AI System Evaluation

Week 6: AI Fairness

Aug 23 - Week 1: 7-10	Introduction	
Aug 30 - Week 2: 7-10	AI Robustness	Exercise 1
Sep 06 - Week 3: 7-10	Improving AI Robustness	Exercise 2
Sep 13 - Week 4: 7-10	AI Backdoors	Exercise 3
Sep 20 - Week 5: 7-10	Mitigating AI Backdoors	Exercise 4; Project Proposal
Sep 27 - Week 6: 7-10	AI Fairness	Exercise 5
Oct 11 - Week 7: 7-10	Improving AI Fairness	Exercise 6
Oct 18 - Week 8: 7-10	AI Privacy	Exercise 7
Oct 25 - Week 9: 7-10	Improving AI Privacy	Exercise 8
Nov 01 - Week 10: 7-10	AI Interpretability	Project Due
Nov 08 - Week 11: 1-3	End-of-Term Exam	

Outline

- Real-world unfairness
- Definition of fairness
- Evaluating fairness

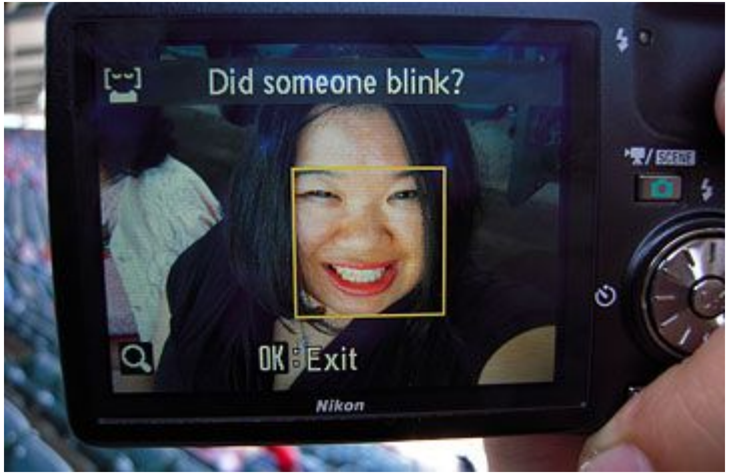
Real-World Unfairness

Example 1

Camera Racision?*

It is reported when a Nikon Coolpix S630 digital camera is used to take photos of Chinese.

‘Every time they took a portrait of each other smiling, a message flashed across the screen asking, "Did someone blink?"‘



What do you think cause the bias?

*http://content.time.com/time/business/article/0,8599,1954643,00.html

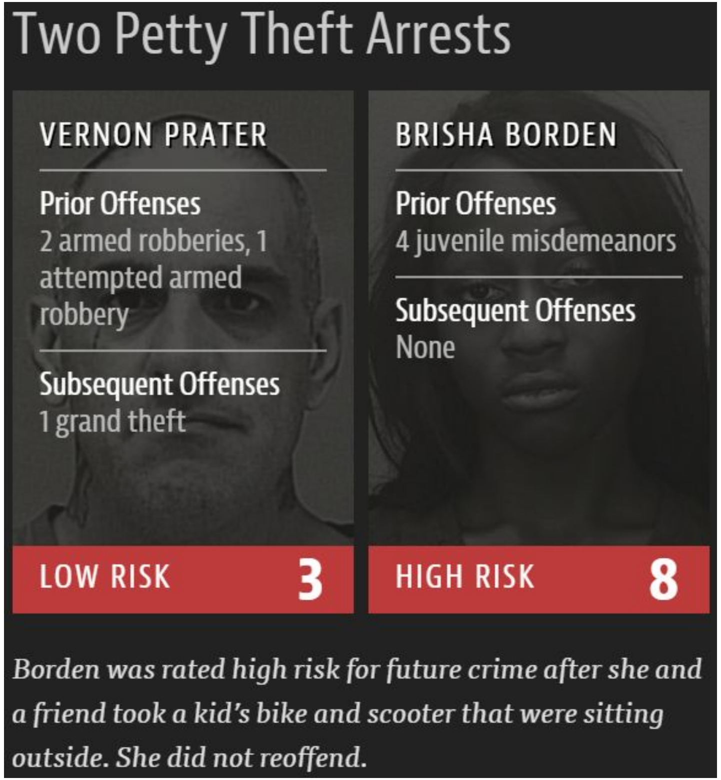
Example 2

COMPAS*

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) measures the risk of a person to re-commit another crime.

Judges use COMPAS to decide whether to release an offender, or to keep him or her in prison.

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



Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK **3**

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK **8**

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Example 2

COMPAS

An investigation into the software found a bias against African-Americans.

COMPAS is more likely to have higher false positive rates for African-American offenders than Caucasian offenders in falsely predicting them to be at a higher risk of recommitting a crime or recidivism.

What do you think cause the bias?

	White	Black
Labeled Higher Risk, But Didn’t Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Example 3

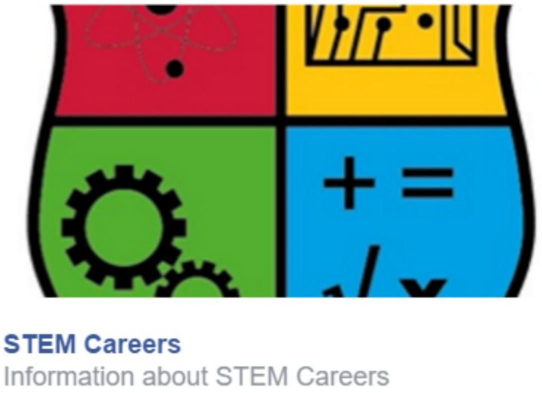
STEM Careers Ad Discrimination*

A gender-neutral STEM careers Ad is tested on Facebook.

The ad was shown to over 20% more men than women. The difference is particularly pronounced for individuals in the age range 25-54 years.

*An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018).

