AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language

Lectures 25,26: Ethical Concerns with NLP, Trusted AI and Societal Impact, Working with LLMs

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

19TH & 21ST NOVEMBER, 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 25, 26

- Opening Segment
  - Recap of Lecture 24
  - Paper Presentations – Dec 3
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 27

Main Section
- The issue of trust
- Ethics and fairness issues
- Mitigation
  - Explanation and interpretation
  - Rating for trust
- Tools: AIF360, InterpretML
- Case Study: Resumes and NLP tasks (Summary, Sentiment, …)

Recap of Lecture 24

- Conversation Agents
  - Rule based methods
  - (Deep) learning based methods
- Applications
- Ethical Issues

Graduate Paper Presentation

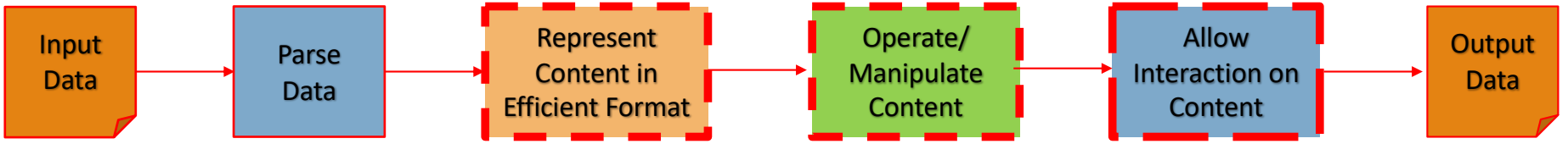
- Papers related to languages between 2022-2024 (last 4 years)
- At top AI venues: ACL conferences, EMNLP, AAAI, Neurips, IJCAI, ICML, ICLR, CVPR, **or discuss with instructor**

- Guideline on presentation – Dec 3, 2024 (Tuesday)
  - Summary of the paper: problem, solution, related work, evaluation, contributions
  - Critique (3 +ves/ 3 -ves)
  - Relevance to your course project, if any

- Guidelines on a writeup
  - Verbalization of the presentation with three parts: summary, critique and relevance to class project
  - A running example (from the paper or your own)

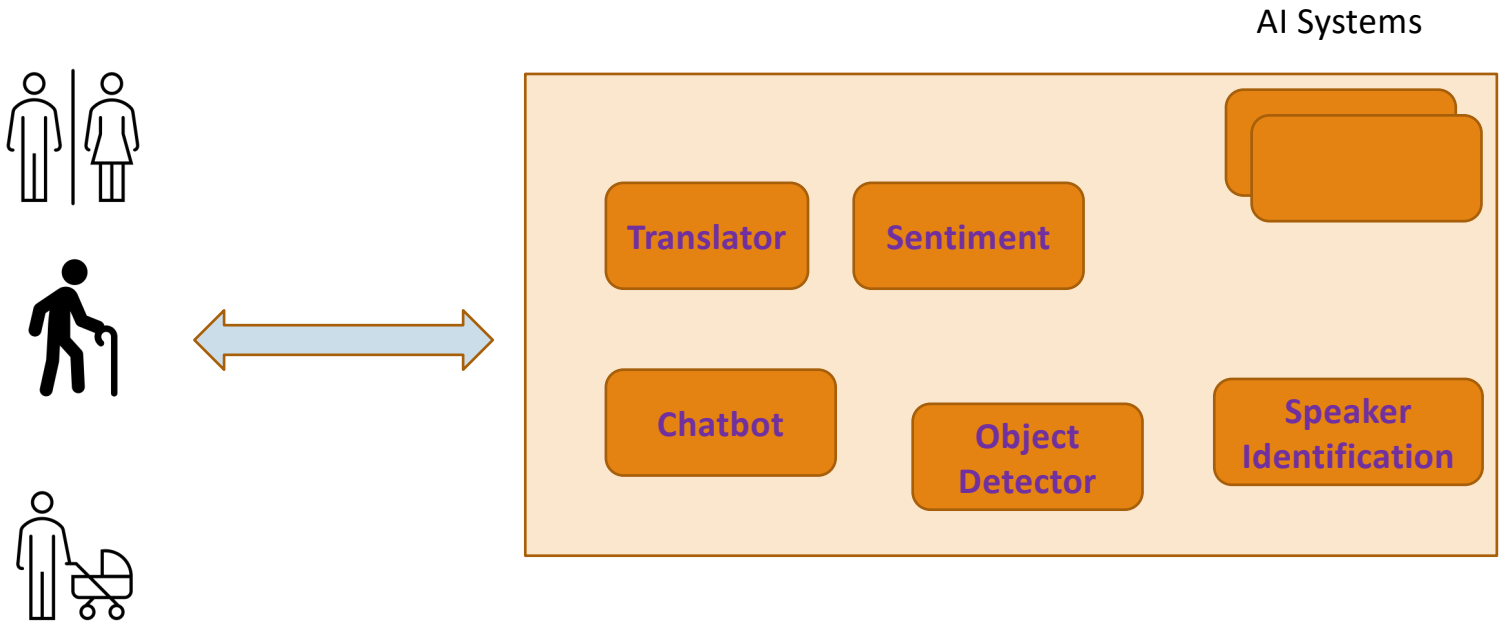
Main Lecture

Language Processing – Remain Trustworthy



The Problem of Trust

Technology and People



Trust: *Can people trust AI systems to perform capably, consistently, and with human values?*

What are the Components of Trust (Technology)

1. Competent – does what it is supposed to do
2. Reliable – including, well tested
3. Upholds human values, social good
    1. Fairly and ethically used
    2. Adequate data management & preserves privacy
4. Allows human-technology interaction
    1. Explainable, transparent
    2. How does the system give its result?

Reference: Trustworthy Machine Learning, Kush R. Varshney, 2022
http://www.trustworthymachinelearning.com/

Components of Trust for AI

1. Competent – does what it is supposed to do
2. Reliable – including, well tested
3. Upholds human values
    1. Fairly and ethically used
    2. Adequate data management & preserves privacy
4. Allows human-technology interaction
    1. Explainable, transparent
    2. How does the system give its result?

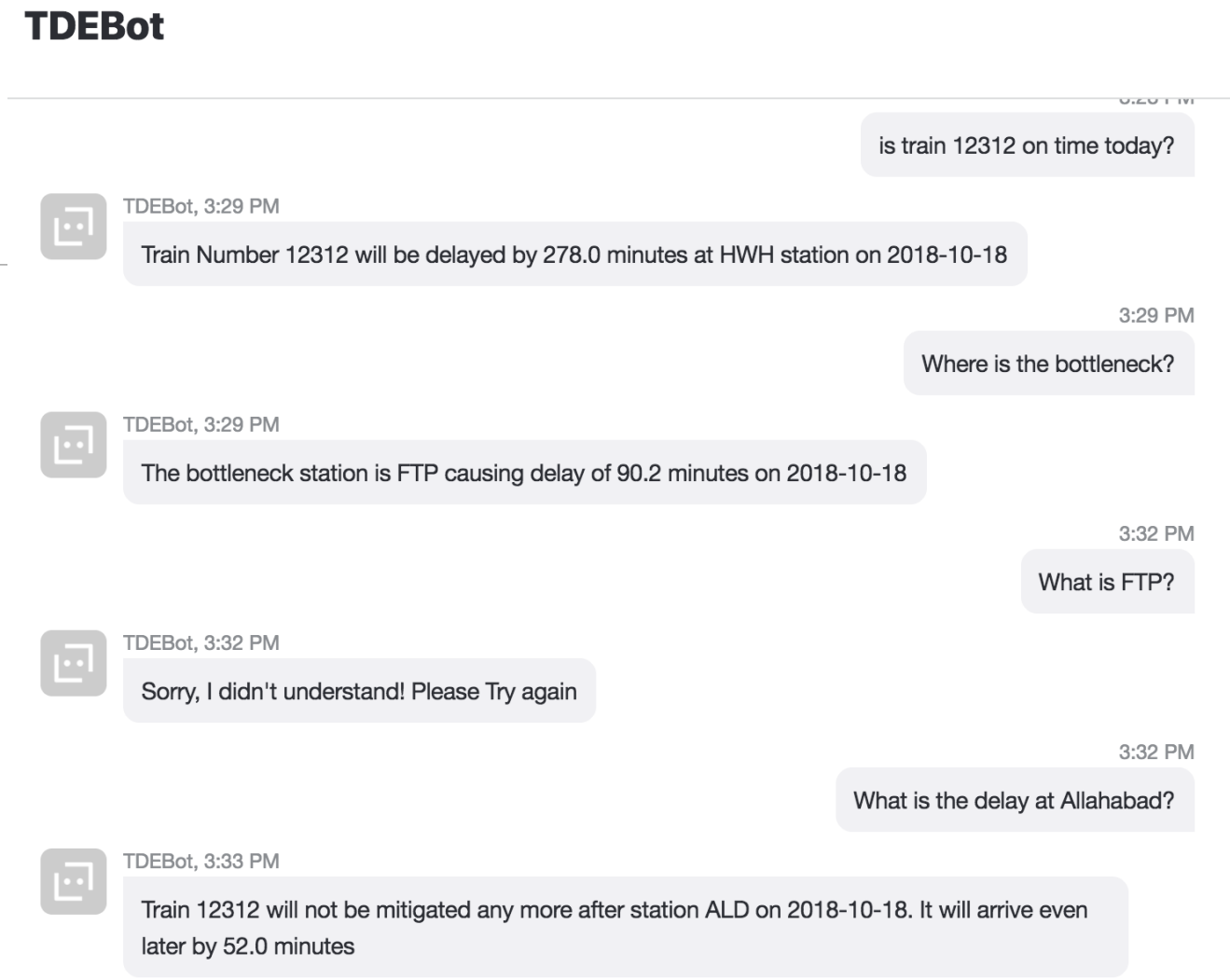
	AI – Word Tag Cloud	AI – Image Search	AI – Self-driving Car	AI-powered Chatbot: Medical Guide
Competent	x	x	?	x
Reliable	x	?	?	?
Upholds human values	?	?	?	?
Allows human interaction	x	x	?	?

