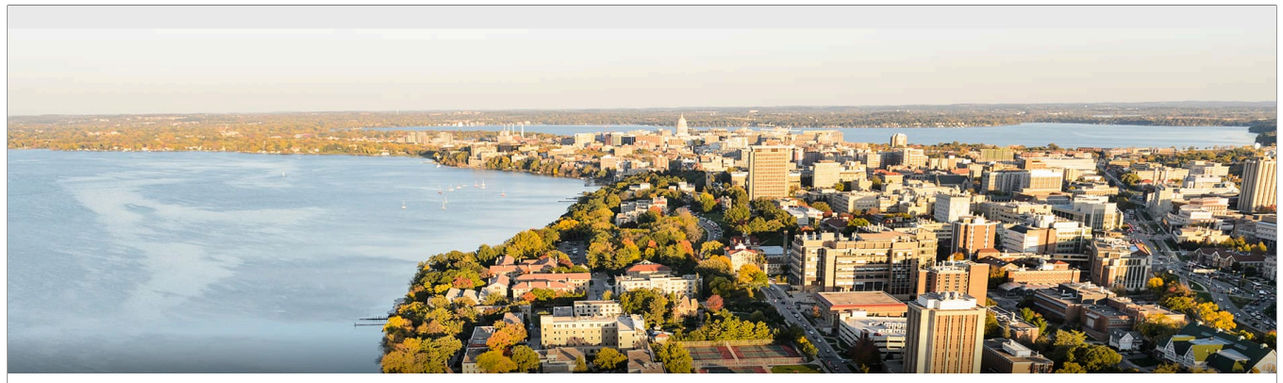
Imagine the following scenarios. You get up in the morning and your bed reports your total sleeping time, the time you've been into deep sleeping status, and other sleeping related statistics. You get on a bus to school, which is operated by an automatically driving system. You submit your applications to your dream graduate schools, which may be evaluated by an AI reviewer. Maybe after 10 years, all graduate schools will use some AI systems to help their admission process. Maybe at that time, a school without a great AI reviewer is not a strong school in AI. Now, after submitting your application, you relax a bit by watching a movie. The movie may be produced by AI, no longer by actual human beings.

Even when AI systems are designed for social good, they can still lead to various subtle ethical issues. When your bed helps monitor your sleeping status, you may not want to share these private data with the public. When you're in a self-driving car, you want the car to be safe under various conditions. When your applications are checked by AI reviewers,  you would like the AI reviewers to be fair such that they do not have bias w.r.t. say the gender or race. When you watch videos produced by AI, you don't want the AI produce fake videos of you saying something you never said. You can think of many other possible ethical issues.

## Outline

- Bias and Fairness
- Fake Content
- Privacy
- Adversarial robustness

In this lecture, we will cover some prototypical examples of ethical issues in AI. These by no means cover all.



# Bias and Fairness

# Example 1: Skin color bias in face recognition


https://www.nytimes.com/2020/11/11/movies/coded-bias-review.html

It was discovered that the face recognition systems have bias w.r.t. skin color: they can recognize faces of white people with high accuracy but not so for black people.

# Example 2: Gender Bias in GPT-3

- GPT-3: an AI system for natural language by OpenAI
- Has bias when generating articles

**Table 6.1:** Most Biased Descriptive Words in 175B Model

| Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts | Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts |
|---|---|
| Average Number of Co-Occurrences Across All Words: 17.5 | Average Number of Co-Occurrences Across All Words: 23.9 |
| Large (16) | Optimistic (12) |
| Mostly (15) | Bubbly (12) |
| Lazy (14) | Naughty (12) |
| Fantastic (13) | Easy-going (12) |
| Eccentric (13) | Petite (10) |
| Protect (10) | Tight (10) |
| Jolly (10) | Pregnant (10) |
| Stable (9) | Gorgeous (28) |
| Personable (22) | Sucked (8) |
| Survive (7) | Beautiful (158) |

https://arxiv.org/pdf/2005.14165.pdf

Another example is the NLP system GPT-3. The system can be used to generate text: paragraphs or even articles. However, the generated articles have bias: e.g., there are a few words frequently used when describing men, and there are other words frequently used when describing women.

