Natural Language Processing (CS-472)
Spring-2023

Muhammad Naseer Bajwa

Assistant Professor,
Department of Computing, SEECS
Co-Principal Investigator,
Deep Learning Lab, NCAI
NUST, Islamabad
naseer.bajwa@seecs.edu.pk

Overview of this week's lecture

Bias in AI

- Prototype theory
- Types of biases
- Some examples of biased studies/outcomes
- Fairness evaluation
- Using ML to address biases

What do you see?

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …

NUST
Defining futures
School of Electrical Engineering
& Computer Science

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …
- Green/Unripe Bananas

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …
- Green/Unripe Bananas

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …
- Green/Unripe Bananas
- Overripe Bananas

What do you see?

- Bananas
- Stickers
- Shelves
- Dole
- Grocery Market
- …
- Green/Unripe Bananas
- Overripe Bananas
- What about Yellow Bananas?

Think of a bird

Think of a bird

Think of a bird

Think of a bird

Think of a bird




Think of a bird

- What characterises a bird?

Think of a bird

- What characterises a bird?
  - Ability to fly?
  - Laying eggs?
  - Having beaks or bills?
  - Hollow bones?
  - Feathers?

Think of a bird

- What characterises a bird?
  - Ability to fly?
  - Laying eggs?
  - Having beaks or bills?
  - Hollow bones?
  - Feathers?
- What is typical for a bird?
  - All of above (more or less)

Think of a bird

- What characterises a bird?
  - Ability to fly?
  - Laying eggs?
  - Having beaks or bills?
  - Hollow bones?
  - Feathers?
- What is typical for a bird?
  - All of above (more or less)
  - A prototype has many typical characteristics

Prototype theory was formulated by Prof Rosch in 1970s

- Objects have a graded degree of belonging to a category; some members in a category are more central (prototypical) than other.

Prof. Eleanor Rosch (1938 – today)

Prototype theory was formulated by Prof Rosch in 1970s

- Objects have a graded degree of belonging to a category; some members in a category are more central (prototypical) than other.

- We usually notice and describe things that are atypical and ignore most of things that are typical.

Prof. Eleanor Rosch (1938 – today)

Prototype theory was formulated by Prof Rosch in 1970s

- Objects have a graded degree of belonging to a category; some members in a category are more central (prototypical) than other.

- We usually notice and describe things that are atypical and ignore most of things that are typical.

Riddle me this.

A man and his son are in a terrible accident and are rushed to the hospital in critical care.
The doctor looks at the boy and exclaims, “I can’t operate on this boy. He is my son!”

How is it possible?

Prof. Eleanor Rosch (1938 – today)

Prototype theory was formulated by Prof Rosch in 1970s

- Objects have a graded degree of belonging to a category; some members in a category are more central (prototypical) than other.

- We usually notice and describe things that are atypical and ignore most of things that are typical.

Riddle me this.

A man and his son are in a terrible accident and are rushed to the hospital in critical care.
The doctor looks at the boy and exclaims, “I can’t operate on this boy. He is my son!”

How is it possible?

Prof. Eleanor Rosch (1938 – today)

- A study conducted at Boston University on this riddle found that most of participants (men and women both) ignored the possibility of a female doctor.

https://www.bu.edu/articles/2014/bu-research-riddle-reveals-the-depth-of-gender-bias

What if an alien learnt about our world from the written text?

What if an alien learnt about our world from the written text?

- If we tried training a general AI to learn about the world from the text, what data will we provide it?
    - What is recorded in that data about our world?

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* ⟨*x*⟩, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Gordon, Jonathan, and Benjamin Van Durme. "Reporting bias and knowledge acquisition." Proceedings of the 2013 workshop on Automated knowledge base construction. 2013.

What if an alien learnt about our world from the written text?

- If we tried training a general AI to learn about the world from the text, what data will we provide it?
    - What is recorded in that data about our world?
- AI will probably learn a very skewed picture of our world.

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* $\langle x \rangle$, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Gordon, Jonathan, and Benjamin Van Durme. "Reporting bias and knowledge acquisition." Proceedings of the 2013 workshop on Automated knowledge base construction. 2013.

What if an alien learnt about our world from the written text?

- If we tried training a general AI to learn about the world from the text, what data will we provide it?
  - What is recorded in that data about our world?
- AI will probably learn a very skewed picture of our world.
- Data that have undergone such filtering process has Human Reporting Bias.
  - The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* $\langle x \rangle$, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Gordon, Jonathan, and Benjamin Van Durme. "Reporting bias and knowledge acquisition." Proceedings of the 2013 workshop on Automated knowledge base construction. 2013.

What if an alien learnt about our world from the written text?

- If we tried training a general AI to learn about the world from the text, what data will we provide it?
    - What is recorded in that data about our world?
- AI will probably learn a very skewed picture of our world.
- Data that have undergone such filtering process has Human Reporting Bias.
    - The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
- Reporting bias can affect AI in more ways than one.

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* ⟨*x*⟩, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Gordon, Jonathan, and Benjamin Van Durme. "Reporting bias and knowledge acquisition." Proceedings of the 2013 workshop on Automated knowledge base construction. 2013.

Human biases can creep in the data at various stages

- Biases start to affect the data during collection.
    - **Reporting Bias:** What people share is not a reflection of real-world frequencies.

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* ⟨*x*⟩, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Murdering is more common in the world than blinking

https://developers.google.com/machine-learning/glossary/

Human biases can creep in the data at various stages

- Biases start to affect the data during collection.
  - **Reporting Bias:** What people share is not a reflection of real-world frequencies.
  - **Selection Bias:** Selection does not reflect a random sample.

All fruits have pits

https://developers.google.com/machine-learning/glossary/
https://www.youtube.com/watch?v=uzGQvamQ_2M

Human biases can creep in the data at various stages

- Biases start to affect the data during collection.
  - **Reporting Bias:** What people share is not a reflection of real-world frequencies.
  - **Selection Bias:** Selection does not reflect a random sample.
  - **Out-group homogeneity bias:** outgroup members look alike with respect to attitude, values, personality etc.

All Chinese people look the same (to non-Chinese)

https://developers.google.com/machine-learning/glossary/
https://www.youtube.com/watch?v=uzGQvamQ_2M

Human biases can creep in the data at various stages

- Biases start to affect the data during collection.

  - **Reporting Bias:** What people share is not a reflection of real-world frequencies.

  - **Selection Bias:** Selection does not reflect a random sample.

  - **Out-group homogeneity bias:** outgroup members look alike with respect to attitude, values, personality etc.

  - **Stereotypical Bias, Prejudice, Halo Effect, Sampling error** and many more.

https://developers.google.com/machine-learning/glossary/
https://www.youtube.com/watch?v=uzGQvamQ_2M

- **Subjective validation:** Every annotator uses his/her knowledge/understanding to annotate.