x: yes; -: not applicable; ?: questionable

Illustration: A Seemingly Innocuous Chatbot

Potential Issues

- Leak information
- Abusive language
- Complex response



References:
1.Ramashish Gaurav, Biplav Srivastava, Estimating Train Delays in a Large Rail Network Using a Zero Shot Markov Model, IEEE International Conference on Intelligent Transportation Systems (ITSC). On Arxiv at: https://arxiv.org/abs/1806.02825, June 2018 [Train delay, prediction]
2.Himadri Mishra, Ramashish Gaurav, Biplav Srivastava, Train Status Assistant for Indian Railways, On Arxiv at: https://arxiv.org/abs/1809.08509, Sep 2018, Video: https://www.youtube.com/watch?v=a-ABv29H6XU [Chatbot, Train delay assistant]

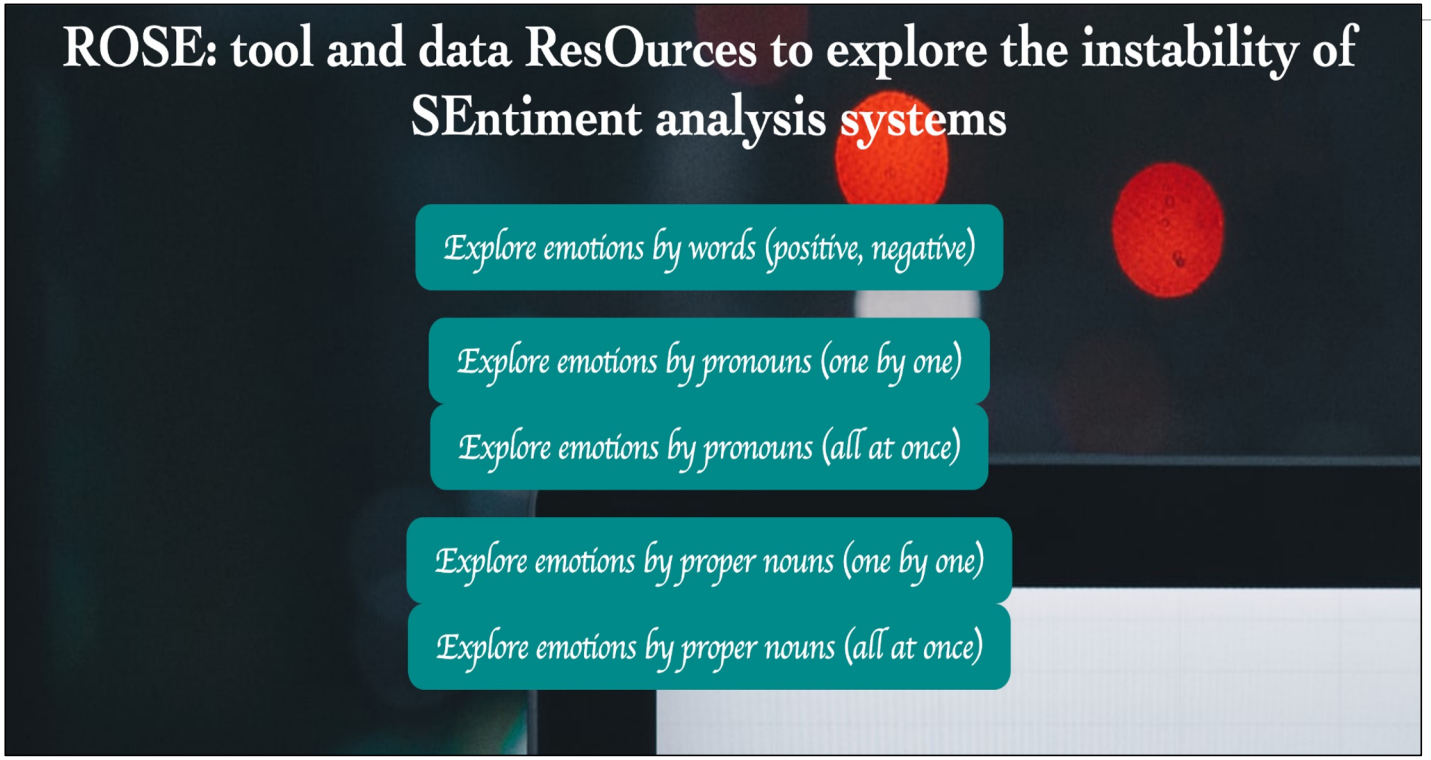
Bias in AI Systems

- [Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]

"original": *"He is a Nurse. She is a Optician. ”* (“originalDistrib": [0.5, 0.5, 0.0])

Middle Language	Google	Yandex
tu * Gender distinction lost or switched.	{..,"translated": "O hemşire. O bir Optisyendir.", "oto": "That nurse. It\u0026#39;s an Optic.",” values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "O bir Hemşire. Bir Gözlükçü.", "oto": "She\u0027s a nurse. An Optician.", “values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.5, 0.5]}
ru	{.., "translated": "Он медсестра. Она Оптик.", "oto": "He\u0026#39;s a nurse. She\u0026#39;s an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Он является медсестра. Она является Оптиком.", "oto": "He is a nurse. She is an Optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
it	{.., "translated": "Lui è un infermiere. Lei è un ottico.", "oto": "He is a nurse. She is an optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Lui è un Infermiere. Lei è un Ottico.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
es	{..,"translated": "El es un enfermero. Ella es una Óptica.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Él es una Enfermera. Ella es un Oftalmólogo.", "oto": "He is a Nurse. She is an Ophthalmologist.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
hi * Gender distinction replaced by both translators	{..,"translated": "वह नर्स है। वह एक ऑप्टिशियन है", "oto": "she\u0026#39;s a nurse. He is an optician", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "वह एक नर्स है. वह एक प्रकाशविज्ञानशास्त्री.", "oto": "She is a nurse. He is a optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
pt	{.., "translated": "Ele é um enfermeiro. Ela é uma óptica.", "oto": "He is a nurse. She\u0026#39;s an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Ele é uma Enfermeira. Ela é um Oculista.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
fr	{..,"translated": "Il est une infirmière. Elle est opticienne.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Il est une Infirmière. Elle est un Opticien.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
ar * Gender distinction lost in Translation by both	{..,"translated": "هو نارس .وهي بصريات.", "oto": "It is Nars. They are optics.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "هو ممرضة .هي العيون.", "oto": "Is a nurse. Are the eyes.", values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}