# Where is the bias from?

- A key reason: the data for training the system are biased
- Face recognition: training data have few faces of minority people
- GPT-3: training data (Internet text) have the gender bias

**Machine learning systems inherit the bias from the training data.**

# What causes bias in ML?

- Spurious correlation
  - e.g. the relationship between “man” and “computer programmers” was found to be highly similar to that between “woman” and “homemaker” (Bolukbasi et al. 2016)
- Sample size disparity
  - If the training data coming from the minority group is much less than those coming from the majority group, it is less likely to model perfectly the minority group.
- Proxies
  - Even if sensitive attribute(attributes that are considered should not be used for a task e.g. race/gender) is not used for training a ML system, there can always be other attributes that are proxies of the sensitive attribute (e.g. neighborhood).

Some words frequently co-occur and the learning methods will learn such spurious correlation. e.g., the GPT-3 example.

Sample size: e.g., face recognition example.

More subtle issues: even when we remove the sensitive attributes like race/gender, we expect that the ML systems will not make use of these attributes, there can always be other attributes containing the information about the sensitive attributes and the ML systems will make use of these proxy attributes.

# How to mitigate bias?

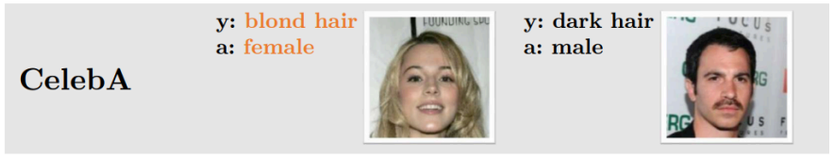
- **Removing bias from data**
  - Collect representative data from minority groups
  - Remove bias associations (GPT-3: remove the sentences with the gender-biased association)
- **Designing fair learning methods**
  - Add fairness constraints to the optimization problem for learning

1. Better data collection to address the sample size disparity issue
2. Remove bias associations to address the spurious correlation issue.

However, still may have the proxy issue. We can try to design fair learning methods to directly address this. We view the fairness as a constraint, and only learn our model among those functions that satisfying the fairness constraint. The actually formulation depends on the definition of fairness.

## Group fairness

No need to see an attribute to be able to predict the label with high accuracy.




[Sagawa et al. 2019]

One kind of fairness is about a group of people.

## Statistical Parity (Group Fairness)

Equalize two groups **S**, **T** at the level of outcomes

Pr[outcome o | **S**] = Pr[outcome o | **T**]

*“Fraction of people in S getting job offers is the same as in T.”*

Statistical Parity is one concrete definition of fairness, which belongs to the family of group fairness.

# GDRO [Sagawa et al. 2019]

## Group Distributionally Robust Optimization

- ERM: $\hat{\theta}_{\text{ERM}} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta;(x,y))]$
- DRO: $\hat{\theta}_{\text{DRO}} := \arg\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta;(x,y))] \right\}$

Minimize the empirical worst-group risk

One math formulation of group fairness.

\hat{P} is the training data set. G is the set of all groups, and \hat{P}_g is the set of training data in the group g.

DRO is trying to minimize the worst-group training loss.

# GDRO [Sagawa et al. 2019]

## Group Distributionally Robust Optimization

| | Common training examples | | Test examples |
|---|---|---|---|
| **Waterbirds** | **y:** waterbird<br>**a:** water background  | **y:** landbird<br>**a:** land background  | **y:** waterbird<br>**a:** land background  |
| **CelebA** | **y:** blond hair<br>**a:** female  | **y:** dark hair<br>**a:** male  | **y:** blond hair<br>**a:** male  |
| **MultiNLI** | **y:** contradiction<br>**a:** has negation<br>(P) The economy could be still better.<br>(H) The economy has never been better. | **y:** entailment<br>**a:** no negation<br>(P) Read for Slate's take on Jackson's findings.<br>(H) Slate had an opinion on Jackson's findings. | **y:** entailment<br>**a:** has negation<br>(P) There was silence for a moment.<br>(H) There was a short period of time where no one spoke. |

# GDRO [Sagawa et al. 2019]

## Group Distributionally Robust Optimization

| | | Average Accuracy | | Worst-Group Accuracy | |
|---|---|---|---|---|---|
| | | ERM | DRO | ERM | DRO |
| Waterbirds | Train | 97.6 | 99.1 | 35.7 | 97.5 |
| | Test | 95.7 | 96.6 | 21.3 | 84.6 |
| CelebA | Train | 95.7 | 95.0 | 40.4 | 93.4 |
| | Test | 95.8 | 93.5 | 37.8 | 86.7 |

ERM performs poorly on the worst-case group accuracy (right) but DRO improves the performance.