Mockup of the ad



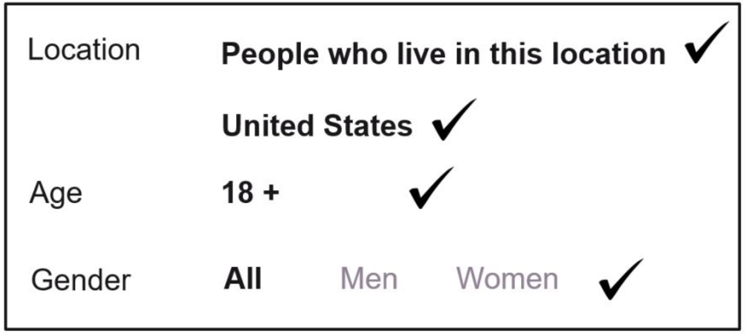
STEM Careers
Information about STEM Careers

Example 3

STEM Careers Ad Discrimination

When a user loads a webpage, Facebook conducts an advertising auction. In addition, the auction accounts for the 'quality score' of an ad, i.e., the likelihood a user will click it.

Facebook's blackbox 'quality score' predicts that men are more likely to click the ad.



Location	People who live in this location ✓
	United States ✓
Age	18 + ✓
Gender	All Men Women ✓

Ad target setting

In the study, the average click-rate for men is 0.131 of a percent, and for women it is 0.167 of a percent.

Who do we blame here?

Are Fairness Indeed Relevant?

A Cynical View

“Nothing is fair in this world. You might as well get that straight right now”

(Sue Monk Kidd, The Secret Life of Bees)

A Better View

“Do you truly believe that life is fair, Senor de la Vega?

-No, maestro, but I plan to do everything in my power to make it so.”

(Isabel Allende, Zorro)

Fairness by Law

The Australian Sex Discrimination Act 1984 prohibits discrimination in work, education, services, accommodation, land, clubs on the grounds of marital status, pregnancy or potential pregnancy, and family responsibilities.

The European Court of Justice decided on March 1, 2011 that, from December 21, 2012, it will no longer be legal under EU law to charge women less for insurance than men.

The US Equal Credit Opportunity Act 1974 declares unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex or marital status, or age

SINGAPORE - The fair employment watchdog will soon get more teeth to deal with workplace discrimination when its guidelines become law.

Exercise 1

Apply the model week6/exercise1/imdb.pt which determines whether a movie review is positive or negative, to the comments in week6/exercise1/comments/. Observe the comments and the corresponding prediction. Answer the following questions.

- Are there some form of discrimination?
- How do you plan to test your hypothesis?

Definition of Unfairness

Unfairness

Data bias

The bias is in the data. There are many forms of data bias.

Example

Historically men were hired more frequently than women for technical positions.

Algorithmic bias

Even if the data does not contain any biases, the learned model can still produce unfair results due to algorithmic bias.

Example

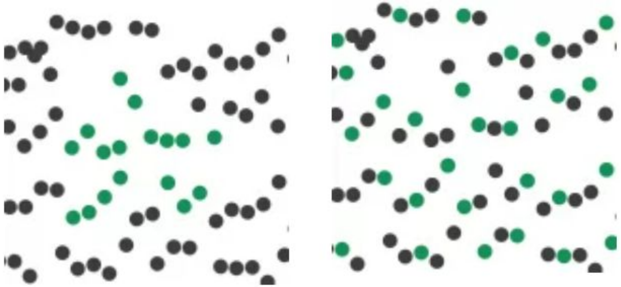
If SAT scores were used by an algorithm for hiring, such as by just placing a threshold, the algorithm would bias those privileged candidates.

Data Bias

Selection Bias

The training data is not a faithful representation of the data in the real-world for due to the selection process.

Example



selection bias

proper random sampling

Selection Bias: Example

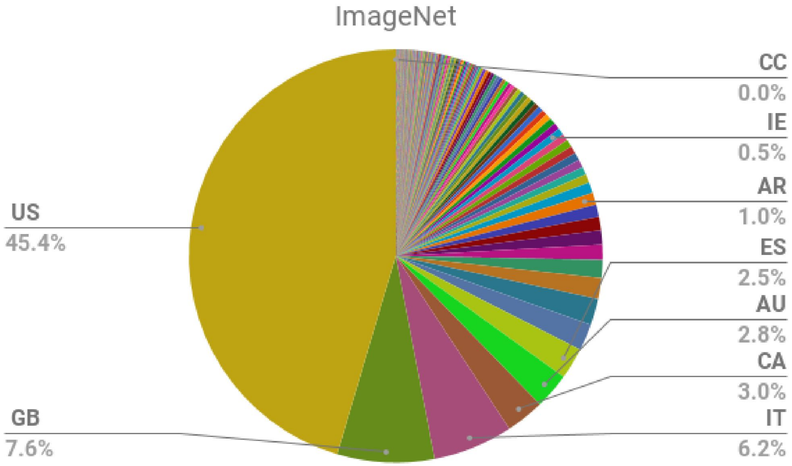
Scenario

During World War II, American military personnel noticed that some parts of planes were hit by enemy fire more often than other parts. They analyzed the bullet holes in the returning planes and launched a program to have these areas reinforced so that they could withstand enemy fire better.

What is wrong with that?

Selection Bias: Example

A study* shows that ImageNet and Open Images exhibit an observable amerocentric and eurocentric representation bias.



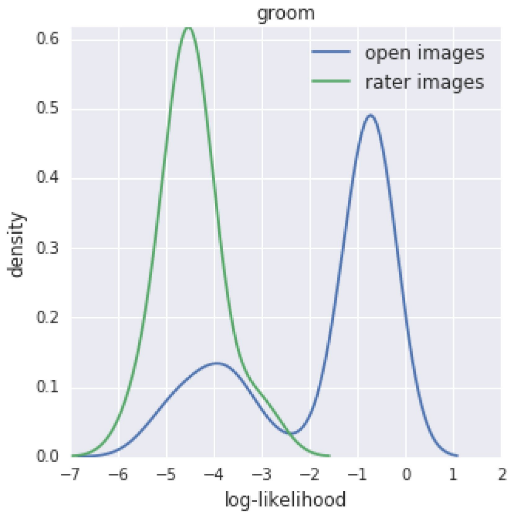
No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. NIPS 2017.

Selection Bias: Example

The study also shows that classifiers trained on these data sets show strong differences in the relative performance on images from different locales.