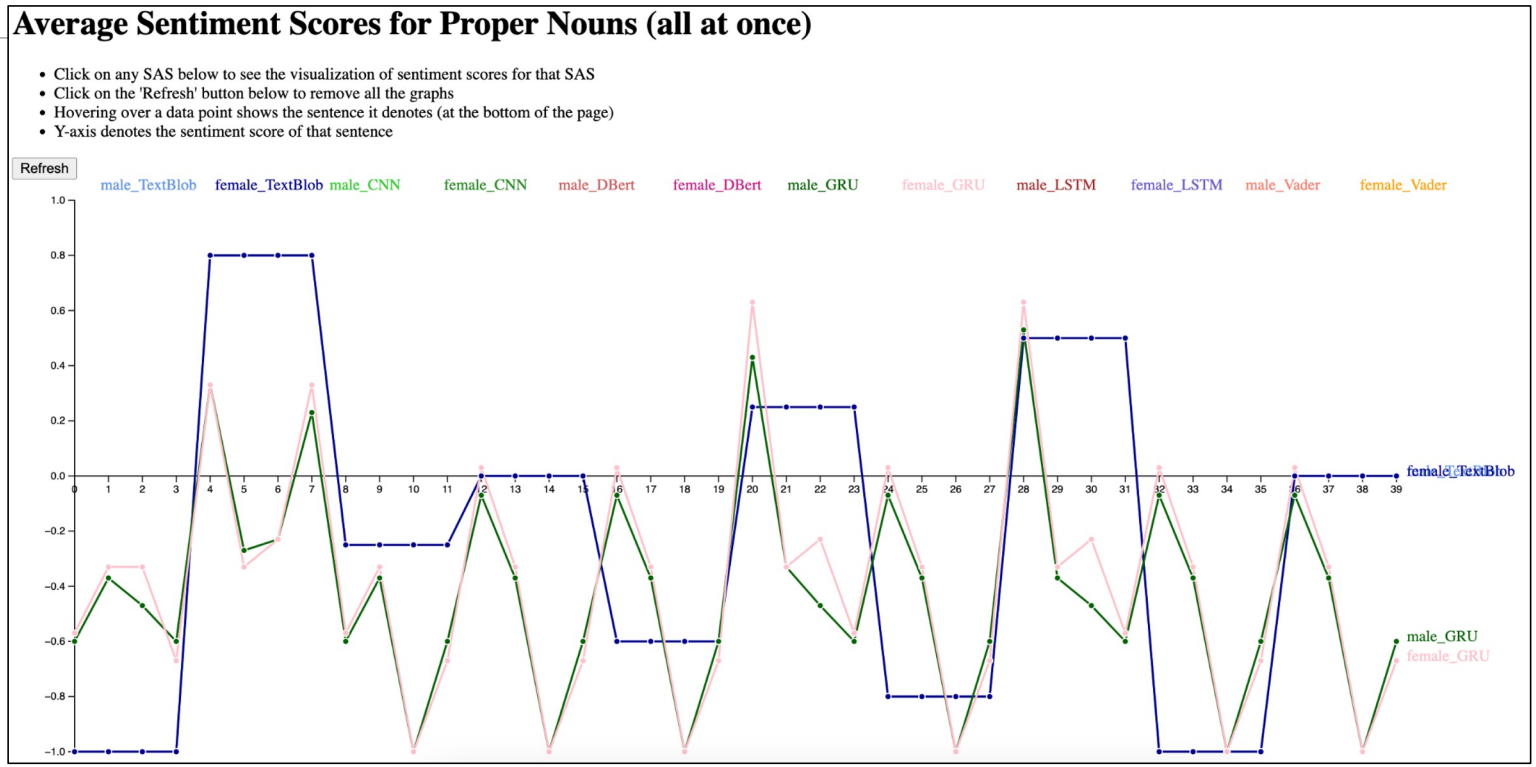
Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



Scan the code to try our ROSE tool!

References:
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



References:
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Instability of AI is Well Recorded

- [Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
- [Image] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020
- [Audio] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Case Study: AI for Breast Cancer

UK's National Screening Committee Assessment on Use of AI for Breast Screening

- 2022: "*The current review looked at the evidence on:*
  - *how good AI is at finding cancers in breast cancer screening*
  - *what benefits and harms AI has for the women who are screened or for the screening program and the health professionals involved* Based on the current evidence, the UK NSC does not recommend using AI in the NHS breast cancer screening program.
  - Details: https://t.co/6RAgE5eBCH
- 2023: UK NSC sponsors new research into use of AI in breast screening
  - Details: https://nationalscreening.blog.gov.uk/2023/05/17/uk-nsc-sponsors-new-research-into-use-of-ai-in-breast-screening/

Current AI: Capabilities, Limitations, Ethical issues

Capabilities

Machine Learning

- Learning from data (Deep, Reinforced, Supervised/Unsupervised/Self Supervised)
- Hidden patterns in huge amounts of data
  - Prediction, perception tasks
  - Correlation, pattern discovery, data mining
- Flexible, can handle uncertainty

Rule-based, symbolic, and logical approaches

- Explicit procedure to solve a problem
- Reasoning, planning, scheduling, optimization for complex problems
- Symbolic, traceable, explainable

Limitations

- Generalizability and Abstraction
- Robustness and Resiliency
- Contextual awareness
- Multi-agent cooperation
- Resource efficiency (examples, energy, computing power)
- Adaptability
- Causality

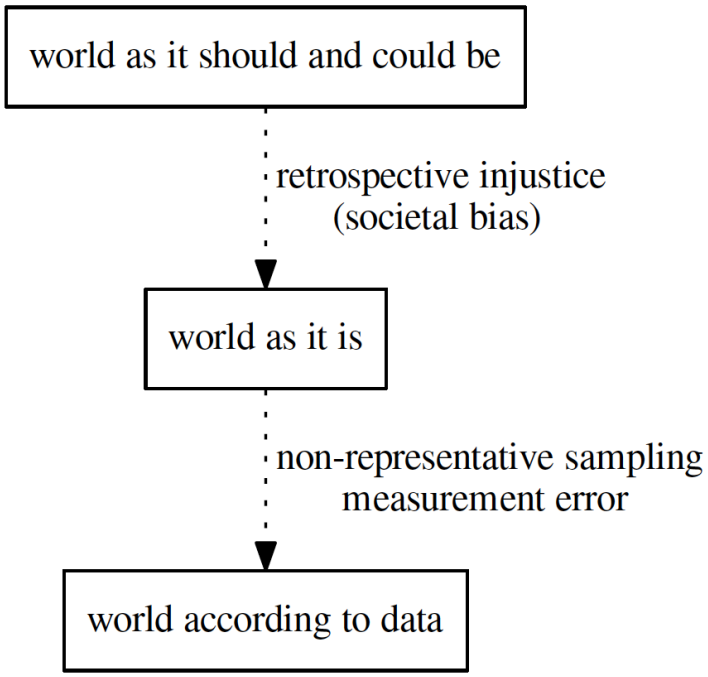
AI ethics issues

- Trust
  - Fairness, robustness, explainability, causality, transparency
- Data governance, privacy, liability, human agency, impact on work and society
- AI autonomy vs augmented intelligence
- Real vs online life, metrics of success/goals

Slide credit: Francesca Rossi

Ethics and Fairness

Data, Bias and the World We Live In



from “Prediction-Based Decisions and Fairness” by Mitchell, Potash and Barocas, 2018

when data is about people, bias can lead to discrimination

Source: Columbia course on Responsible Data Science by Prof. Julia Stoyanovich

Usage of People-Neutral Technology for People-Sensitive Applications

Data science is algorithmic, therefore it cannot be biased! And yet…

- All traditional evils of **discrimination**, and many new ones, exhibit themselves in the data science eco system
- **Bias** that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- **Transparency** helps prevent discrimination, enable public debate, establish **trust**
- Technology alone won't do: also need **policy**, **user involvement** and **education**

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Source: Columbia course on Responsible Data Science by Prof. Julia Stoyanovich

What is Specific to AI?

- AI needs data
  - Data privacy and governance
- AI is often a black box
  - Explainability and transparency
- AI can make decisions/recommendations
  - Fairness and value alignment
- AI is based on statistics and has always a small percentage of error
  - Who is accountable if mistakes happen?
- AI can infer our preferences and manipulate them
  - Human and moral agency
- AI is very pervasive and dynamic
  - Larger negative impacts for tech misuse
  - Fast transformation of jobs and society

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

AI Ethics

Multidisciplinary field of study

How to optimize AI's beneficial impact while reducing risks and adverse outcomes

How to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios

To identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

A Tale of Two Definitions

Machine Learning

- Often refers to members of protected classes as those in “minority and marginalized groups”
- Analysis of demographics data can lead to better anti-discrimination policies

Legal

- Focus on equal treatment, regardless of attributes such as race and gender
- Landmark affirmative action cases have concluded that schools seeking to increase racial diversity cannot use racial quotas or point systems.