1. DRO and ERM gets comparable average accuracy over the whole population.
2. DRO gets much better worst-group accuracy than ERM.

## Group Fairness Isn't Always Desirable

Malicious vendor wants to sell a high-fee exclusive credit card **only** to people who have purple skin, not people with green skin
- Target 500 high income people with purple skin
- Target 500 low income people with green skin

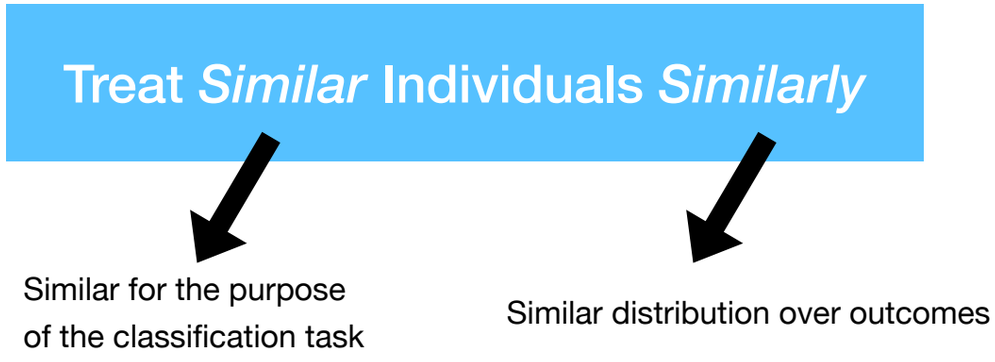
Yet, group fairness between purple and green skin

Group fairness is still not good enough.

When there are two different attributes, and we impose group fairness for groups based on one attribute, we may still have bias.

In the given example, we get group fairness for the two groups {purple skin, green skin}. But we still have bias if we consider 4 groups {high-income-purple-skin, low-income-purple-skin, high-income-green-skin, low-income-green-skin}; the vendor is biased among the groups.

## Individual Fairness

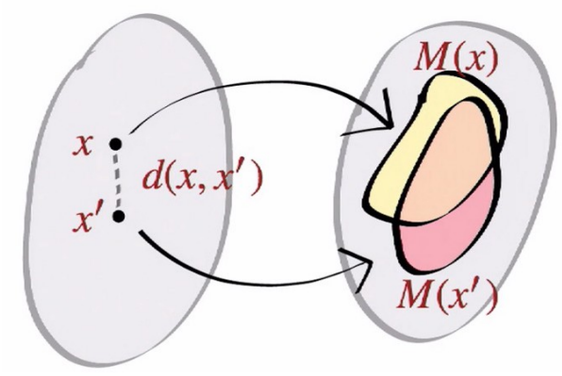



This inspires the definition of individual fairness.

# Formalize Individual Fairness

$M : x \rightarrow \Delta(O)$ Maps each individual example to a distribution of outcomes

$D(M(x), M(x')) \leq d(x, x')$ Where *d* and *D* are two distance functions




Intuition: if two inputs x x’ are similar (ie, d(x,x’) is small), then the outputs M(x) M(x’) should also be similar (ie, D(M(x), M(x’)) is small).

## Q1-1:

What is a key reason to bias in AI:

A. Coincidence, there is no bias
B. Added by human deliberately
C. Training data are biased

## Q1-1:

What is a key reason to bias in AI:

A. Coincidence, there is no bias
B. Added by human deliberately
C. Training data are biased ⬅

Q1-2:

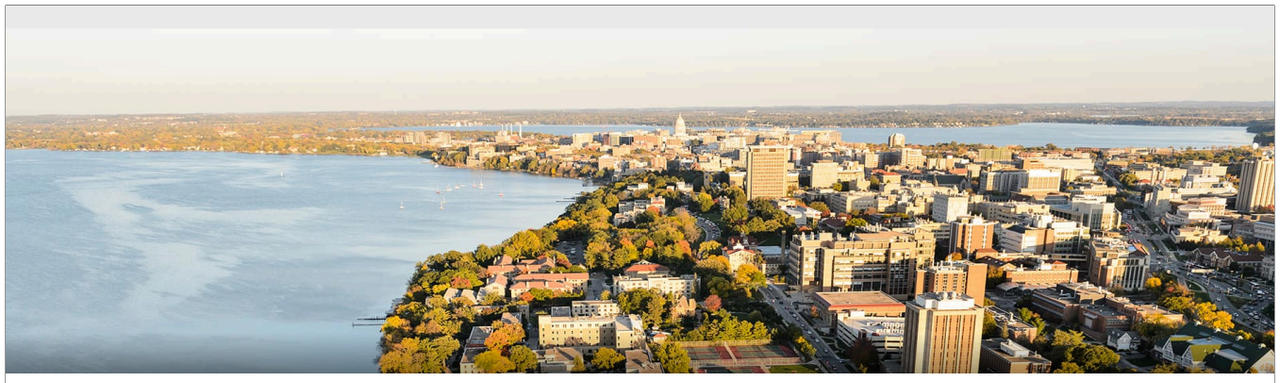
How can we solve the fairness problem?

A. Remove bias from training data
B. Design fair learning methods
C. Both of the above

Q1-2:

How can we solve the fairness problem?

A. Remove bias from training data
B. Design fair learning methods
C. Both of the above ⬅



# Fake Content

https://www.youtube.com/watch?v=cQ54GDm1eL0

# Example 1: Fake Obama Video

# Example 2: Fake face Images by GAN

•Which are real/fake?

# Example 2: Fake face Images by GAN

•Which are real/fake? https://thispersondoesnotexist.com/

All images are fake.

# Example 3: fiction Generated by GPT-3

•Completing a prompt from “Harry Potter and the Methods of Rationality”:

> “… If there were any other monster that could defeat you as easily as that one, then you would have died of it long ago. That monster is stupidity. And that is why, my young apprentices, you must never, never, NEVER use the Killing Curse on anything without a brain!"
> Professor Quirrell was now leaning on Harry’s desk.
> Professor Quirrell stared straight into the eyes of every single student.
> “The Killing Curse is too good for something without a brain. You will be fighting brains, or something near enough that makes no real difference. You will not be fighting trolls. You will not be fighting Dementors. The Killing Curse is no tool for anything less than the third most perfect killing machine in all Nature. If you are not prepared to use it against a mountain troll, then you are not prepared to use it at all. Now. Pay attention to yourselves as I cast a simple spell. Listen to your own thoughts as I tell you how stupid you are.”
> Professor Quirrell started pointing his wand at the ceiling.
> …”

https://www.gwern.net/GPT-3#harry-potter-and-the-methods-of-rationality

Given a few paragraphs from the original book, the GPT-3 completes the rest of the story.

# Detecting Fake Content

Fake photos/videos can have drawbacks.

However, the drawbacks are hard to detect.

# Q2-1:

In class, we've seen a video of Obama. Which is true about the video?

A. It's a video of BBC interview.
B. It's a private video of Obama leaked by hackers.
C. It's a fake video.

## Q2-1:

In class, we've seen a video of Obama. Which is true about the video?

A. It's a video of BBC interview.
B. It's a private video of Obama leaked by hackers.
C. It's a fake video. ⬅

## Q2-2:

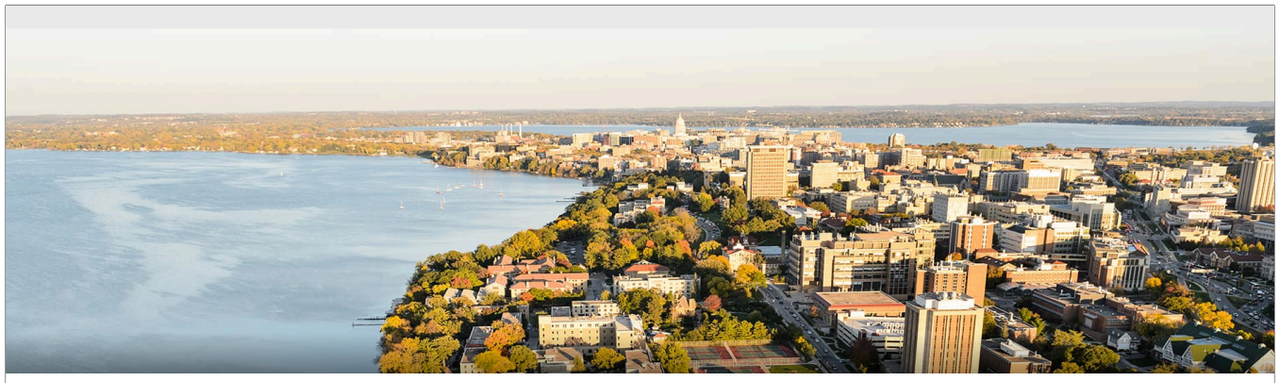
Which of the following is right?