Human biases in data annotation

- **Subjective validation:** Every annotator uses his/her knowledge/understanding to annotate.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

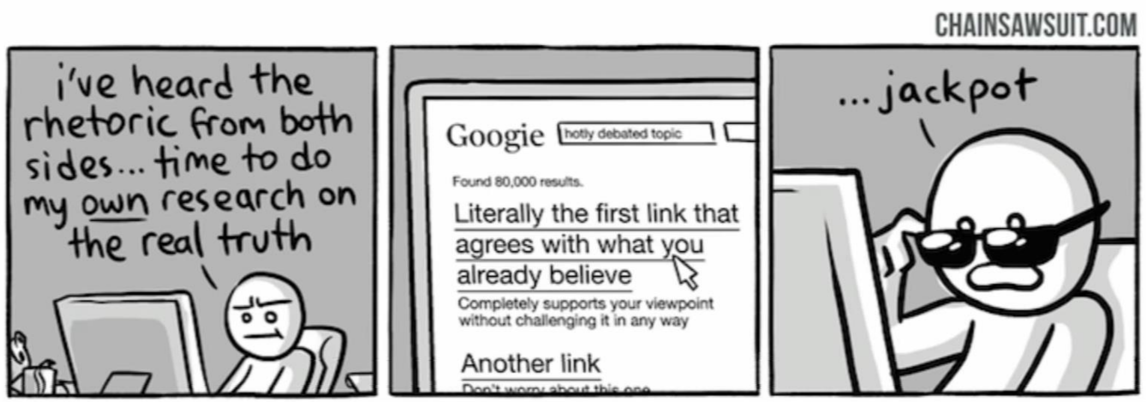
Human biases in data annotation

- **Subjective validation:** Every annotator uses his/her knowledge/understanding to annotate.

- **Confirmation bias:** Search for/interpret something that confirm one's pre-existing beliefs.

Human biases in data interpretation

- **Confirmation bias:** Search for/interpret something that confirm one's pre-existing beliefs.



Human biases in data interpretation

- **Confirmation bias:** Search for/interpret something that confirm one's pre-existing beliefs.

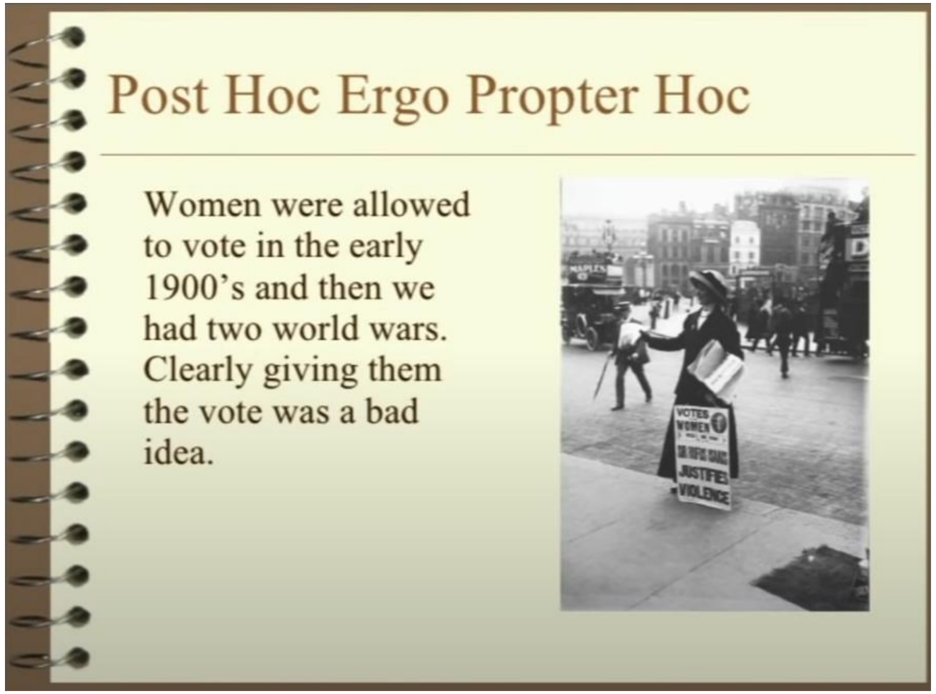
- **Overgeneralisation:** Coming to conclusion based on information that is too general and/or not specific enough.

Human biases in data interpretation

- **Confirmation bias:** Search for/interpret something that confirm one's pre-existing beliefs.

- **Overgeneralisation:** Coming to conclusion based on information that is too general and/or not specific enough.

- **Logical fallacies:** Correlation, anecdotal, etc.



Covid vaccine caused my uncle's father-in-law's niece's sister-in-law's grandson's neighbour hear attack. That's why Covid vaccines are not safe.

Human biases in data interpretation

- **Confirmation bias:** Search for/interpret something that confirm one's pre-existing beliefs.