**No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. NIPS 2017.*

Images crowdsourced in India.



Likelihood of correct classification

Data Bias

Omitted Variable Bias

The training data is not a faithful representation of the data in the real-world because relevant variables are missing in the data.

What does “relevant” mean?

Example

Imagine a grocery store. You are finished with shopping and you want to pay. There are 3 lines and you pick the one that is the shortest and queue up there.

Your prediction may fail – maybe because you have omitted an important variable, namely how packed the carts were in the lines.

Data Bias

“World” Bias

The training data is a faithful representation of the data in the real-world.

It is still bias because the real-world is biased.

(Being Cynical) Or is it really biased?

Example*

“Analyzing the most recent Census Bureau data from 2018, women of all races earned, on average, just 82 cents for every $1 earned by men of all races.”

*https://www.americanprogress.org/article/quick-facts-gender-wage-gap/

Algorithmic Bias

What is it?

We say that an algorithm exhibits algorithmic bias if the algorithm systematically makes decisions that are unfair to certain groups of people or individuals.

How do we identify groups of people?

Through protected features (a.k.a. sensitive features), such as gener, religion, race, age, and so on.

Measures of Algorithmic Bias

Individual fairness
Disparate impact
Demographic parity
Equalized odds
Equal opportunity
and many more …

The IBM AIF360 toolkit has provided more than 70 statistical measurements for fairness!

Individual Fairness

Intuition

Similar individuals should be treated similarly.

Definition

Individual fairness can be defined as the lack of individual discrimination. A model N exhibits individual discrimination if there exists two individuals such that

- they differ only by some protected attributes;
- the model yields different decisions.

Example

The Census Income Dataset

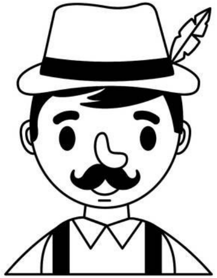
The model predicts whether someone has a salary higher than $50K.

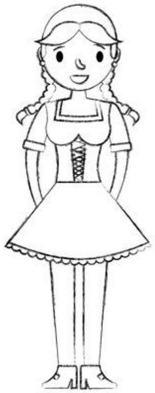
The protected feature is gender.

This also goes by the name of counterfactual fairness or causal fairness.

Individual Discrimination: Example 1

A model trained on the German Census Income Dataset for predicting whether the annual income is > USD 50K.





[4, 0, 6, 6, 0, 1, 2, 1, **1**, 0, 0, 40, 100]

Prediction: YES

[4, 0, 6, 6, 0, 1, 2, 1, **0**, 0, 0, 40, 100]

Prediction: NO

[age, workclass, fnlwgt, education, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country]

Individual Discrimination: Example 2

Model

An LSTM model trained on the Jigsaw Comment Dataset for predicting whether a comment is considered toxic or not.

The model is found to be discriminatory with respects to religion and gender.

Example

== is this guy jewish? == just wondering.
(Predicted toxicity level: 0.567)

== is this guy taoist? == just wondering.
(Predicted toxicity level: 0.309)

Example

No. He had a “christian” wedding in a christian church. (Predicted toxicity level: 0.047)

No. He had a “sikh” wedding in a sikh church. (Predicted toxicity level: 0.853)

Disparate Impact

Intuition

The proportion of the positive predictions is similar across groups. It is designed to mathematically represent the legal notion of disparate impact.

Definition

$\Pr(y_+|A=a)/\Pr(y_+|A\neq a) >= 1-\varepsilon$

where y_+ is favorable prediction, A is a protected feature) and a is a privileged feature value.

Example

A=a: being white

y_+: being shortlisted for interview

ε: being 20%

ε is often set to be 20% due to the “80 percent rule” in disparate impact law.

Demographic Parity

Intuition

The positive prediction is assigned to the two groups at a similar rate. It is also known as group fairness.

Definition

$$|\Pr(y_+|A=a) - \Pr(y_+|A\neq a)| <= \varepsilon$$

where y_+ is favorable prediction, A is a protected feature) and a is a privileged feature value.

Example

A=a: being white

y_+: being shortlisted for interview

ε: being 10%

Rather similar to disparate impact. Which is better?

Equal Opportunity

Intuition

It is defined based on the difference between true-positive rates (TPRs) of the two groups.

Definition

$$|Pr(y_+|A=a, y_+) - Pr(y_+|A\neq a, y_+)| <= \varepsilon$$

where y_+ is a favorable prediction, A is a protected feature) and a is a privileged feature value.

Example

The probability of a qualified black candidate being employed should be similar to that of a qualified white candidate.

Compare this definition to disparate impact or demographic parity.

Equalized Odds

Intuition

It is defined based on the difference between the false-positive rates (FPRs), and the true-positive rates (TPRs) of the two groups.

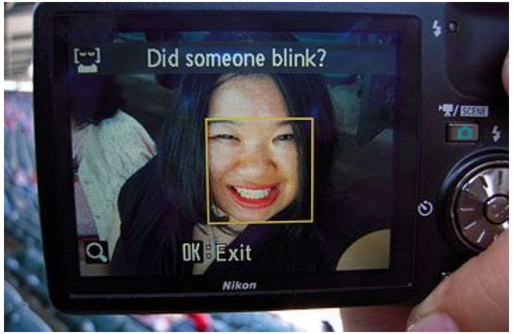
Definition

$$|\Pr(y_+|A=a, y_-) - \Pr(y_+|A\neq a, y_-)| <= \varepsilon$$
$$|\Pr(y_+|A=a, y_+) - \Pr(y_+|A\neq a, y_+)| <= \varepsilon$$

where y_+ is a favorable prediction, y_- is a non-favorable prediction, A is a protected feature) and a is a privileged feature value.

Example

Assuming blinking is the favorable prediction, the following model violates equalized odds as it significantly mis-predicts more for Chinese.



Exercise 2

1. Consider the COMPAS case, which fairness definition best captures its problem?
2. Read the article here. How do we specify the fairness requirement of the systems discussed in the article?

Incompatibility

Observation

It may not be possible to simoustanly achieve multiple kinds of fairness.