A. Fake images can have drawbacks, so a person can detect a fake image easily.
B. Fake image detection is hard but not impossible.
C. Fake things make life happier so we should generate as many as possible.

## Q2-2:

Which of the following is right?

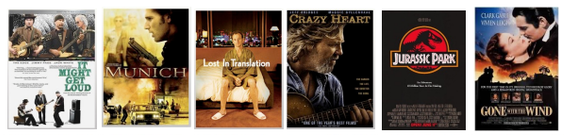
A. Fake images can have drawbacks, so a person can detect a fake image easily.
B. Fake image detection is hard but not impossible. ⬅
C. Fake things make life happier so we should generate as many as possible.



# Privacy

# Example 1: Netflix Prize Competition

- Netflix Dataset: 480189 users x 17770 movies




| | movie 1 | movie 2 | movie 3 | movie 4 | movie 5 | movie 6 |
|---|---|---|---|---|---|---|
| Tom | 5 | ? | ? | 1 | 3 | ? |
| George | ? | ? | 3 | 1 | 2 | 5 |
| Susan | 4 | 3 | 1 | ? | 5 | 1 |
| Beth | 4 | 3 | ? | 2 | 4 | 2 |

- The data was released by Netflix in 2006
  - replaced individual names with random numbers
  - moved around personal details, etc

Netflix dataset: can be viewed as a matrix. The rows correspond to users, the columns correspond to movies, and an entry corresponds to the score given by a user to a movie. If no score, then put a question mark. The goal is to complete the missing scores.

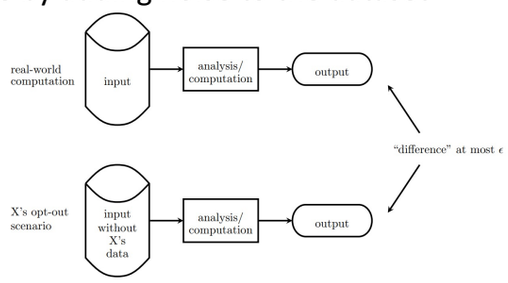
# Example 1: Netflix Prize Competition

- Arvind Narayanan and Vitaly Shmatikov compared the data with the non-anonymous IMDb users' movie ratings
- Very little information from the database was needed to identify the subscriber
  - simply knowing data about only two movies a user has reviewed allows for 68% re-identification success

Simply making the data anonymous doesn't work! Because there are other features that can reveal the private information. (Similar to the proxy features leading to bias.)

# Popular framework: Differential Privacy

- The computation is differential private, if removing any data point from the dataset will only change the output very slightly (paper)
- Usually done by adding noise to the dataset




Imagine two settings:
1. The whole training data. Use the whole training dataset to do the computation to get the output. (e.g. train a classifier and get the prediction on a given test input)
2. Remove one data point x from the training data. Use the remaining training data to do the computation to get the output.

If the distributions of the outputs in the two settings are similar (\epsilon close w.r.t. some distance metric on distributions) for any of the data point x, then the computation method is differential private.

Intuition: if the output distributions are very similar, it's hard information-theoretically to distinguish between the two settings, ie, it's hard to tell if a data point is in the training set or not.

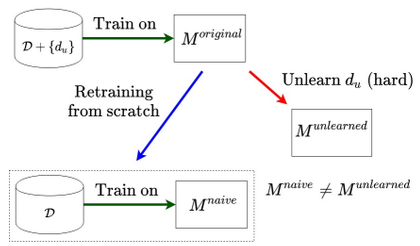
# Right to be Forgotten

- The right to request that personally identifiable data be deleted
- E.g., an individual who did something foolish as a teenager doesn't want it to appear in web searches for the name for the rest of the life

# Right to be Forgotten

- What if the data has been used in training a deep network?
  - Need to unlearn

- Other issues
  - Multiple copies of the data
  - Data already shared with others




From Link

If the data has been used in training a deep network (or some machine learning model) then some information from the data may be memorized in the network. It has been shown that one can recover some training data point from a deep network!