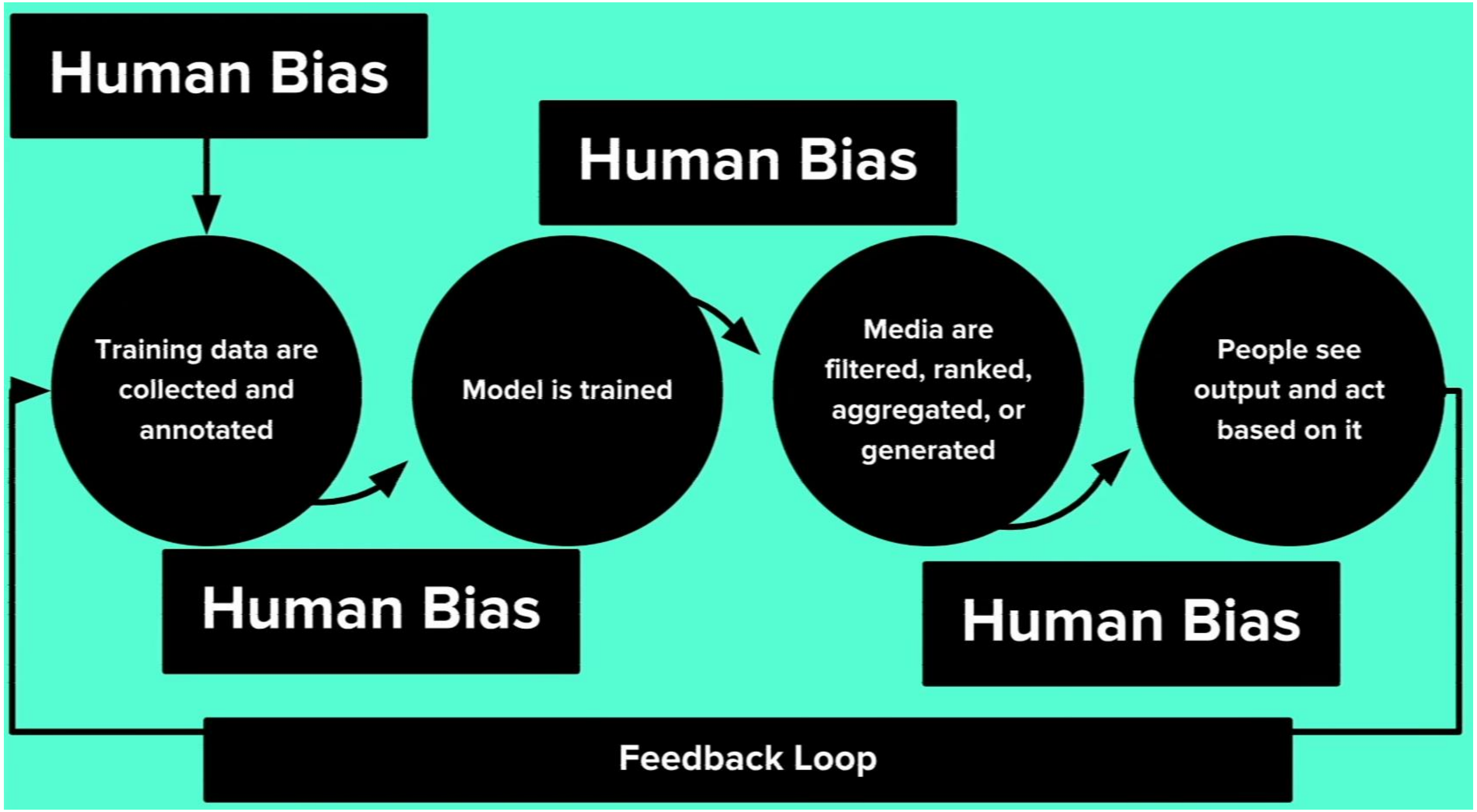
- **Overgeneralisation:** Coming to conclusion based on information that is too general and/or not specific enough.

- **Logical fallacies:** Correlation, anecdotal, etc.

- **Automation bias:** Propensity for humans to favour suggestions made by automated decision-support systems.

Biases, Biases Everywhere

Human biases are everywhere



Biases can be good, bad, or neutral

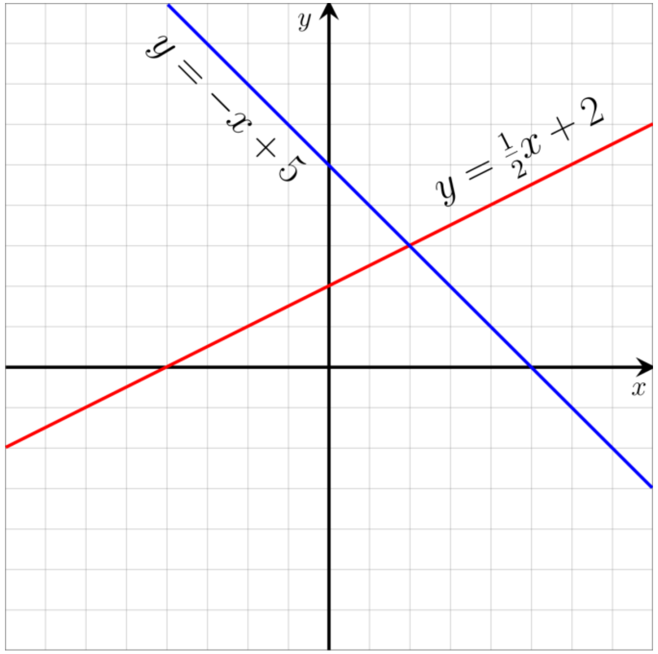
- Biases are indicative of representation of data we store in our minds.

Biases can be good, bad, or neutral

- Biases are indicative of representation of data we store in our minds.

- Biases in ML and Statistics

  - Predictive loss is also a type of bias

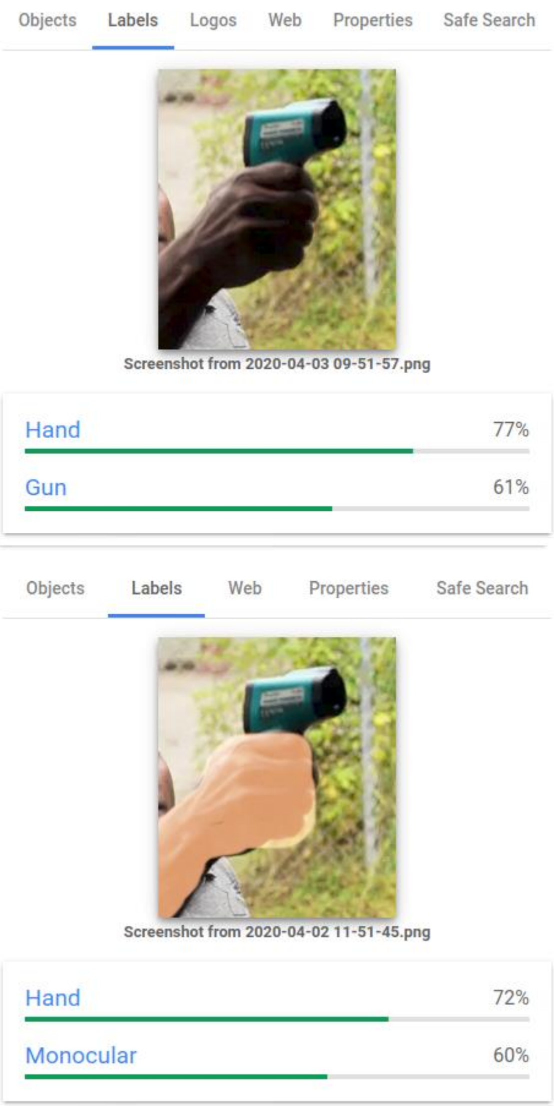


Biases can be good, bad, or neutral

- Biases are indicative of representation of data we store in our minds.
- Biases in ML and Statistics
  - Predictive loss is also a type of bias
- Cognitive Biases
  - Confirmation bias, Recency bias, Optimism Bias

Biases can be good, bad, or neutral

- Biases are indicative of representation of data we store in our minds.
- Biases in ML and Statistics
  - Predictive loss is also a type of bias
- Cognitive Biases
  - Confirmation bias, Recency bias, Optimism Bias
- Algorithmic Biases
  - Unjust, unfair, or prejudicial treatment of people related to race, income, religion, gender, or any other characteristics, when and where they manifest in algorithmic systems or algorithmically aided decision-making.



https://algorithmwatch.org/en/google-vision-racism/

AI is not biased on its own but it can learn biases from biased data

"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will amplify injustice."