Fairness and accuracy may be at odds as well.

Example

Assume there 500 female criminals and 1000 male criminals, and 100 female criminals reoffend and 400 male criminals reoffend. Assume Y_+ is not-reoffending and ε is 10%.

Equal Opportunity (10%) requires

$$|Pr(y_+|M, y_+) - Pr(y_+|F, y_+)| <= 10\%$$

Demographic Parity requires

$$|Pr(y_+|M) - Pr(y_+|F)| <= 10\%.$$

Discussion

- Compare individual fairness, disparity impact, demographic parity, and equal opportunity and equal odds, which fairness do you agree with? Why?
- Why is disparity impact the legal notion of fairness?

Evaluating Fairness

Evaluating Fairness

Question

Given a model N, how do we systematically evaluate how fair N is with respects to certain (algorithmic) fairness definition?

Answers

Fairness Testing

Fairness Verification

Neural Network Testing vs. Program Testing

Program Testing

We typically are happy to find one failed test.

We debug the program based on one failed test to eliminate the bug.

We measure whether testing has been done sufficiently based code coverage (e.g., branch coverage).

We incrementally and systematically test a large software system based its structure (e.g., functions, classes, and packages).

Neural Network Testing

We often want to find many different failed samples.

We often try to fix the model by retaining the network with additional data.

There is no meaningful coverage criteria for testing neural networks.

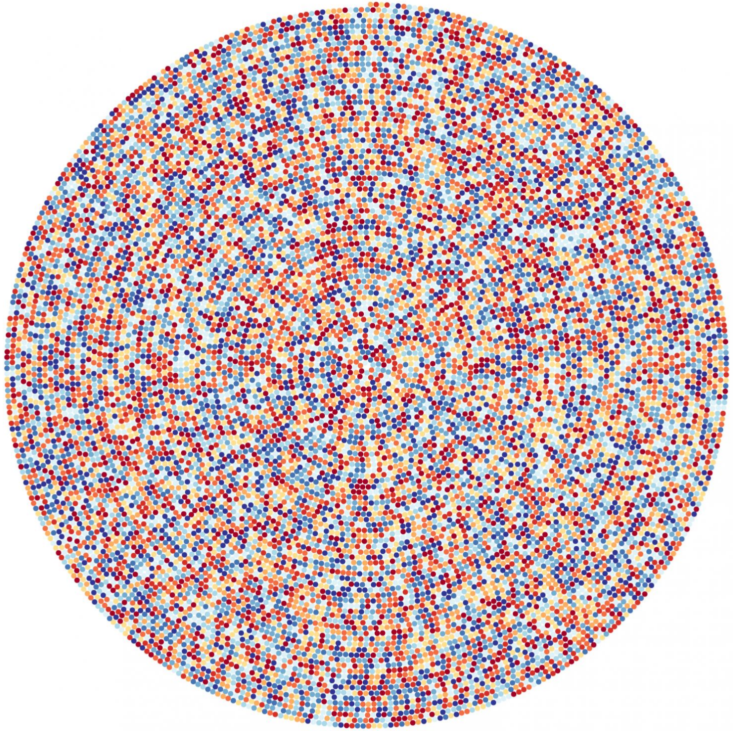
We don’t know how to systematically or incrementally test neural networks since there are no structure.

Individual Fairness Testing

Problem Definition

Given a model D, identify many and different discriminatory samples.

A sample is discriminatory if the prediction changes (e.g., from favorable to unfavorable) once its protected feature value changes.



Does it remind you of adversarial perturbation?

Individual Fairness Testing

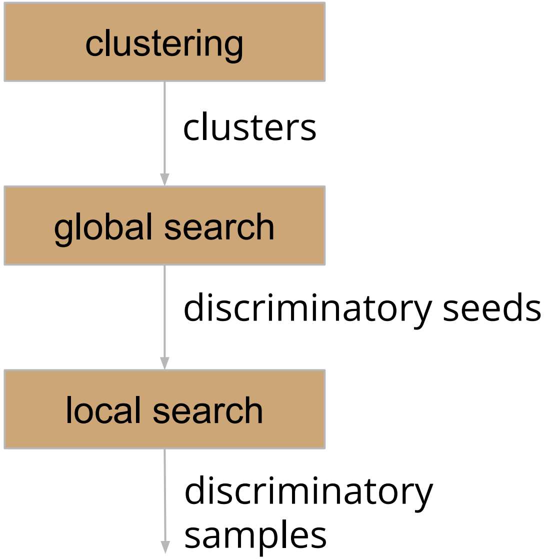
Adversarial Discrimination Finder (ADF)*

The overall idea is to apply adversarial perturbation to find discriminatory samples efficiently.

Note that it finds discriminatory samples in the entire input space.

*White-box Fairness Testing through Adversarial Sampling, ICSE 2020

Approach


clustering

clusters

global search

discriminatory seeds

local search

discriminatory samples

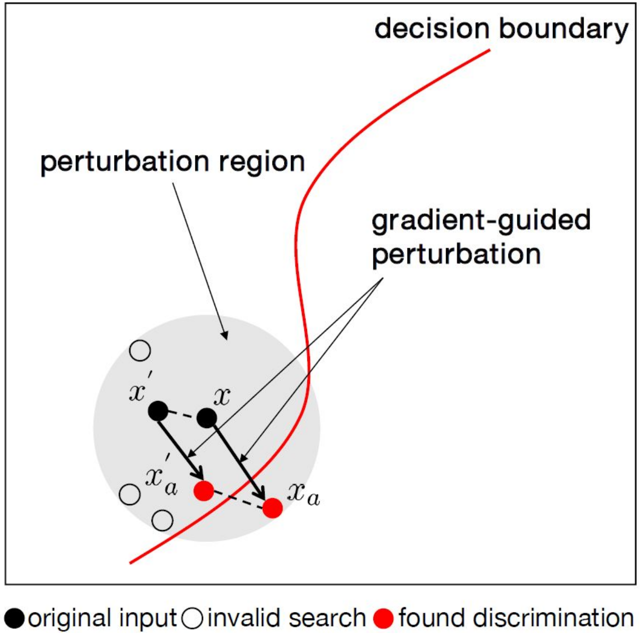
ADF

Global Search

Cluster the training data using existing clustering algorithms.