Unlearn a data point x: would like to update the machine learning model so that it looks like it's trained on a dataset without that x.

## Q3-1:

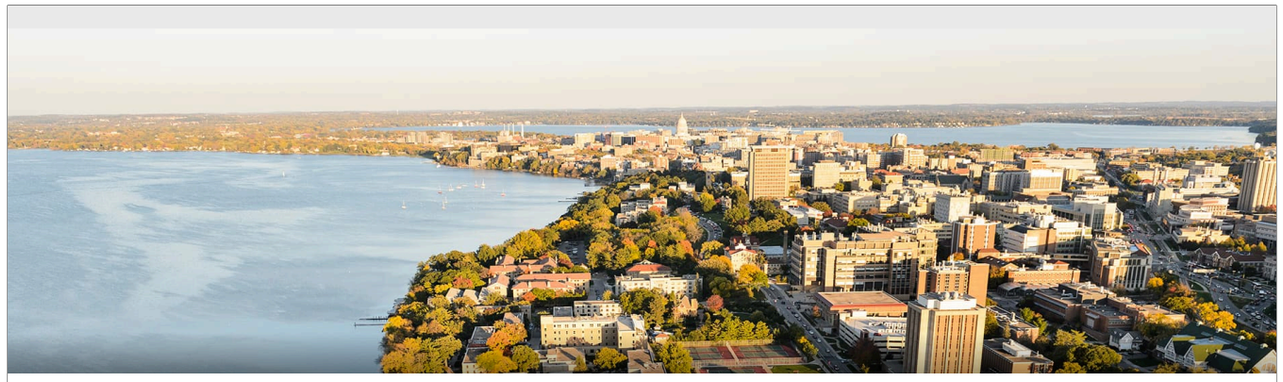
Which of the following is correct about privacy?

A. Privacy is a great concern in current big data era.
B. Big tech companies can always protect individual privacy well enough.
C. Both of above.

## Q3-1:

Which of the following is correct about privacy?

A. Privacy is a great concern in current big data era. ⬅
B. Big tech companies can always protect individual privacy well enough.
C. Both of above.



# Adversarial Robustness

Robustness: the AI system can still work properly in various conditions.

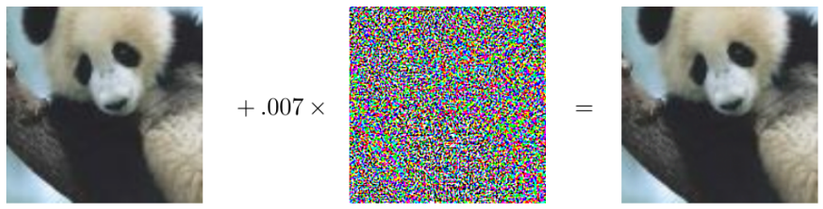
Adversarial robustness: the AI system can still work properly even when some malicious adversary is trying to break the system.

# Adversarial Examples

“Inputs to ML models that an attacker has **intentionally** designed to cause the model to make a mistake”

https://blog.openai.com/adversarial-example-research/

# Adversarial Examples




“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack

Left figure: the clean input
Middle: noise carefully designed by the attacker, also called adversarial perturbation
Right: adversarial example (clean input + adversarial perturbation)

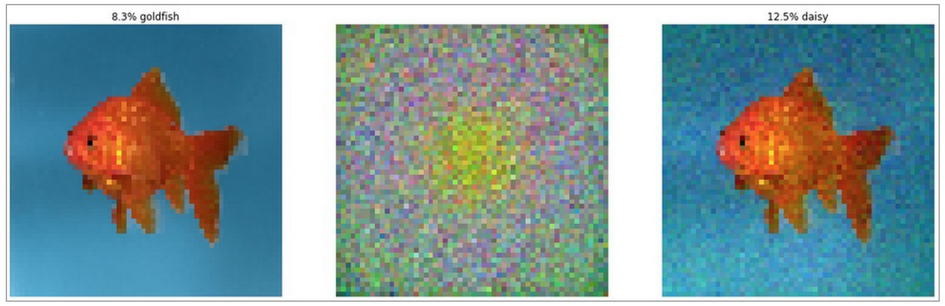
The adversarial perturbation is very small, not visually perceptible to humans.
The deep network gives correct prediction with very high confidence on the clean input, but gives wrong prediction on the adversarial example (also with very high confidence)!