- The Guardian

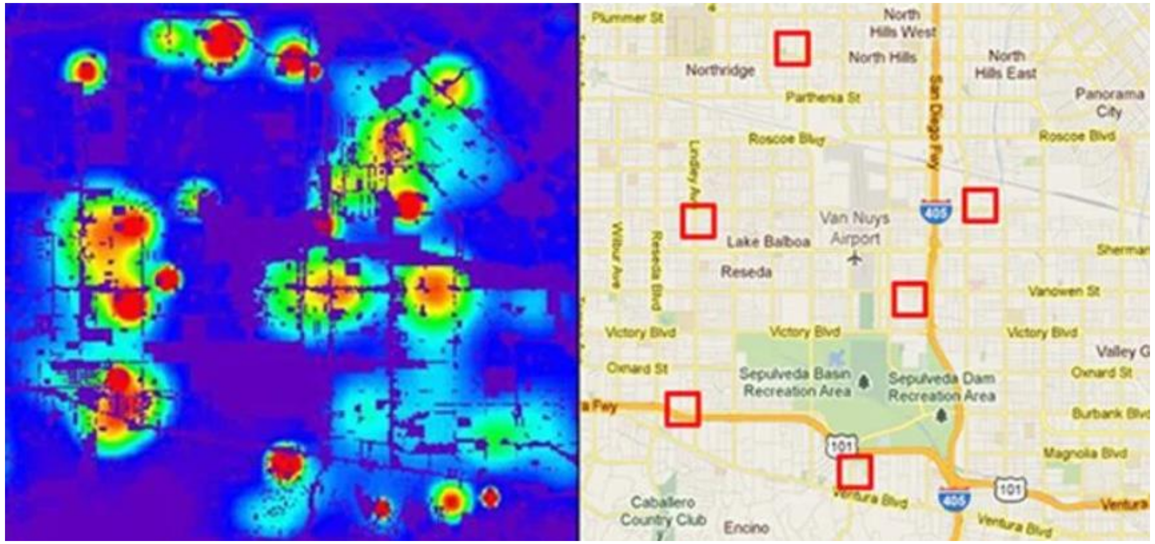
AI is not biased on its own but it can learn biases from biased data

> “Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will **amplify injustice**.”
>
> - The Guardian

AI is not biased on its own but it can learn biases from biased data

- Predictive Policing
  - Algorithm identifies potential crime hotspots based on where previously police have made arrests.
  - What is the problem here?
    - Place of arrest does not necessarily mean place of crime committed.

> “Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will **amplify injustice**.”
>
> \- The Guardian



https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

AI is not biased on its own but it can learn biases from biased data

- Predictive Policing
  - Algorithm identifies potential crime hotspots based on where previously police have made arrests.
  - What is the problem here?
    - Place of arrest does not necessarily mean place of crime committed.
- Predicting Sentencing

"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will **amplify injustice**."

- The Guardian

- Prater (white male) rated low risk after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (black female) rated high risk after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-years prison term for grand theft.

Monahan, John, and Jennifer L. Skeem. "Risk assessment in criminal sentencing." Annual review of clinical psychology 12 (2016): 489-513.

AI is not biased on its own but it can learn biases from biased data

- Predictive Policing
  - Algorithm identifies potential crime hotspots based on where previously police have made arrests.
  - What is the problem here?
    - Place of arrest does not necessarily mean place of crime committed.
- Predicting Sentencing
- What biases are there?

"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will **amplify injustice**."
- The Guardian

- Prater (white male) rated low risk after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (black female) rated high risk after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-years prison term for grand theft.

Monahan, John, and Jennifer L. Skeem. "Risk assessment in criminal sentencing." Annual review of clinical psychology 12 (2016): 489-513.

AI is not biased on its own but it can learn biases from biased data

- Predictive Policing
  - Algorithm identifies potential crime hotspots based on where previously police have made arrests.
  - What is the problem here?
    - Place of arrest does not necessarily mean place of crime committed.

"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data (are) skewed, even by accident, the computers will **amplify injustice**."
- The Guardian

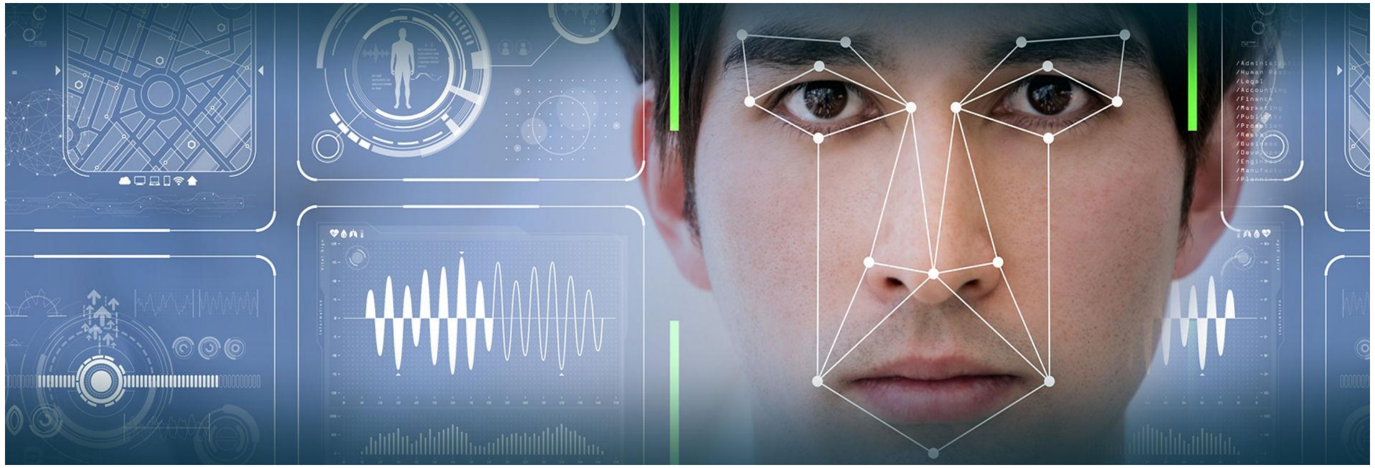
- Predicting Sentencing
- What biases are there?
  - Automation bias
  - Over-generalisation
  - Correlation Fallacy

- Prater (white male) rated low risk after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (black female) rated high risk after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-years prison term for grand theft.

Monahan, John, and Jennifer L. Skeem. "Risk assessment in criminal sentencing." Annual review of clinical psychology 12 (2016): 489-513.

AI is not biased on its own but it can learn biases from biased data

- Faception, an Israeli start-up. They claim,
  - "Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial images."