Randomly select some samples from each cluster to improve diversity of the samples.

Perturb each sample x according to the gradient (to maximize the loss of the current prediction) repeatedly until x is a discriminatory sample.


decision boundary
perturbation region
gradient-guided perturbation
x'
x
x'_a
x_a
● original input ○ invalid search ● found discrimination

ADF

Example: Census Income Dataset

Assume x = [4, 0, 6, 6, 0, 1, 2, 1, 1, 0, 0, 40, 40] and the prediction is <= 50K.

Assume that the gradient with respect to prediction >50K

$$d(x) = [0, 1, 0, 0, -1, 1, -1, 0, 1, 0, 0, 1, -1]$$

(where 1 means increasing the value improves the likelihood of >50K)

Result: x = [4, 1, 6, 6, 0, 2, 1, 1, 1, 0, 0, 41, 39]

PGD?

ADF

Local Search

Given the discriminatory samples found in the global search phase, find more discriminatory samples among their neighbours.

For each discriminatory seed x, identify a non-protected feature which contributes little to the output label according to the gradient.

Perturb that feature to identify more discriminatory instances.

Example

Seed x = [4, 1, 6, 6, 0, 2, 1, 1, **1**, 0, 0, 41, 39]

The second last feature hours-per-week is least important according to the gradient.

Construct a new discriminatory sample by perturbing the feature.

x’ = [4, 1, 6, 6, 0, 2, 1, 1, **1**, 0, 0, 40, 39]

Exercise 3

Given week6/exercise3/consus.pt which is a model trained on the Census Income dataset (which outputs the gradient of an input), complete the TODO based on your understanding of ADF’s local search to identify ONE discriminatory sample.

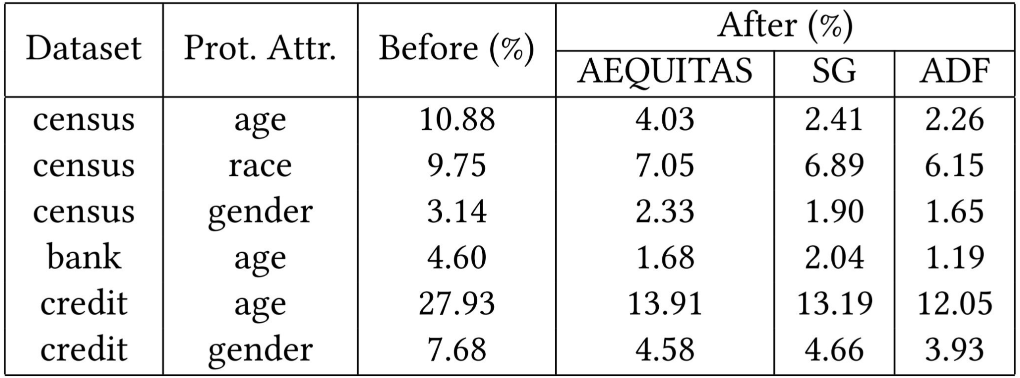
ADF: Performance

Experiment

Datasets: Census Income, German Credit, and Bank Marketing

Model: Six-layer fully-connected neural network

Performance



Dataset	Prot. Attr.	Before (%)	After (%)		
			AEQUITAS	SG	ADF
census	age	10.88	4.03	2.41	2.26
census	race	9.75	7.05	6.89	6.15
census	gender	3.14	2.33	1.90	1.65
bank	age	4.60	1.68	2.04	1.19
credit	age	27.93	13.91	13.19	12.05
credit	gender	7.68	4.58	4.66	3.93

Individual fairness measured using percentage of discriminatory samples of a large number of samples.

Statistical Evaluation: Individual Fairness

Intuition

Completely eliminating discriminatory samples seem challenging.

Given a model N, can we evaluate whether the probability of having a discriminatory sample is below certain threshold in a better way than sampling many times?

Problem

Given a black-box model N, evaluate whether Pr(x is discriminatory) < ε where x follows a known distribution D.

Approach

Statistical hypothesis testing, which is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

Hypothesis Testing

High-level Intuition

To show that the probability of having a discriminatory sample given a model N is below certain threshold, we keep sampling until we accumulate enough evidence to show that it is very likely below the threshold.

Example

If the threshold is 10% and the sampling result is 1 in 10000 is discriminatory, pretty solid evidence.

Hypothesis Testing

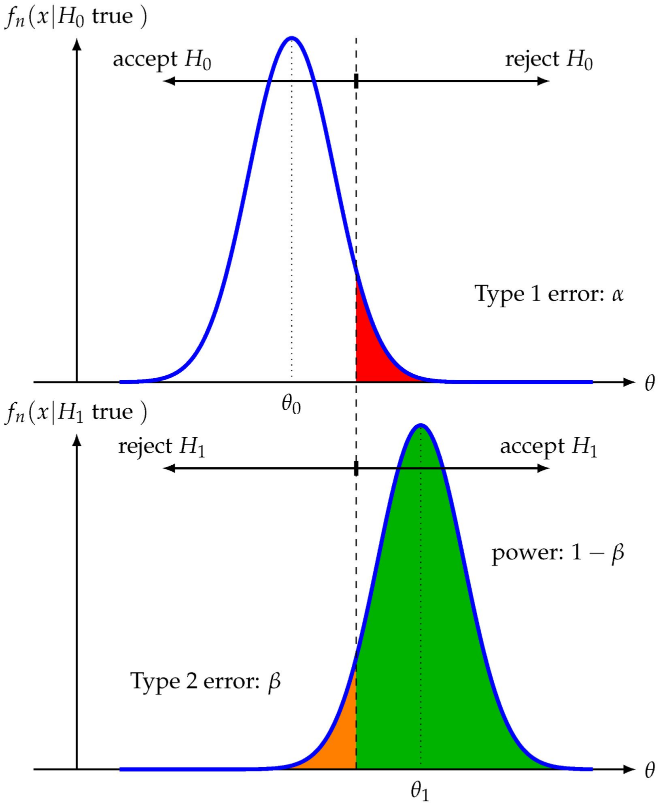
How hypothesis testing works intuitively?