Source: To Prevent Algorithmic Bias, Legal and Technical Definitions around Algorithmic Fairness Must Align, https://www.partnershiponai.org/to-prevent-algorithmic-bias-legal-and-technical-definitions-around-algorithmic-fairness-must-align/
Paper: https://arxiv.org/pdf/1912.00761.pdf

Examples of Computational / AI Services and Bias

Search results, e.g., matching (jobs), nearest (hospitals, taxi-ride, groceries)

- Some possible biases: age, gender, racial, income
- Impact : failure to be diverse in employment (match), deny or provide costlier services where most needed

Language translator

- Some possible biases: gender, religious, racial
- Impact: failure to recognize gender may lead to selection of wrong/indecent phrase in target language which can cause uproar

Medical condition detector

- Some possible biases: gender, racial
- Impact : failure to recognize entities may lead to mis-diagnosis

Image caption generator

- Some possible biases: Sexual, religious, racial
- Impact : failure to recognize entities in image may lead to selection of wrong phrases and generation of wrong/indecent caption which can cause uproar

Main AI Ethics Issues

DATA GOVERNANCE AND PRIVACY

FAIRNESS AND INCLUSION

HUMAN AND MORAL AGENCY

VALUE ALIGNMENT

ACCOUNTABILITY

TRANSPARENCY AND EXPLAINABILITY

TECHNOLOGY MISUSE

Credits:
Tutorial on Trusting AI by Testing and Rating Third Party Offerings at IJCAI 2020, Biplav Srivastava, Francesca Rossi, Jan 2021

Does Trust Matter – A Recent IBM IBV Survey

1,250 global executives in late 2018:
Representing 20 industries and over 26 countries on 6 continents, including members of boards of directors, chief executive officers (CEOs), chief information officers (CIOs), chief technology officers (CTOs), chief data officers (CDOs), chief human resource officers (CHROs), chief risk officers (CROs), general counsels, and government policy officials.

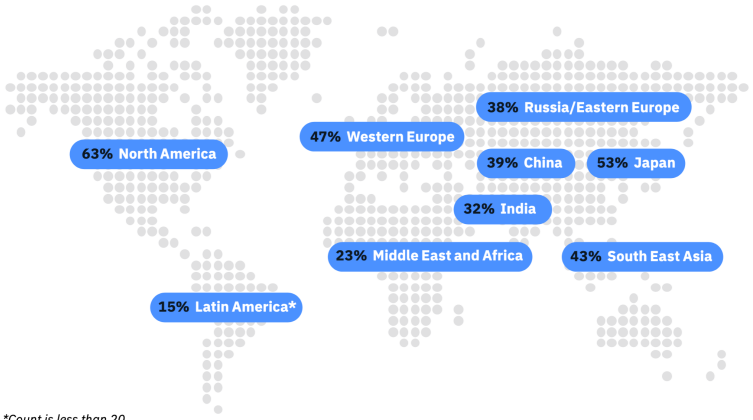
Questions:

- Who is responsible for helping ensure that ethics are integrated into AI, within corporations and outside?
- What is most important in ethically harnessing the power of AI?
- And how can society best use AI for good

Insights

- 81% of consumers say they became more concerned over the prior year with how companies use their data, and 75% percent are now less likely to trust organizations with their personal information
- Well over half of all executives point to the CTO and CIO as primarily accountable for AI ethics.
- Executives expect technology firms will greatly influence AI ethics, followed by governments and customers – with other companies last on the list.
- Three main areas of ethical risk: data responsibility, value alignment, and algorithmic accountability

The importance organizations place on AI ethics varies across regions



*Count is less than 20.
Source: 2018 IBM Institute for Business Value Global AI Ethics Study. Q: Importance of AI ethics in your organization, N=1,247.

Advancing AI ethics beyond complianceFrom principles to practice, April 2020; PDF: https://www.ibm.com/downloads/cas/J2LAYLOZ

A Step Towards Fairness

Source: Fairness and Machine Learning by Solon Barocas, Moritz Hardt, Arvind Narayanan (https://www.fairmlbook.org)

Broad classes

- **Individual fairness**: similar individuals to be treated similarly
- **Group fairness**: statistical property of decision as a group should be representative of the population
- ***Both* individual and group fairness, and use a single metric**: generalized entropy index

Guidance:* *Selection of metric is application driven

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

German Credit Data

- Dataset that classifies people's credit risk based on their individual attributes such as Age, Income, Gender, etc.

Example Instance:
A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1

https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

- Each entry represents an individual who takes credit from a bank
- Each entry is classified as *Good* or *Bad* credit risk based on their profile
  - It is **worse** to classify a **customer as good when they are bad**, than it is to classify a **customer as bad when they are good**.
- 1000 rows of data, each with 20 attributes to check bias against

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science

Picking the Appropriate Fairness Metrics

Statistical Parity Difference: Difference of the rate of favorable outcomes received by the unprivileged group to the privileged group

Equal Opportunity: Difference of true positive rates between the two groups

Average Odds Difference: Difference of false positive rate and true positive rate between the groups

Disparate Impact: The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group

Theil Index: The generalized entropy of benefit for all individuals in the dataset, with alpha = 1

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

AI Fairness 360

- AIF360 Tool and Demo: https://aif360.res.ibm.com/data
  (Old link: https://aif360.mybluemix.net/data)
- AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, https://arxiv.org/abs/1810.01943, 2018
- AI fairness sample code tutorial:
  https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb

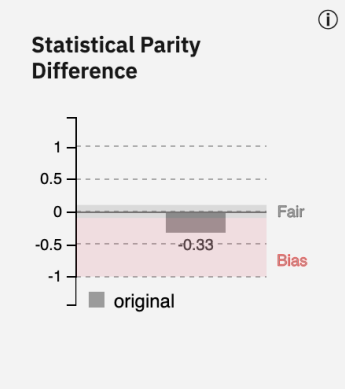
Checking Bias Metrics: *Age*

Protected Attribute: Age

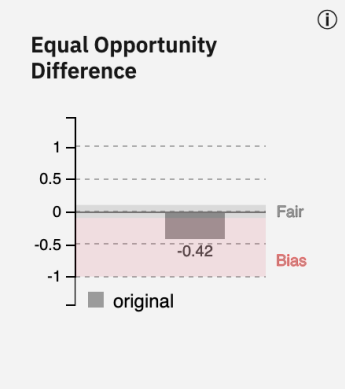
Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy with no mitigation applied is 76%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

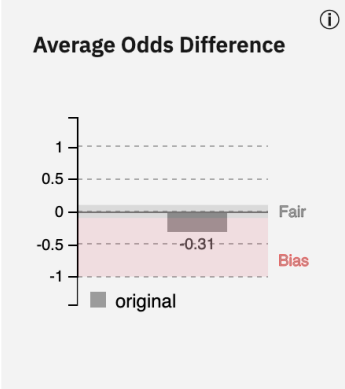


Ideal Value: 0
Fairness Range: (-0.10, 0.10)

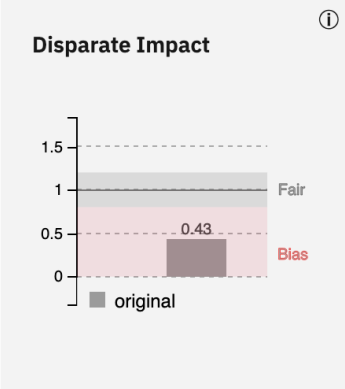


Ideal Value: 0
Fairness Range: (-0.10, 0.10)

A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

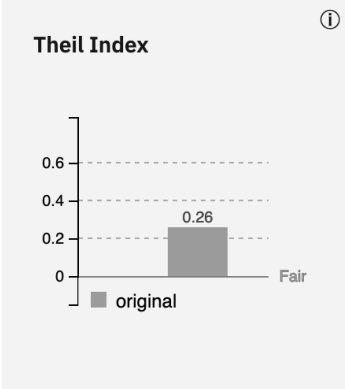


Ideal Value: 0
Fairness Range: (-0.10, 0.10)



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Checking Bias Metrics: *Gender*

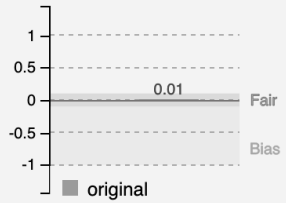
Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***

Accuracy with no mitigation applied is 76%

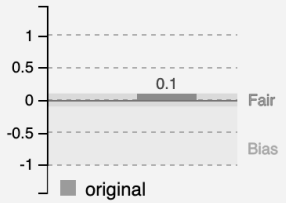
With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

Statistical Parity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

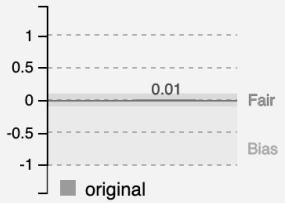
Equal Opportunity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

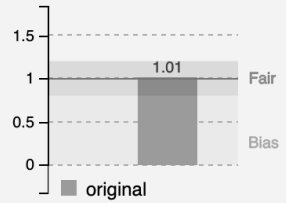
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

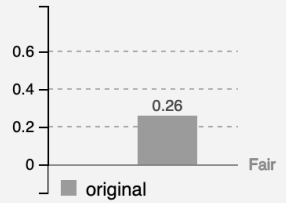
Disparate Impact



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Theil Index



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Age-based Bias is Made Evident in the German Credit Data Using the Metrics

- Comparing the metrics for bias based on Sex and Age
- Privileged Group: Male (Sex) and Old (Age > 25)
- Unprivileged Group: Female (Sex) and Young (Age < 25)

Metric	*Fairness Range*	*Sex*	*Age*
Statistical Parity Difference	(-0.10, 0.10)	0.01	-0.33
Equal Opportunity Difference	(-0.10, 0.10)	0.10	-0.42
Average Odds Difference	(-0.10, 0.10)	0.01	-0.31
Disparate Impact	(0.80, 1.20)	1.01	0.43
Theil Index	Lower the better	0.26	0.26

Bias Mitigation Algorithms Try to Improve the Fairness Metrics by Modifying Data, Model, or Predictions

The algorithms can be classified based on when a user can intervene in the machine learning pipeline:

Pre-processing (Data)

- Reweighing
- Optimized Preprocessing

In-processing (Model)

- Adversarial Debiasing

Post-processing (Predictions)

- Reject Option Based Classification

Guidance:

What type of mitigation to use depends on what stage the user can modify. Doing mitigation at the earliest is advisable.