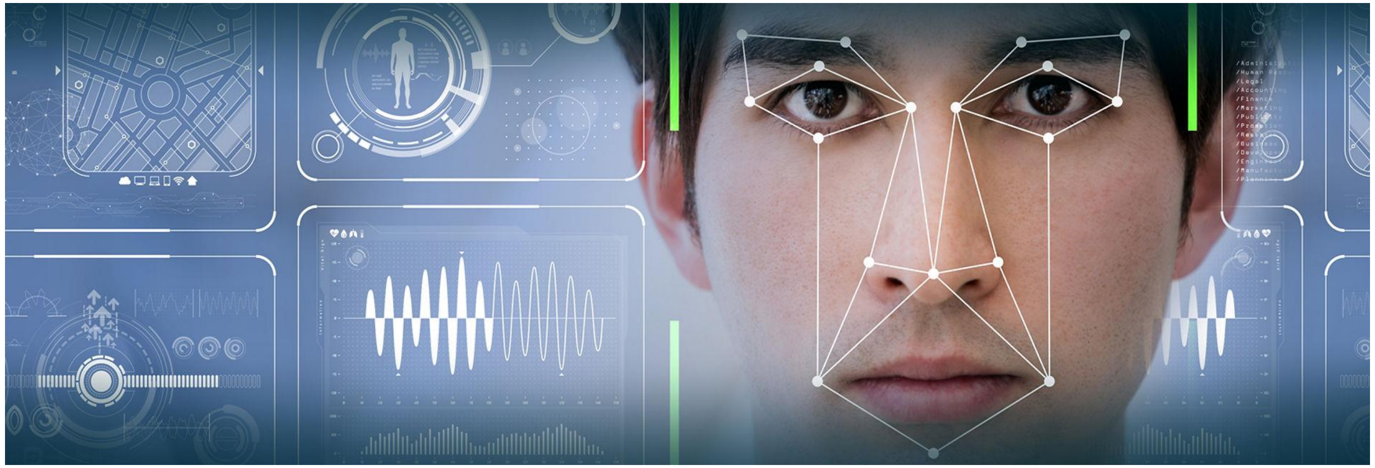


https://www.faception.com/

AI is not biased on its own but it can learn biases from biased data

- Faception, an Israeli start-up. They claim,
    - “Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial images.”

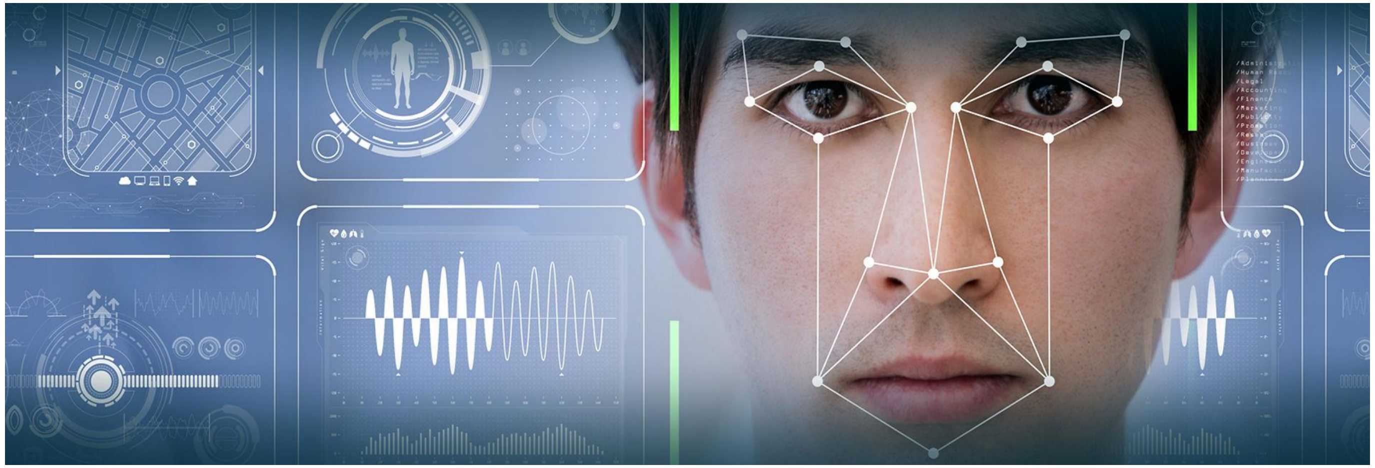
- Offering specialised engines for recognising,
    - High IQ
    - White Collar Offender
    - Paedophile
    - Terrorist



https://www.faception.com/

AI is not biased on its own but it can learn biases from biased data

- Faception, an Israeli start-up. They claim,
  - “Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial images.”



- Offering specialised engines for recognising,
  - High IQ
  - White Collar Offender
  - Paedophile
  - Terrorist
- No information about their training data, methodology, or quantitative results.
  - Still selling fast.

https://www.faception.com/

AI is not biased on its own but it can learn biases from biased data

- Using 1856 closely cropped images of faces, detects if a person is criminal or not with around 90% accuracy.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Wu, Xiaolin, and Xi Zhang. "Automated inference on criminality using face images." arXiv preprint arXiv:1611.04135 (2016): 4038-4052.

AI is not biased on its own but it can learn biases from biased data

- Using 1856 closely cropped images of faces, detects if a person is criminal or not with around 90% accuracy.

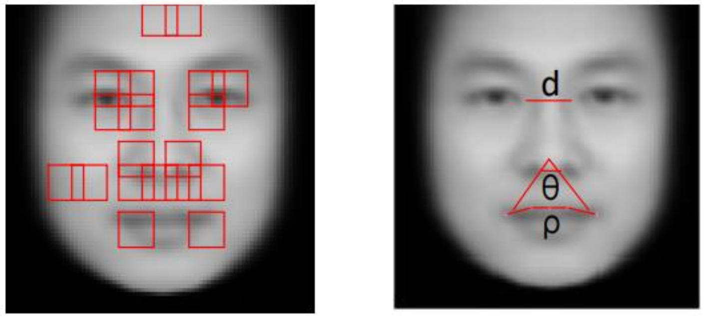
- One interesting finding is,

    - … angle theta from nose tip to two mouth corners in on average 19.6% smaller for criminals than for non-criminals.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.