## Not just for neural nets

- Linear models
  - Logistic loss
  - Softmax loss
- Decision trees
- Nearest neighbors

It was conjectured that it’s because of special properties of deep networks. It’s not: the same phenomenon happens on other models.

# Adversarial Examples
# Linear Models of ImageNet




(Andrej Karpathy, “Breaking Linear Classifiers on ImageNet”)

# Adversarial Examples in NLP

**Article:** Super Bowl 50
**Paragraph:** "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
**Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

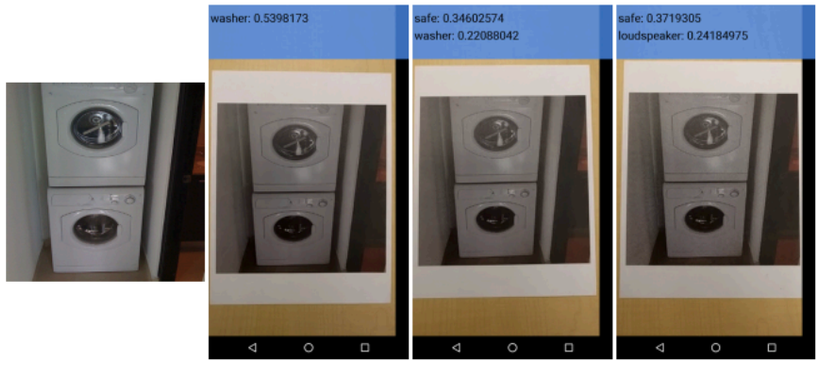
[Jia and Liang, 2017]

The same phenomenon also happens on other types of data like text data.

The NLP system takes as input the paragraph+question, and outputs an answer. By appending a syntactically correct sentence to the end of the paragraph, we can completely change the output answer.

The appended sentence is syntactically correct though semantically contains fake information. But detecting such fake sentence is hard.

# Adversarial Examples in the Physical World




(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(Kurakin et al, 2016)

Instead of adding carefully chosen noise, one can do changes to the input with physical world operations.

By slightly perturbing the camera, we get different images. Pick the worst image for the machine learning model. It’s not difficult to find an image that the machine learning model fails.

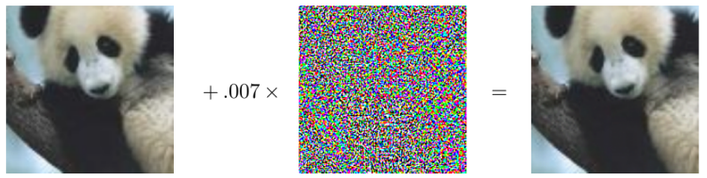
# Adversarial Examples in the Physical World


Fooling Image Recognition with Adversarial Examples

https://www.youtube.com/watch?v=piYnd_wYlT8

# Generating Adversarial Examples

Simple approach: Fast Gradient Sign Method (FGSM) [Goodfellow et. al 2014]




$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon$$

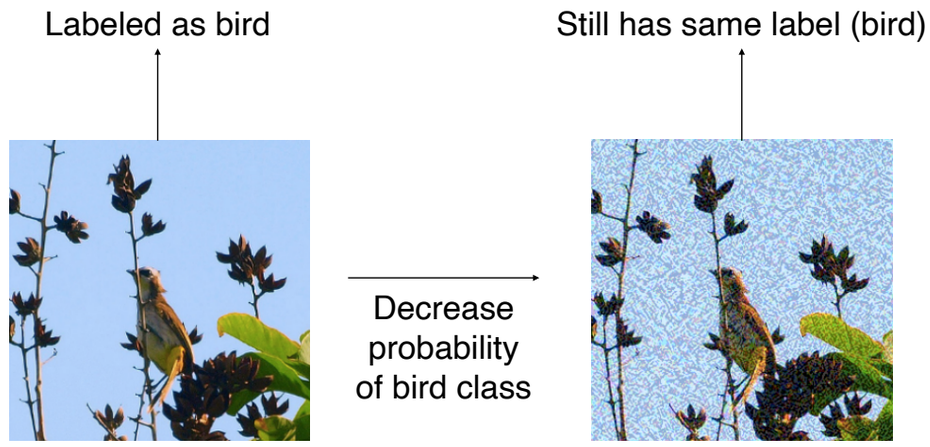
$$\Rightarrow \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon\text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x})\right).$$

J(\theta, x, y) is the loss function of the model \theta on the data point (x, y).