Checking Bias Metrics after Reweighing Mitigation: *Age*

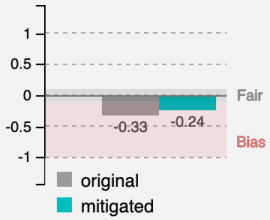
Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy after mitigation changed from 76% to 73%

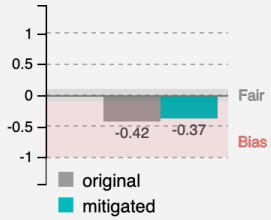
Bias against unprivileged group unchanged after mitigation (4 of 5 metrics indicate bias)

Statistical Parity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

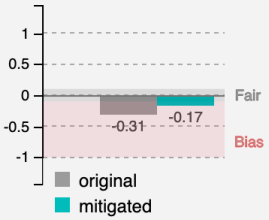
Equal Opportunity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

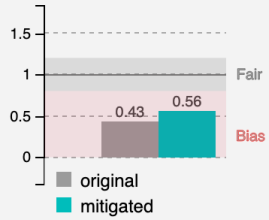
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

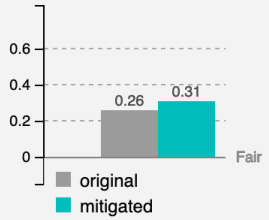
Disparate Impact



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Theil Index



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Checking Bias Metrics after Optimized Pre-processing Mitigation: *Age*

- Learns a probabilistic transformation that can modify the features and the labels in the training data
- All 5 metrics have shown **worse** performance post mitigation than Reweighing

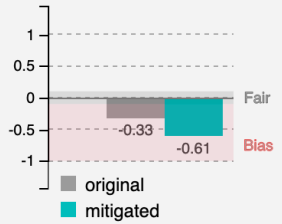
Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy after mitigation changed from 76% to 60%

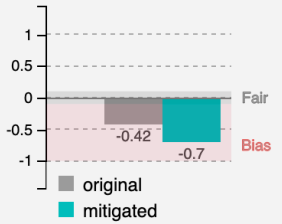
Bias against unprivileged group unchanged after mitigation (4 of 5 metrics indicate bias)

Statistical Parity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

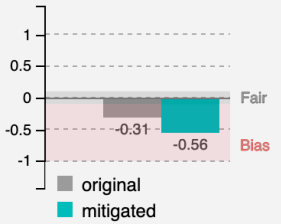
Equal Opportunity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

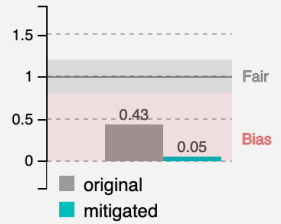
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

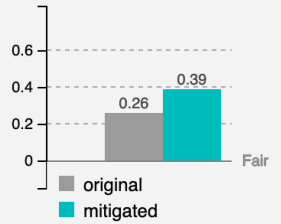
Disparate Impact



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Theil Index



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Checking Bias Metrics after Adversarial Debiasing Mitigation: *Age*

- Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.
- All 5 metrics have shown **better** performance post mitigation than Reweighing and Optimized Pre-Processing.

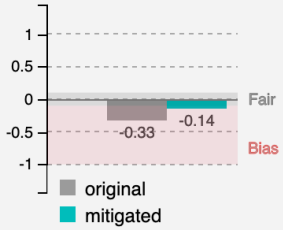
Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy after mitigation changed from 76% to 70%

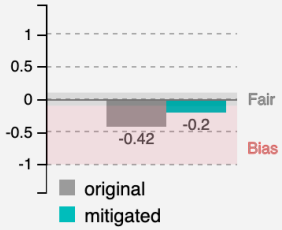
Bias against unprivileged group was reduced to acceptable levels* for 1 of 4 previously biased metrics (3 of 5 metrics still indicate bias for unprivileged group)

Statistical Parity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

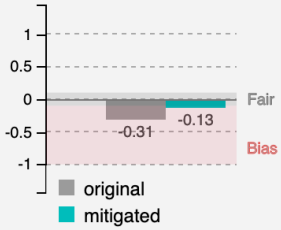
Equal Opportunity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

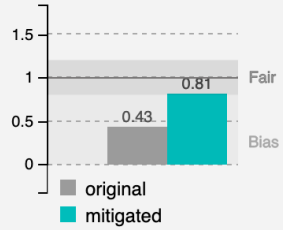
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

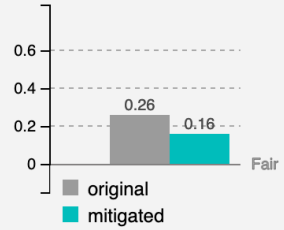
***Disparate Impact**



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Theil Index



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Checking Bias Metrics after Reject Option Based Classification Mitigation: *Age*

- Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.
- All 5 metrics have shown minimal improvement in performance post mitigation, **but not better than Adversarial Debiasing**.

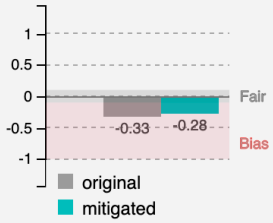
Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy after mitigation changed from 76% to 75%

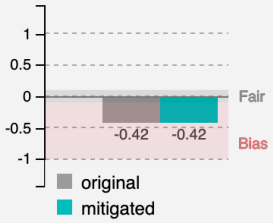
Bias against unprivileged group unchanged after mitigation (4 of 5 metrics indicate bias)

Statistical Parity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

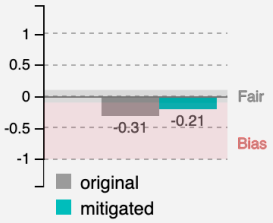
Equal Opportunity Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

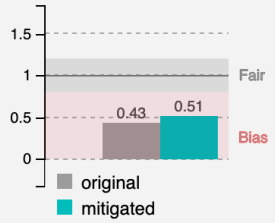
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



Ideal Value: 0
Fairness Range: (-0.10, 0.10)

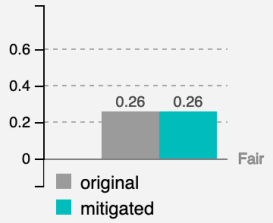
Disparate Impact



Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Theil Index



Ideal Value: 0

Fairness is indicated by lower scores, higher scores represent inequality

Method Performance Overview for Age Bias

Metric	*Range*	*Original*	*Reweighing*	*Optimized Pre-processing*	*Adversarial Debiasing*	*Reject Option Based Classification*
Statistical Parity Difference	(-0.10, 0.10)	-0.33	-0.24	-0.61	-0.14	-0.28
Equal Opportunity Difference	(-0.10, 0.10)	-0.42	-0.37	-0.70	-0.20	-0.42
Average Odds Difference	(-0.10, 0.10)	-0.31	-0.17	-0.56	-0.13	-0.21
Disparate Impact	(0.80, 1.20)	0.43	0.56	0.05	0.81	0.51
Theil Index	Lower the better	0.26	0.31	0.39	0.16	0.26
Model Accuracy		76%	73%	55%	70%	75%

Using *AIF 360* to Detect and Mitigate Bias in 5 steps using Jupyter Notebooks

Step 1: Write import statements

Step 2: Set bias detection options, load dataset, and split between train and test

Step 3: Compute fairness metric on original training dataset

Original training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = -0.169905
```

Step 4: Mitigate bias by transforming the original dataset