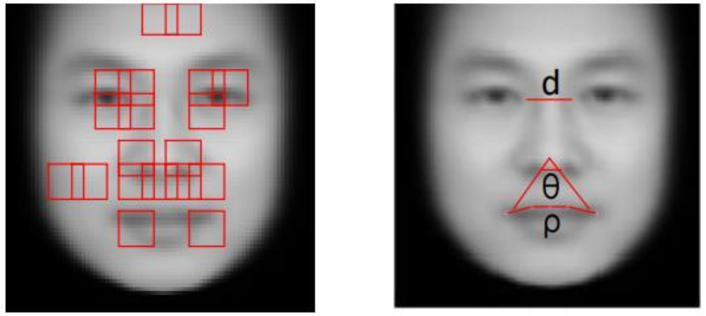
Wu, Xiaolin, and Xi Zhang. "Automated inference on criminality using face images." arXiv preprint arXiv:1611.04135 (2016): 4038-4052.

AI is not biased on its own but it can learn biases from biased data

- Using 1856 closely cropped images of faces, detects if a person is criminal or not with around 90% accuracy.

- One interesting finding is,
    - … angle theta from nose tip to two mouth corners in on average 19.6% smaller for criminals than for non-criminals.

- What biases are at play?
    - Selection bias
    - Experimenter's bias
    - Confirmation bias
    - Correlation fallacy



(a) (b)

Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Wu, Xiaolin, and Xi Zhang. "Automated inference on criminality using face images." arXiv preprint arXiv:1611.04135 (2016): 4038-4052.

NUST Defining futures
School of Electrical Engineering & Computer Science

Evaluate for fairness and inclusion to alleviate the effects of possible biases

- In addition to performing aggregated evaluation, perform **disaggregated evaluation**.
  - For example, for face detection application calculate performance metrics for
    - Men and Women separately
    - Black and White separately

Evaluate for fairness and inclusion to alleviate the effects of possible biases

- In addition to performing aggregated evaluation, perform **disaggregated evaluation**.
  - For example, for face detection application calculate performance metrics for
    - Men and Women separately
    - Black and White separately
- In addition to performing disaggregated evaluation, perform intersectional evaluation.
  - For example, for face detection application calculate performance metrics for
    - Black men and black women

Evaluate for fairness and inclusion to alleviate the effects of possible biases

- In addition to performing aggregated evaluation, perform **disaggregated evaluation**.
    - For example, for face detection application calculate performance metrics for
        - Men and Women separately
        - Black and White separately
- In addition to performing disaggregated evaluation, perform intersectional evaluation.
    - For example, for face detection application calculate performance metrics for
        - Black men and black women
- Intersectionality Theory: Intersecting social identities may relate to systems and structures discrimination.
    - *DeGraffenreid v. General Motors case*

Prof. Kimberlé Crenshaw

How is fairness evaluated quantitatively?

- Fairness may be evaluated using two criteria.
  - Equality of Opportunity:
    - Equal recall across all subgroups
  - Predictive Parity:
    - Equal precision across all subgroups

How is fairness evaluated quantitatively?

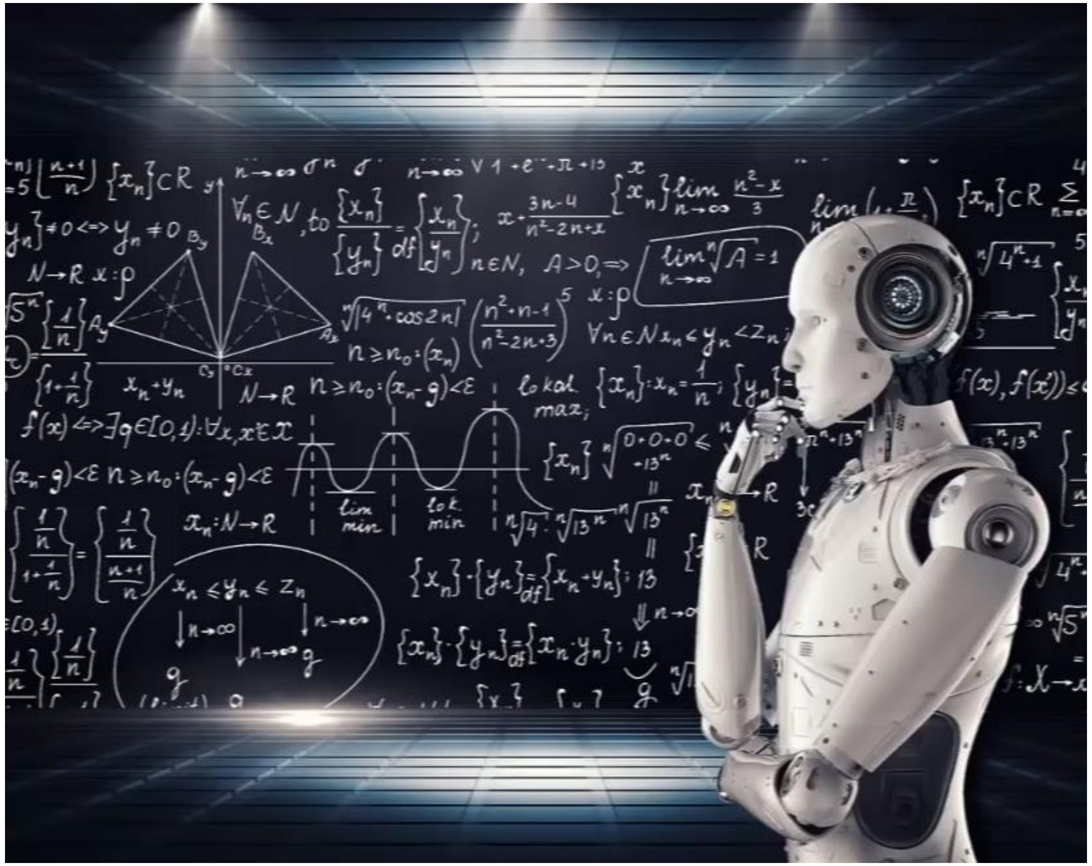
- Fairness may be evaluated using two criteria.
  - Equality of Opportunity:
    - Equal recall across all subgroups
  - Predictive Parity:
    - Equal precision across all subgroups
- One metric might have higher weightage over the other depending upon application.
  - False positives might be better than false negatives in some situations.

How is fairness evaluated quantitatively?

- Fairness may be evaluated using two criteria.
    - Equality of Opportunity:
        - Equal recall across all subgroups
    - Predictive Parity:
        - Equal precision across all subgroups
- One metric might have higher weightage over the other depending upon application.
    - False positives might be better than false negatives in some situations.
    - False negatives may be preferred over false positives in others.