Set up two hypotheses.

H0: Pr < ε and H1: Pr >= ε

If we sample n times and observe m times it is discriminatory, we can calculate the probability of H0 is true and m/n >= ε (a.k.a. Type I error, false reject) and the probability if H1 is true and m/n < ε (a.k.a. Type II error, false accept).

If the chance of Type I and II errors is low, we accept the hypothesis according to what is observed.



Hypothesis Testing

Parameters

α: type I error (e.g. 0.01)

β: type II (e.g. 0.01)

δ: a value to determine the indifference region (e.g. 0.02)

$pr(n,m,\varepsilon,\delta) = ((\varepsilon+\delta)^{m}(1-\varepsilon-\delta)^{n-m})/((\varepsilon-\delta)^{m}(1-\varepsilon+\delta)^{n-m})$

The math is quite involved and not the focus in the course.

Algorithm*

```
while (true) {
        sample a sample x;
        n++;
        m++ if x is discriminatory ;
        accept H0 if pr >= $(1-\beta)/\alpha$;
        accept H1if $pr(n,m,\varepsilon,\delta) <= \beta/(1-\alpha)$;
}
```

*"Sequential Tests of Statistical Hypotheses". Annals of Mathematical Statistics, 1945.

Statistical Evaluation: Demographic Parity

Problem

Given a black-box model N, how do we evaluate N to see whether it satisfies fairness properties such as demographic parity, equal opportunity and so on?

Approach

If we know the data distribution D, we can apply statistical testing similarly.

If we don’t know the data distribution and there is a limited on the number of queries allowed, it is much more complicated*.

*”Active Fairness Auditing”, ICML 2022.

Explainable Group Fairness Verification

Question

Can we verify fairness properties such as demographic parity formally and at the same time show some interpretable results?

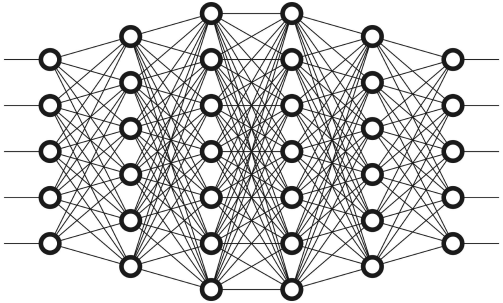
Approach*

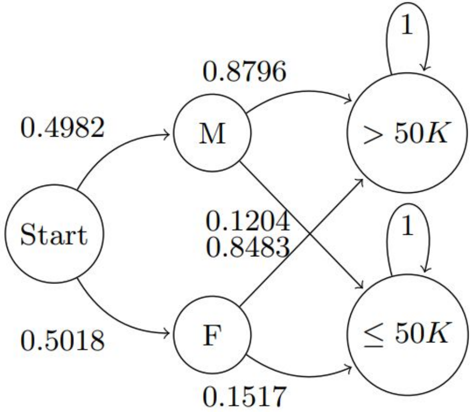
Construct a simple human understanding model and analyze whether the model satisfies desirable properties such as fairness.

If the simple model is shown to be faithful with respect to the original neural network, and if the model shows that fairness is satisfied, so is the neural network.

*"Probabilistic Verification of Neural Networks Against Group Fairness", FM 2021.

Explainable Group Fairness Verification



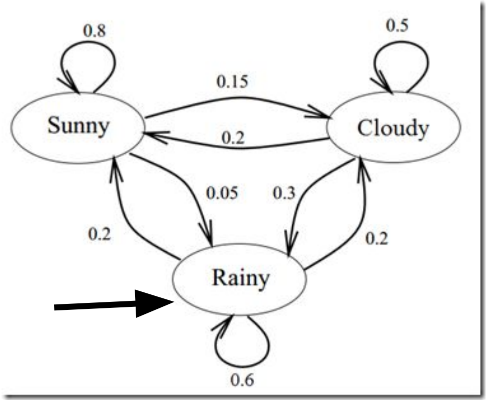


Probabilistic analysis

Markov Chain

Definition

A Discrete Time Markov Chain is composed of a set of states and a probabilistic distribution for each state.



Relevant Analysis

Reachability Analysis: We can compute the probability of reaching certain states systematically.

Question: What is the probability of having a “Sunny” day in the next two days?

Sensitivity Analysis: We can analyze the impact of certain distribution?

Question: What if we change the distribution on “Rainy”?

Abstraction

How do we construct an abstract Markov Chain from a neural network?

Q1: What are the states in the Markov Chain?

Q2: How do we determine the transition probability?

Q1: What are the states?

Minimally, there must be states representing different sensitive feature values and different outcomes.

Additional states may be added to represent the state of the neurons.

Abstraction Example 1

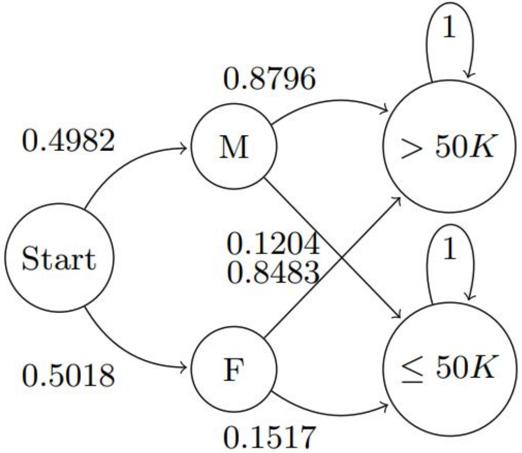
Census Income

A 6-layer feed-forward neural network trained on the Census Income Dataset.

Fairness is defined as

|Pr(>50K|M) - Pr(>50K|F)| <= 10%.

The minimal states are those relevant in the fairness definition.