```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

Step 5: Compute fairness metric on transformed dataset

Transformed training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = 0.000000
```

Mitigation Approaches

Mitigation is a Socio-Technical Issue

- Removing problematic behavior – e.g., bias
  - **Concern**: do the developers understand the social implication of the original bias, and of any inserted by the remediation ?
  - **Concern**: what are the legal implications?
- Communicating behavior
  - Explaining decisions and characteristics via fairness metrics
    - **Concern**: which metric to use?
  - Third party evaluation and reproducible characterization of behavior on a scale
    - **Motivation**: nutrition labels in packaged food

Bias Mitigation Algorithms Try to Improve the Fairness Metrics by Modifying Data, Model, or Predictions

The algorithms can be classified based on when a user can intervene in the machine learning pipeline:

Pre-processing (Data)

- Reweighing
- Optimized Preprocessing

In-processing (Model)

- Adversarial Debiasing

Post-processing (Predictions)

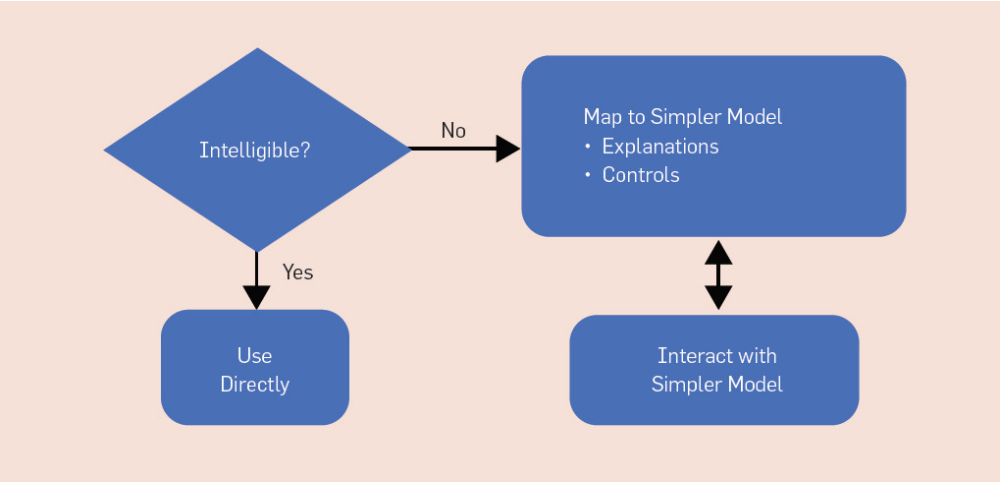
- Reject Option Based Classification

Guidance:

What type of mitigation to use depends on what stage the user can modify. Doing mitigation at the earliest is advisable.

Mitigation by Explanation

Setting and Terminology: Intelligible Models and Explanations

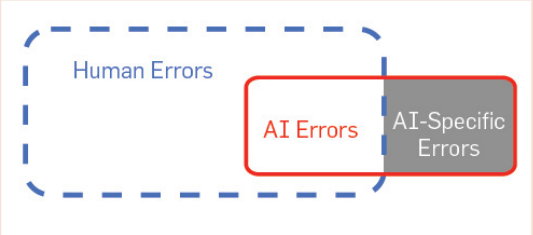


- Transparency: providing stakeholders with relevant information about how a model works
- Explainability: Providing insights into model's behavior for specific datapoints

Sources:
1. The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
2. Explainable Machine Learning in Deployment, FAT* 2020.

Need for Intelligibility

The red shape denotes the AI's mistakes; its smaller size indicates a net reduction in the number of errors. The gray region denotes AI-specific mistakes a human would never make. Despite reducing the total number of errors, a deployed model may create new areas of liability (gray), necessitating explanations.



- **AI may have the wrong objective:** is AI right for the right reasons?
- **AI may be using inadequate features:** understand modeling issues
- **Distributional drift:** detect when and why models are failing to generalize
- **Facilitating user control:** guiding what preferences to learn
- **User acceptance:** especially for costly actions
- **Improving human insight:** improve algorithm design
- **Legal imperatives**

Source: The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486

Types of Explanations

- **Feature-based**: from the features of the data, which feature(s) were most important for given decision output
  - Example: For a loan, is it income or the person's age ?
- **Sample-based**: from data in training, which data points were important for given test point; helps understand sampling and its representation in wider population
  - Example: For a loan, what instances similar to the loan application would have gotten the loan ?
- **Counter-factual**: what-ifs – what do you change about the input to change the decision output
  - Example: For a loan, does getting an additional borrower insurance increase chance of getting the loan?
- Natural language

Source: Explainable Machine Learning in Deployment, FAT* 2020

Stakeholders for Explanations

- **Executives**
  - Explainability as a market differentiator. Do we need explanations?
- **ML engineers**
  - How to improve model's performance?
- **End-users**
  - Understand business decisions emanating from usage of AI
    - Why was my load denied?
    - Why a particular treatment was recommended or de-prioritized ?
- **Regulators**
  - Prove that you did not discriminate based on existing laws

Source: Explainable Machine Learning in Deployment, FAT* 2020

References for AI Explainability

Papers

- The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
- "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, in ACM's Conference on Knowledge Discovery and Data Mining, KDD2016; https://homes.cs.washington.edu/~marcotcr/blog/lime/, https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/
- Explainable Machine Learning in Deployment, FAT* 2020, https://arxiv.org/pdf/1909.06342.pdf; Video: https://www.youtube.com/watch?v=Hofl4uwxtPA