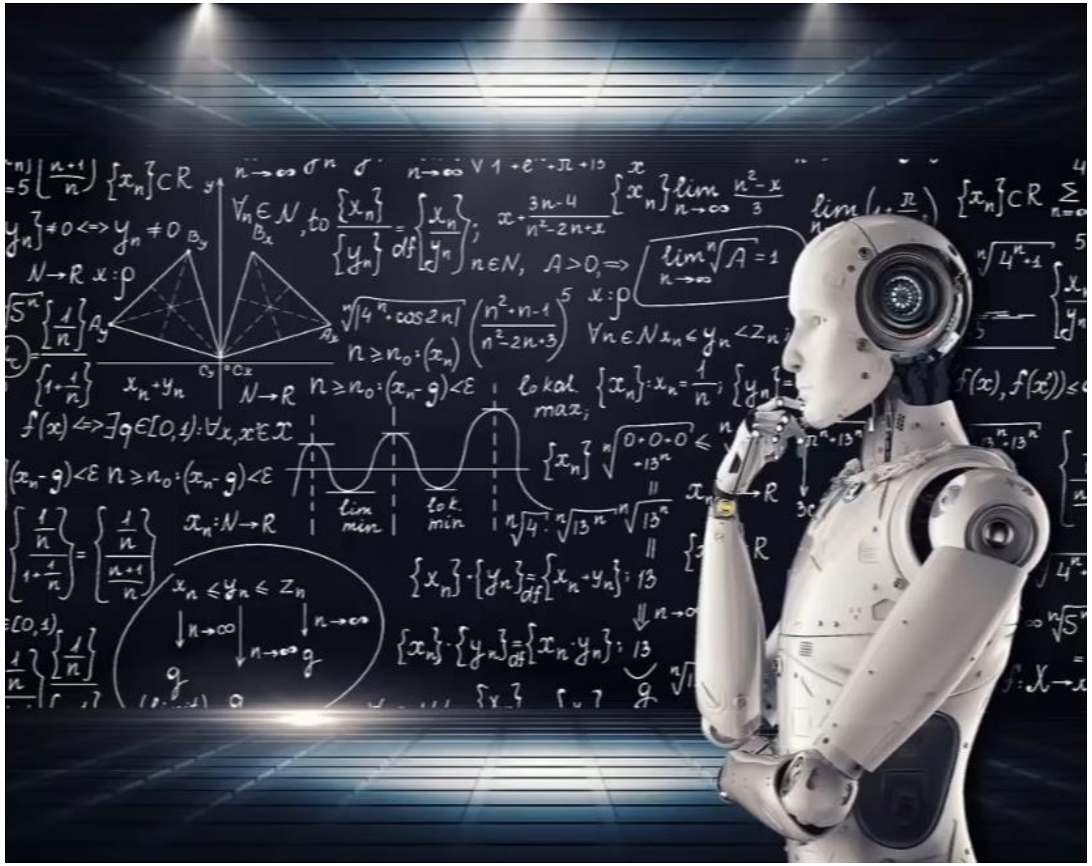
Many factors can lead AI to read unjust outcomes

- Lack of insights into sources of biases in the data and model.
- Lack of insights into the feedback loops.



Many factors can lead AI to read unjust outcomes

- Lack of insights into sources of biases in the data and model.
- Lack of insights into the feedback loops.
- Lack of careful, disaggregated evaluations.
- Human biases in interpreting and accepting results.



Many factors can lead AI to read unjust outcomes

- Lack of insights into source
- Lack of insights into the fee
- Lack of careful, disaggrega
- Human biases in interpretin

Remember!
AI is not inherently biased.
You're biased

How can we prevent biases in AI?

- Understand your data. It matters. Like, really matters.
  - Correlations, limitations, skews of data should be well-understood before using them.
  - Either design models that take care of those correlations or use data-augmentation to address skews.

How can we prevent biases in AI?

- Understand your data. It matters. Like, really matters.
  - Correlations, limitations, skews of data should be well-understood before using them.
  - Either design models that take care of those correlations or use data-augmentation to address skews.
- In absence of pre-define train/test sets, use multiple random train/test sets or use k-fold cross validation.
  - If possible use a separate small test set for hard/important cases.

How can we prevent biases in AI?

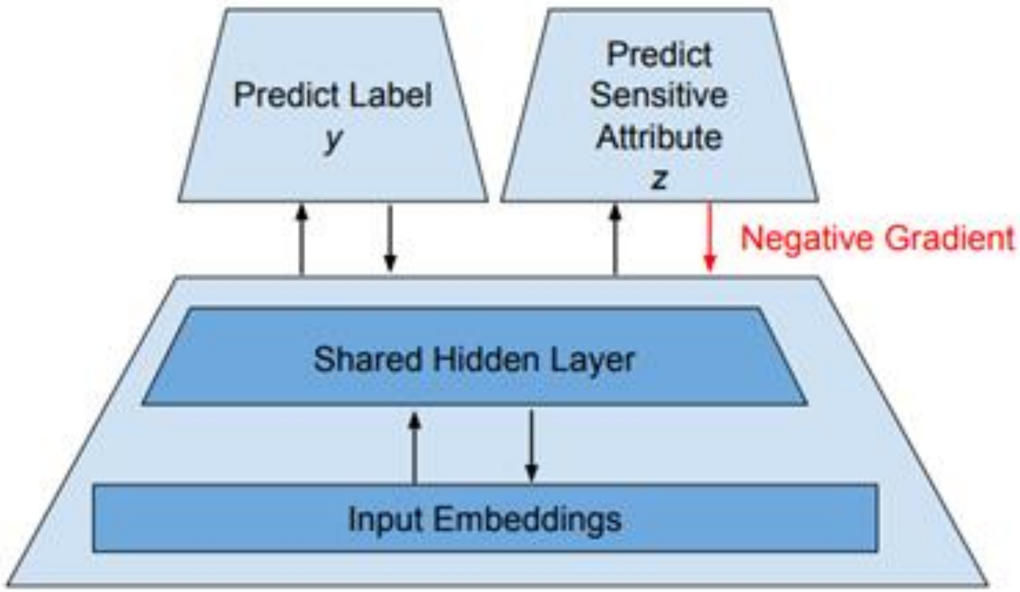
- Understand your data. It matters. Like, really matters.
  - Correlations, limitations, skews of data should be well-understood before using them.
  - Either design models that take care of those correlations or use data-augmentation to address skews.
- In absence of pre-define train/test sets, use multiple random train/test sets or use k-fold cross validation.
  - If possible use a separate small test set for hard/important cases.
- Combine inputs from multiple sources.

How can we prevent biases in AI?

- Understand your data. It matters. Like, really matters.
    - Correlations, limitations, skews of data should be well-understood before using them.
    - Either design models that take care of those correlations or use data-augmentation to address skews.
- In absence of pre-define train/test sets, use multiple random train/test sets or use k-fold cross validation.
    - If possible use a separate small test set for hard/important cases.
- Combine inputs from multiple sources.
- **Remember:** A dataset without any bias has not been curated yet.

Multitask adversarial learning maybe used to mitigate biases

- Jointly predict
  - Output decision D.
  - Attribute Z that you would like to remove from D.
  - Negate the effect of the undesired attribute.