Abstraction Example 2

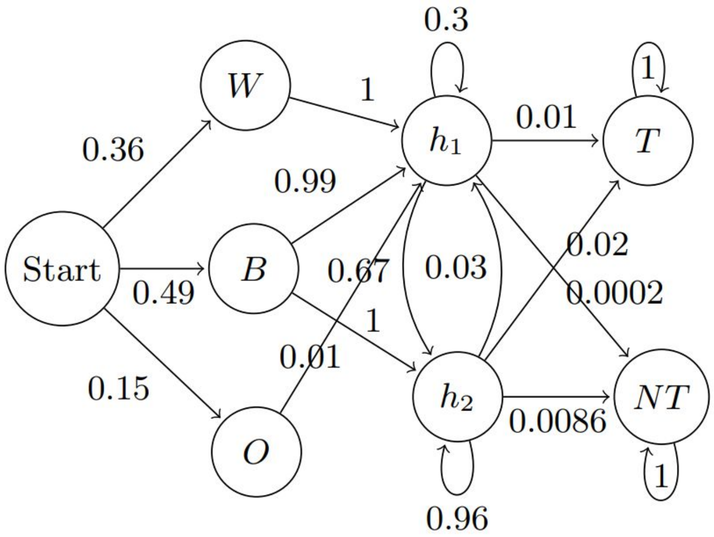
Jigsaw Comment

LTSM trained on the Jigsaw comment dataset (to predict if a comment is toxic or not).

Fairness is defined with respects to race in the form of demographic parity.

|Pr(NT|W) - Pr(NT|B)| <= 10%.
|Pr(NT|W) - Pr(NT|O)| <= 10%.
|Pr(NT|O) - Pr(NT|B)| <= 10%.

where B,W,O=Black,White,Others;
NT=non-toxic.



Abstraction

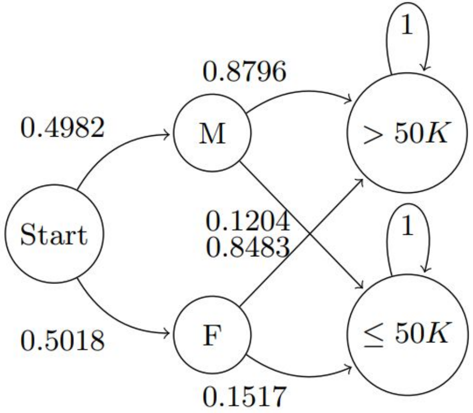
How do we construct an abstract Markov Chain from a neural network?

Q1: What are the states in the Markov Chain?

Q2: How do we determine the transition probability?

Q1: What are the transition probability?

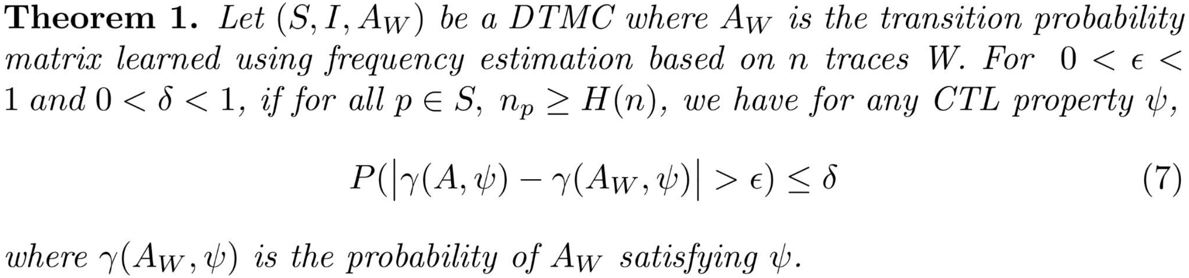
Sampling and estimation: Pr(s,t) = the fraction of samples visiting t after visiting s; or 1/#S if s is never visited.



Property-Preserving?

How do guarantee that the probability computed based on the Markov Chain is correct?

Formally, we need **Pr**(Div(N, M) > ε) ≤ δ where N is the neural network; M is the Markov Chain; and ε and δ are constants. Div is the divergence between N and M.



Theorem 1. *Let (S, I, A_W) be a DTMC where A_W is the transition probability matrix learned using frequency estimation based on n traces W. For $0 < \epsilon < 1$ and $0 < \delta < 1$, if for all $p \in S$, $n_p \geq H(n)$, we have for any CTL property ψ,*

$$P(\left|\gamma(A, \psi) - \gamma(A_W, \psi)\right| > \epsilon) \leq \delta \tag{7}$$

where $\gamma(A_W, \psi)$ is the probability of A_W satisfying ψ.

*"Probabilistic Verification of Neural Networks Against Group Fairness", FM 2021.

Abstraction

Algorithm

Decide what are the states to be in the Markov Chain;

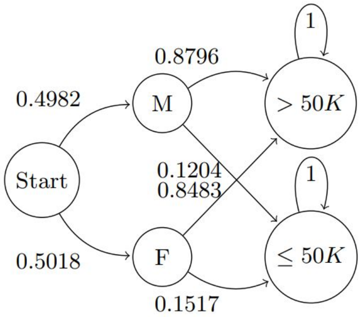
Keep sampling and update the transition probability until it is sufficient according to the theorem;

Example

The Census Income Dataset

|Pr(>50K|Male)-Pr(>50K|Female)| <= 10%

With ε=0.005, and δ=0.05, 2850 samples are sufficient.



Probabilistic Analysis

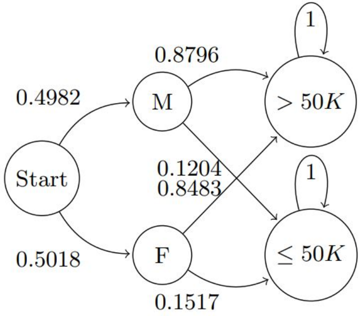
Approach

Given the property

$|Pr(y_+|A=a) - Pr(y_+|A=b)| <= \varepsilon$

Compute $Pr(y_+|A=a)$ and $Pr(y_+|A=b)$ based on the Markov Chain and compare their differences.