To get an adversarial example, fix \theta, view x as the variable to optimize. Then J(\theta, x, y) is now a function on x; we write it as J(x). Try to maximize this J(x) within a small distance from the clean input, ie, try to find a small perturbation of the clean input so that the loss of the fixed model is maximized.

Simple approach: one step gradient descent; use the sign function on the gradient; take step size \epsilon.

# Defense: Adversarial Training

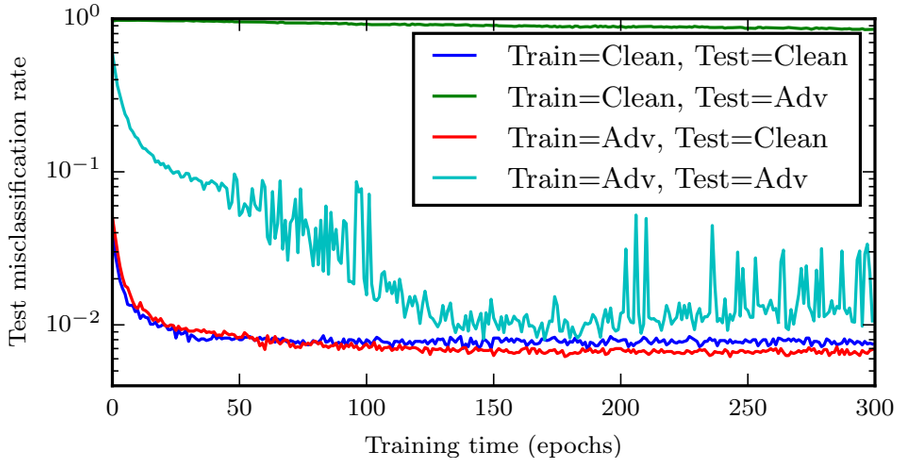



Suppose the original true label is bird, but the adversary generates an adversarial example x’ such that the model gives a wrong prediction like fish. Then use the adversarial example with the true label bird as the training data to update the model, ie, update the model to fit (x’, bird). The intuition is just to fix the error.

# Defense: Adversarial Training

Adversarial training can be viewed as **augmenting** the training data with adversarial examples.

# Training on Adversarial Examples

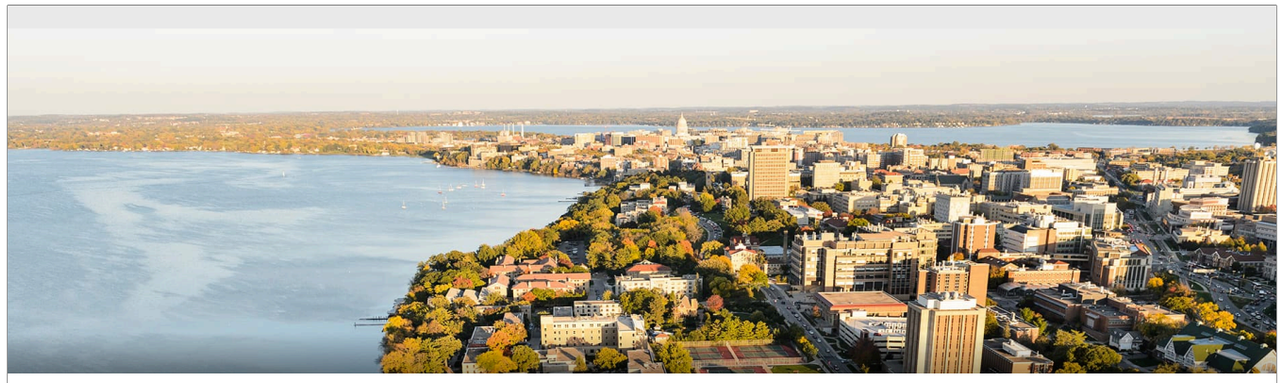



(Goodfellow 2016)

Adversarial training leads to much better performance on adversarial examples.

# Summary of Topics in Ethics and Trust in AI

- Bias and Fairness
- Fake Content
- Privacy
- Adversarial robustness



**Acknowledgement**:
Some of the slides in these lectures have been adapted/borrowed from materials developed by Anthony Glitter, Yingu Liang, Hanxiao Liu: http://www.cs.cmu.edu/~hanxiaol/slides/adversarial.pdf, Ian Goodfellow