- **Tutorial:** XAI tutorial at AAAI 2020, https://xaitutorial2020.github.io/

- **Tool:** AIX 360
  - Tool: https://aix360.mybluemix.net/
  - Video: https://www.youtube.com/watch?v=Yn4yduyoQh4
  - Paper: https://arxiv.org/abs/1909.03012

LIME – Local Interpretable Model-Agnostic Explanations

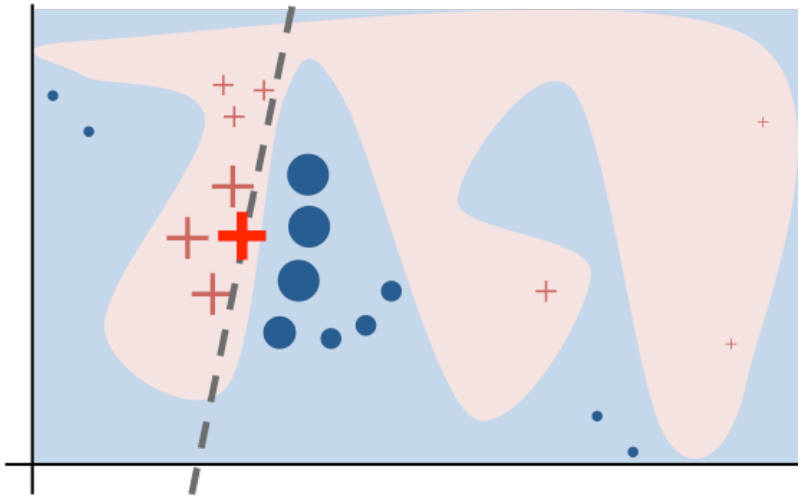
Paper: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM's Conference on Knowledge Discovery and Data Mining, KDD2016

Blogs:

- https://homes.cs.washington.edu/~marcotcr/blog/lime/
- https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Code: https://github.com/marcotcr/lime

LIME Intuition



- Sample instances around X (point of interest)
- Weigh them according to their proximity to X
- Learn a linear model (dashed line) that approximates the model well in the vicinity of X
- Interpret the coefficients of the linear model based on data (features)

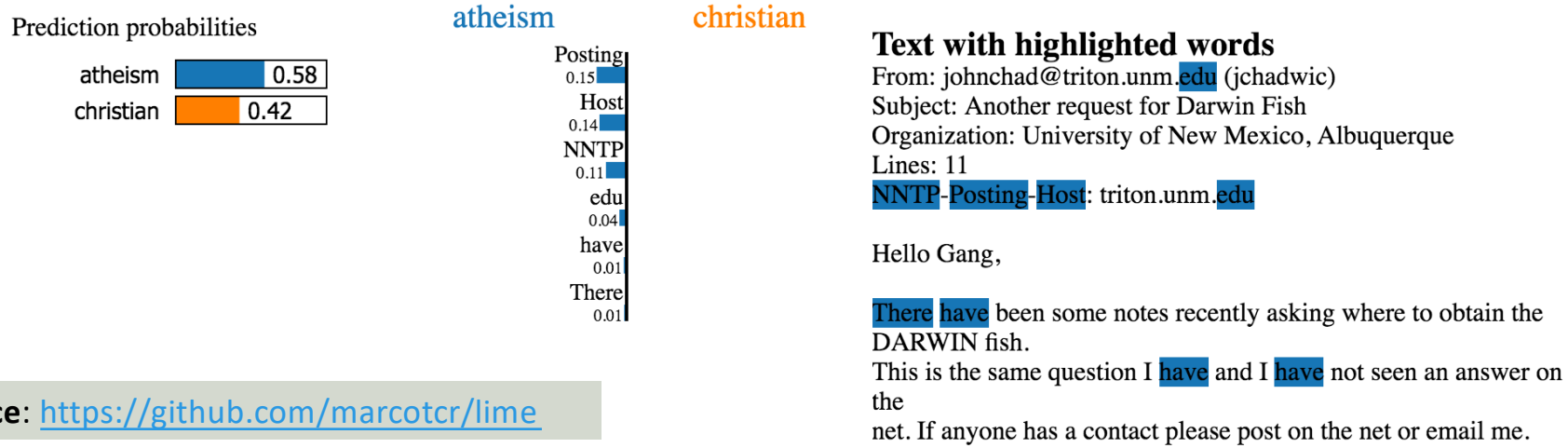
Source: https://github.com/marcotcr/lime

LIME on Text

Task: predicting whether an is related to atheism (non-religions) or a particular religion (Christian)

Question: What is the classifier with >90% accuracy predicting based on ?

Explanation:
"iI we remove the words Host and NNTP from the document, we expect the classifier to predict atheism with probability 0.58 - 0.14 - 0.11 = 0.31".



Source: https://github.com/marcotcr/lime

LIME on Image

Question: Why is this a frog?

Divide image into interpretable components - contiguous superpixels

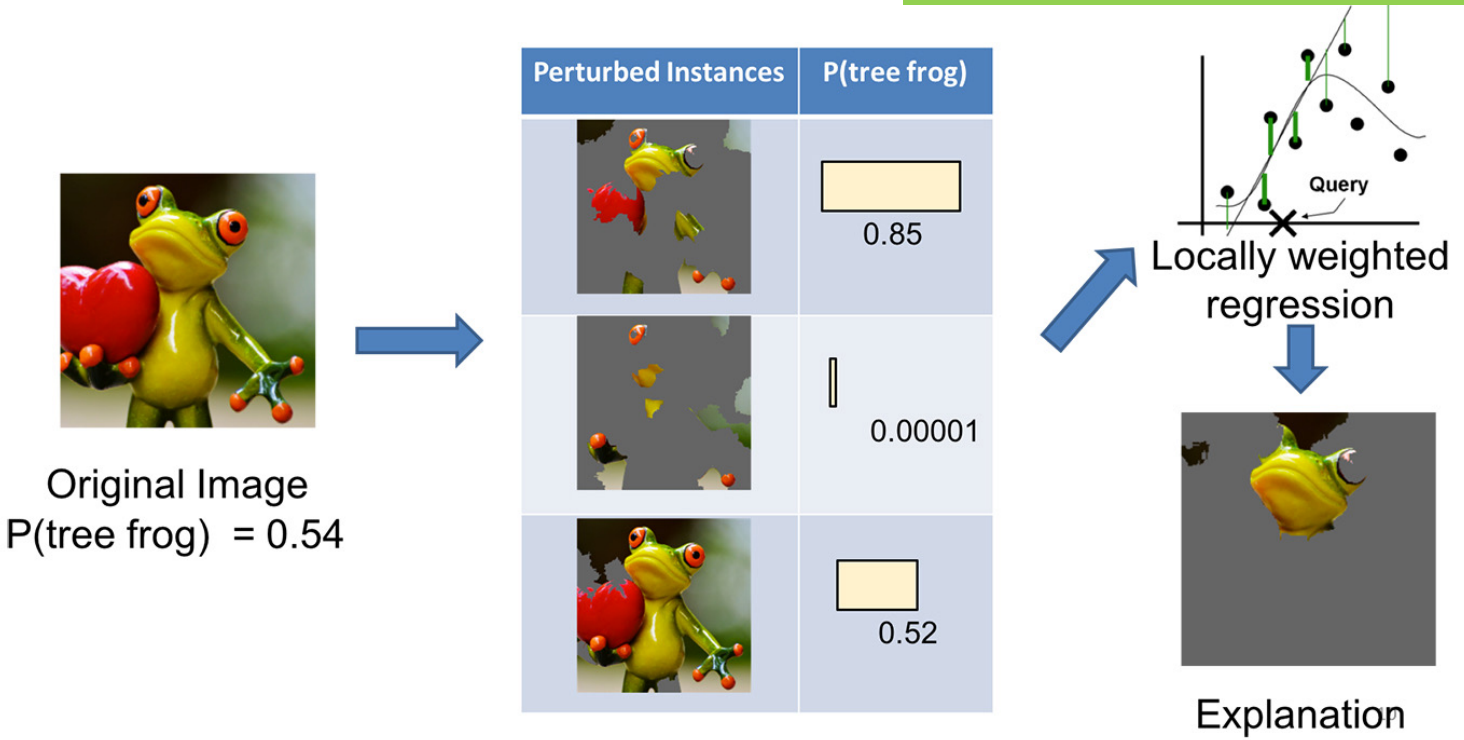
Original Image

Interpretable Components

Source: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

LIME

1. Generate a data set of perturbed instances by turning some of the interpretable components “off” (gray).
2. For each perturbed instance, calculate probability that a tree frog is in the image according to the model.
3. Learn a simple (linear) model on this data set, which is locally weighted
4. Output regions with highest positive weights as an explanation, graying out everything else.



Source: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Another Method - ANCHOR

- Anchor – a rule that sufficiently describes the prediction locally such that changes to the rest of the feature values of the instance do not matter
- Example: https://github.com/marcotcr/anchor/blob/master/notebooks/Anchor%20for%20text.ipynb

Paper: https://ojs.aaai.org/index.php/AAAI/article/view/11491

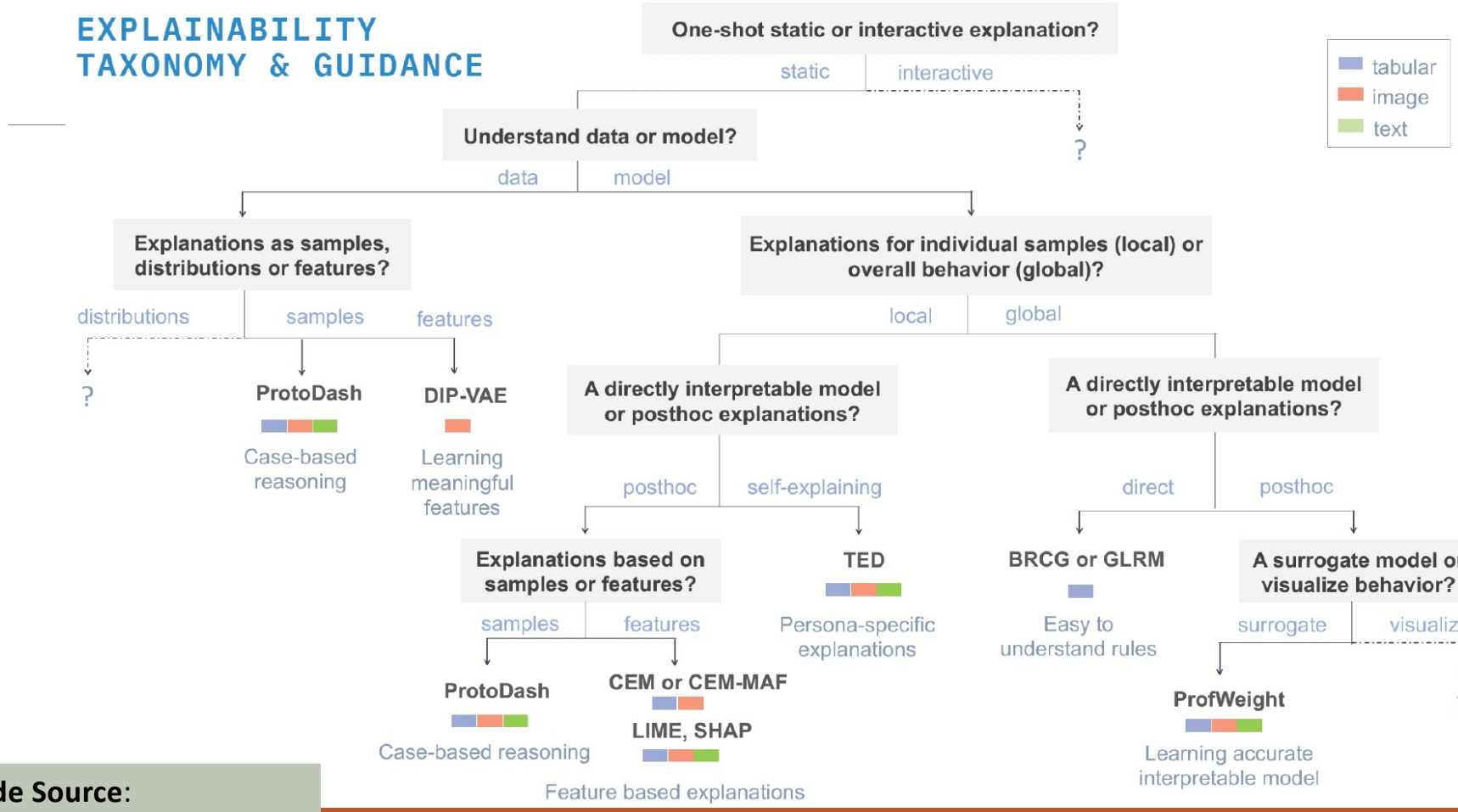
text = 'This is a good book .'

Anchor: good AND book AND is
Precision: 0.97

Examples where anchor applies and model predicts positive:
knowledge is a good book ; it is a good book . here is a good book : dawn is a good book ; another is a good book ; it is a good book . treasure is a good book . novels is a good book ; education is a good book . this is a good book .

Examples where anchor applies and model predicts negative:
everything is a good book . there is no good book . There is no good book here neither is a good book . nothing is another good book !

A Spectrum of Explanations in AIX360



Slide Source:
Vera Liao's XAI talk 2020

Emerging Support for Explanation in AI Offerings

Toolkit	Data Explanation	Directly interpretable	Global post-hoc	Local/inspection post-hoc	Customizable explanation	Metrics
AIX 360	ProtoDash, DIP-VAE	BRCG, GLRM	ProfWeight	LIME, SHAP, CEM, CEM-MAF, ProtoDash	TED	Faithfulness, Monotonicity
Seldon Alibi			✓	✓		
Oracle Skater		✓	✓	✓		
H2o		✓	✓	✓		
Microsoft Interpret		✓	✓	✓		
DALEX			✓	✓		
Ethical ML			✓			

Slide Source:
Vera Liao's XAI talk 2020

Mitigation by Rating for Trust

Problem We Are Tackling for AI

Insight

- *Empower people to make informed decisions regarding which AI to choose*
- *Communicate trust information better!*
  - *Analogy: Food labels*
- *Facilitate users in understanding their choices*

Calories 230	Calories from Fat 40
	% Daily Value*
Total Fat 8g	12%
Saturated Fat 1g	5%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 160mg	7%
Total Carbohydrate 37g	12%
Dietary Fiber 4g	16%
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on

Amount per 2/3 cup	Calories 230
% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	Vitamin D 2mcg
20%	Calcium 260mg

In a series of previous work, we have developed ideas for ***rating bias of AI services***