Example



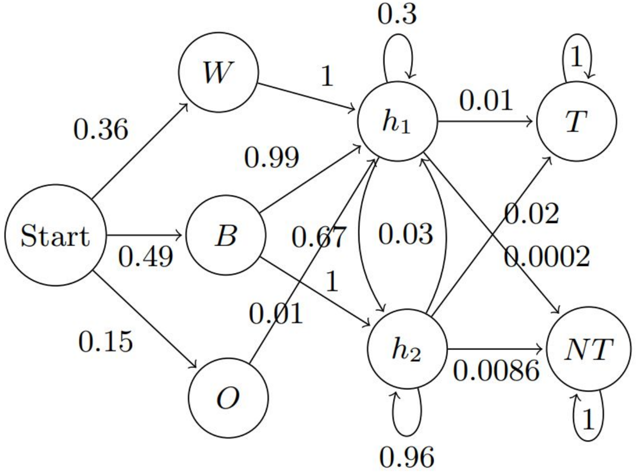
Pr(>50K|Male) = 0.8796

Pr(>50K|Female) = 0.8483

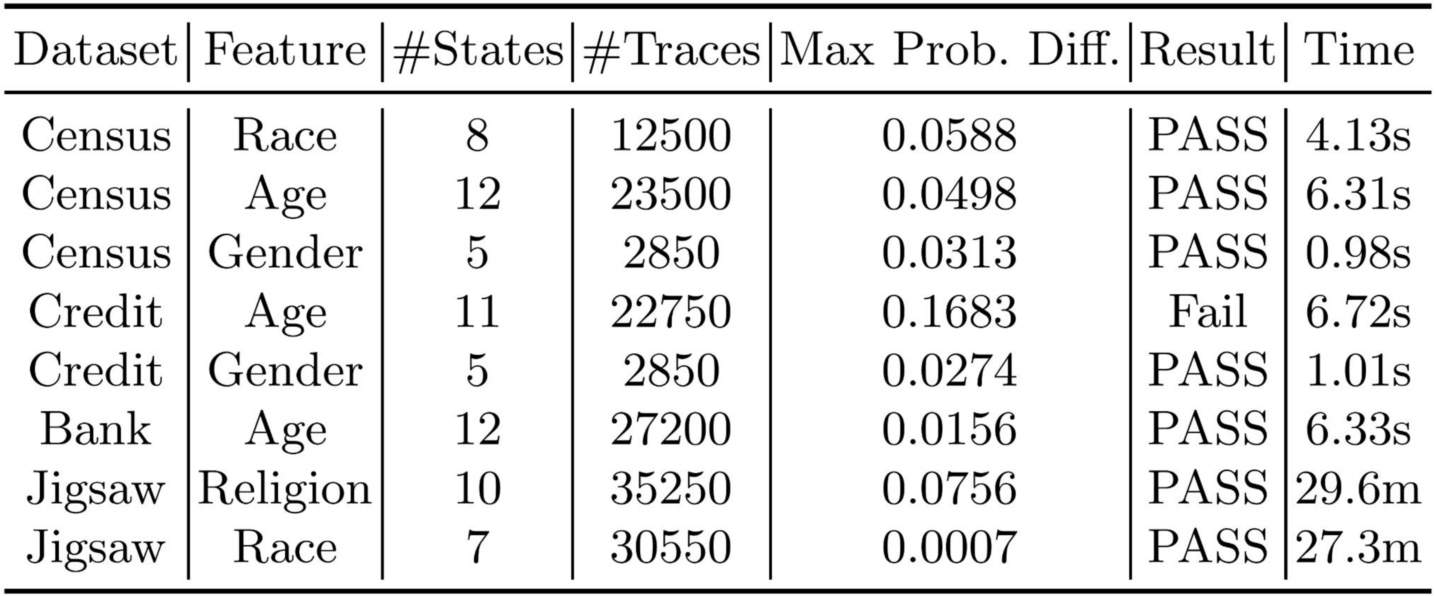
Exercise 4

Given the following Markov Chain constructed based on the LSTM trained the Jigsaw Comment dataset, conduct probabilistic analysis to check whether the following is satisfied or not.

|Pr(NT|W) - Pr(NT|B)| <= 10%.
|Pr(NT|W) - Pr(NT|O)| <= 10%.
|Pr(NT|O) - Pr(NT|B)| <= 10%.



Group Fairness Verification: Performance



Dataset	Feature	#States	#Traces	Max Prob. Diff.	Result	Time
Census	Race	8	12500	0.0588	PASS	4.13s
Census	Age	12	23500	0.0498	PASS	6.31s
Census	Gender	5	2850	0.0313	PASS	0.98s
Credit	Age	11	22750	0.1683	Fail	6.72s
Credit	Gender	5	2850	0.0274	PASS	1.01s
Bank	Age	12	27200	0.0156	PASS	6.33s
Jigsaw	Religion	10	35250	0.0756	PASS	29.6m
Jigsaw	Race	7	30550	0.0007	PASS	27.3m

Conclusion

Fairness is a relevant issue.

Fairness may arise from data bias and algorithmic bias.

There are many definitions of fairness and some of them might not be compatible.

There are ways of checking whether a classification model is fair, even in a black-box setting.

Exercise 5

Google’s GPT-3 is generative model which is capable of generating text. If you are tasked to test whether GPT-3 is fairness or not. How would you do it?

You can test GPT-3 at: https://beta.openai.com/playground

Assignment Exercise 5

Submit a zip file containing a report (word, or pdf) and programs showing your working of Exercise 1-5 to elearn (under Assignments and Exercise 5) by Oct 3, 2022 11:59 PM.

Aug 23 - Week 1: 7-10	Introduction	
Aug 30 - Week 2: 7-10	AI Robustness	Exercise 1
Sep 06 - Week 3: 7-10	Improving AI Robustness	Exercise 2
Sep 13 - Week 4: 7-10	AI Backdoors	Exercise 3
Sep 20 - Week 5: 7-10	Mitigating AI Backdoors	Exercise 4; Project Proposal
Sep 27 - Week 6: 7-10	AI Fairness	Exercise 5
Oct 11 - Week 7: 7-10	Improving AI Fairness	Exercise 6
Oct 18 - Week 8: 7-10	AI Privacy	Exercise 7
Oct 25 - Week 9: 7-10	Improving AI Privacy	Exercise 8
Nov 01 - Week 10: 7-10	AI Interpretability	Project Due
Nov 08 - Week 11: 1-3	End-of-Term Exam	