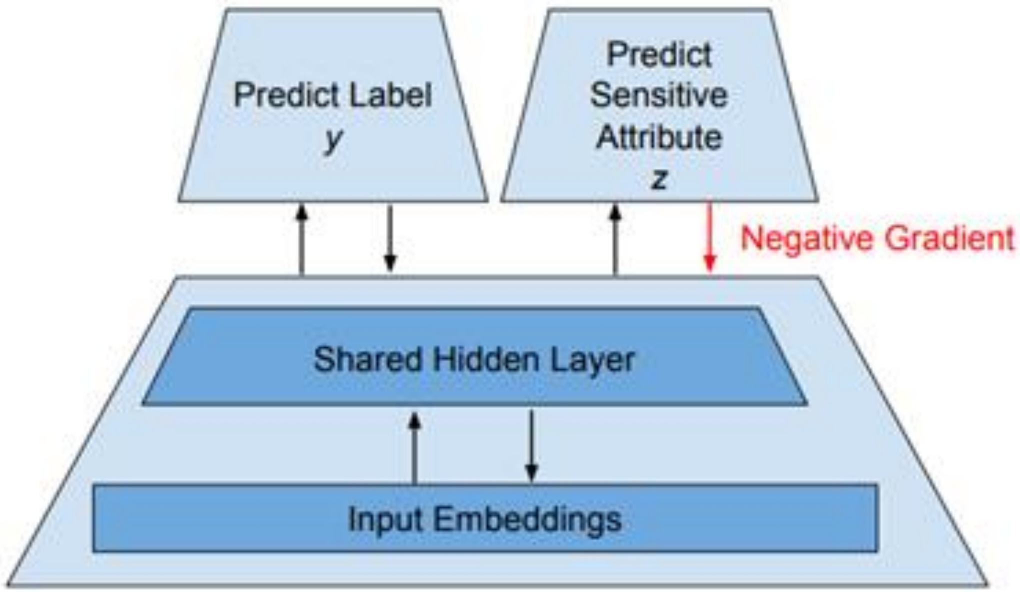


Beutel, Alex, et al. "Data decisions and theoretical implications when adversarially learning fair representations." arXiv preprint arXiv:1707.00075 (2017).

Multitask adversarial learning maybe used to mitigate biases

- Jointly predict
  - Output decision D.
  - Attribute Z that you would like to remove from D.
  - Negate the effect of the undesired attribute.
    - Leads to demographic parity.

$$P(\hat{y} = 1 \mid y = 1, z = 1) = P(\hat{y} = 1 \mid y = 1, z = 0)$$

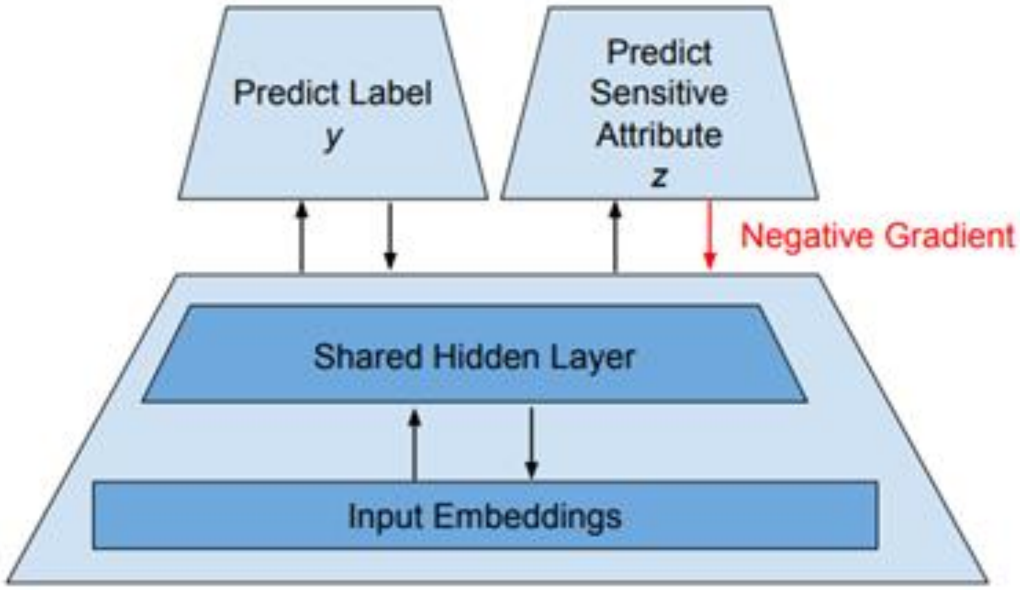


Beutel, Alex, et al. "Data decisions and theoretical implications when adversarially learning fair representations." arXiv preprint arXiv:1707.00075 (2017).

Multitask adversarial learning maybe used to mitigate biases

- Jointly predict
  - Output decision D.
  - Attribute Z that you would like to remove from D.
  - Negate the effect of the undesired attribute.
    - Leads to demographic parity.

$$P(\hat{y} = 1\,|y = 1, z = 1) = P(\hat{y} = 1\,|y = 1, z = 0)$$



- Track demographic parity using two criteria

$$ParityGap = |ProbTrue_1 - ProbTrue_0|$$

$$EqualityGap_y = |ProbCorrect_{y,1} - ProbCorrect_{y,0}|$$

Beutel, Alex, et al. "Data decisions and theoretical implications when adversarially learning fair representations." arXiv preprint arXiv:1707.00075 (2017).

AI models may falsely identify bias when there is none

- Sometimes model may falsely associate frequently attacked identities with toxicity.

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Frequency of identity terms in toxic comments and overall

Dixon, Lucas, et al. "Measuring and mitigating unintended bias in text classification." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

AI models may falsely identify bias when there is none

- Sometimes model may falsely associate frequently attacked identities with toxicity.

- The model associates toxicity with the terms instead of the context in which they are used.

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Frequency of identity terms in toxic comments and overall

Dixon, Lucas, et al. "Measuring and mitigating unintended bias in text classification." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

AI models may falsely identify bias when there is none

- Sometimes model may falsely associate frequently attacked identities with toxicity.

- The model associates toxicity with the terms instead of the context in which they are used.

- Solution? Synthetic dataset with template based text.

  - Sort of data augmentation to provide non-toxic use of terms.

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Table 2: Phrase template examples.

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Frequency of identity terms in toxic comments and overall

Dixon, Lucas, et al. "Measuring and mitigating unintended bias in text classification." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

Do you have any problem?

Some material (images, tables, text etc.) in this presentation has been borrowed from different books, lecture notes, and the web. The original contents solely belong to their owners, and are used in this presentation only for clarifying various educational concepts. Any copyright infringement is ***not at all*** intended.