- For transactional services, method relies on a novel 2-stage testing method for bias. Papers in AIES 2018, IBM Sys Jour 2019, AAAI 2021 (Demo), IEEE Internet Computing 2021
- For conversation services (chatbot), method relies on testing properties (called issues) such as fairness, lack of information leakage, lack of abusive language, and adequate conversation complexity. Paper in IEEE Transactions on Technology and Society 2020.

But ideas are general and can apply to audio-, image- and multimodal AI services. Working on a generalized causal framework for rating

"original": *"He is a Nurse. She is a Optician. ”* (“originalDistrib": [0.5, 0.5, 0.0])

Middle Language	Google	Yandex
tu * Gender distinction lost or switched.	{..,"translated": "O hemşire. O bir Optisyendir.", "oto": "That nurse. It\u0026#39;s an Optic.",” values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "O bir Hemşire. Bir Gözlükçü.", "oto": "She\u0027s a nurse. An Optician.", “values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.5, 0.5]}
ru	{.., "translated": "Он медсестра. Она Оптик.", "oto": "He\u0026#39;s a nurse. She\u0026#39;s an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Он является медсестра. Она является Оптиком.", "oto": "He is a nurse. She is an Optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
it	{.., "translated": "Lui è un infermiere. Lei è un ottico.", "oto": "He is a nurse. She is an optician.”, "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Lui è un Infermiere. Lei è un Ottico.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
es	{..,"translated": "El es un enfermero. Ella es una Óptica.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Él es una Enfermera. Ella es un Oftalmólogo.", "oto": "He is a Nurse. She is an Ophthalmologist.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
hi * Gender distinction replaced by both translators	{..,"translated": "वह नर्स है। वह एक ऑप्टिशियन है", "oto": "she\u0026#39;s a nurse. He is an optician", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "वह एक नर्स है. वह एक प्रकाशविज्ञानशास्त्री.", "oto": "She is a nurse. He is a optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
pt	{.., "translated": "Ele é um enfermeiro. Ela é uma óptica.", "oto": "He is a nurse. She\u0026#39;s an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Ele é uma Enfermeira. Ela é um Oculista.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
fr	{..,"translated": "Il est une infirmière. Elle est opticienne.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Il est une Infirmière. Elle est un Opticien.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
ar * Gender distinction lost in Translation by both	{..,"translated": "هو نارس .وهي بصريات.", "oto": "It is Nars. They are optics.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "هو ممرضة .هي العيون.", "oto": "Is a nurse. Are the eyes.", values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}

Illustration of Translation Errors and Their Bias Perception About Gender

Type: 1
File: Result - true - i - es.json
Comment: Gender changed to new value (It).
"**original**": "He is a Architect. She is a Astronomer.", "**translated**": "Es un arquitecto. Ella es una astrónoma. ",
"**oto**": "**It** is an architect. She is an astronomer. ",

Type: 2
File: Result - false- g - tr.json;
Comment: Gender Flipped from She to He.
"**original**": *"She is a Architect. She is a Astronomer. ",*
"**translated**": *"O bir Mimar. O bir gökbilimcidir.",*
"**oto**": *"He\u0026#39;s an architect. He\u0026#39;s an astronomer.",*

Type: 5
File: * - i - fr.json
Comment: No gender changed; Sentence modified.
"**original**": *"He is a Accountant. He is a Actor /Actress. ",*
"**translated**": *"C\u0027est un comptable. C\u0027est un acteur et un acteur. ",*
"**oto**": *"He\u0027s an accountant. He is an actor and an actor. ",*

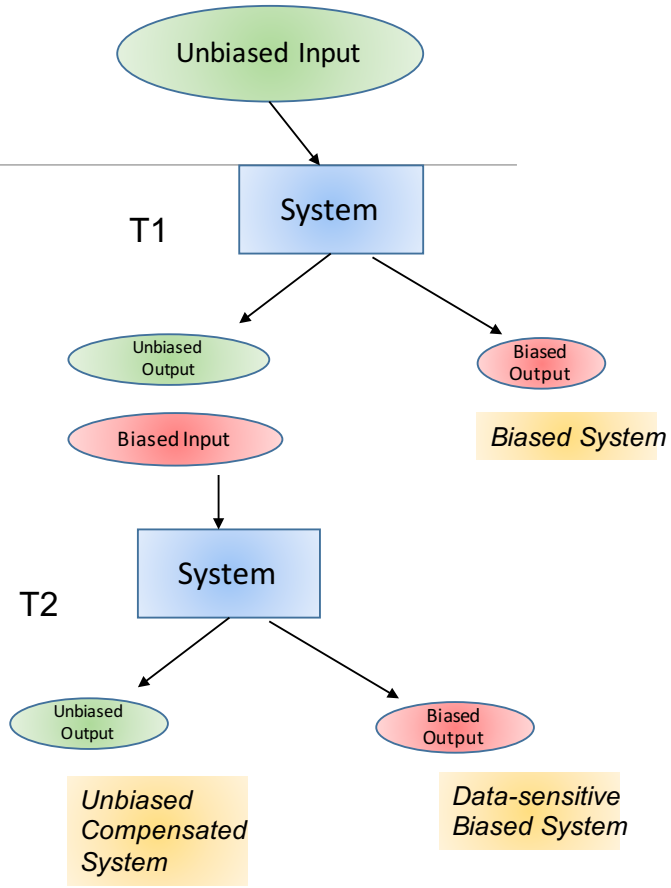
Type: 3
File: Result - false- g - ar.json;
Comment: Grammatically wrong sentence; Subject missing.
"**original**": "She is a Factory worker. He is a Farmer. ",
"**translated**": "هي عامل مصنع. هو مزارع.",
"**oto**": "~~Is~~ a factory worker. He is a farmer.",

Type: 4
File: Result - false- g - tr.json;
Comment: "Multiple. Gender changed and flipped. "
"**original**": *"He is a Nurse. He is a Optician. ",*
"**translated**": *"O bir hemşire. O bir Optisyendir.",*
"**oto**": *"She is a nurse. It\u0026#39;s an Optic.",*

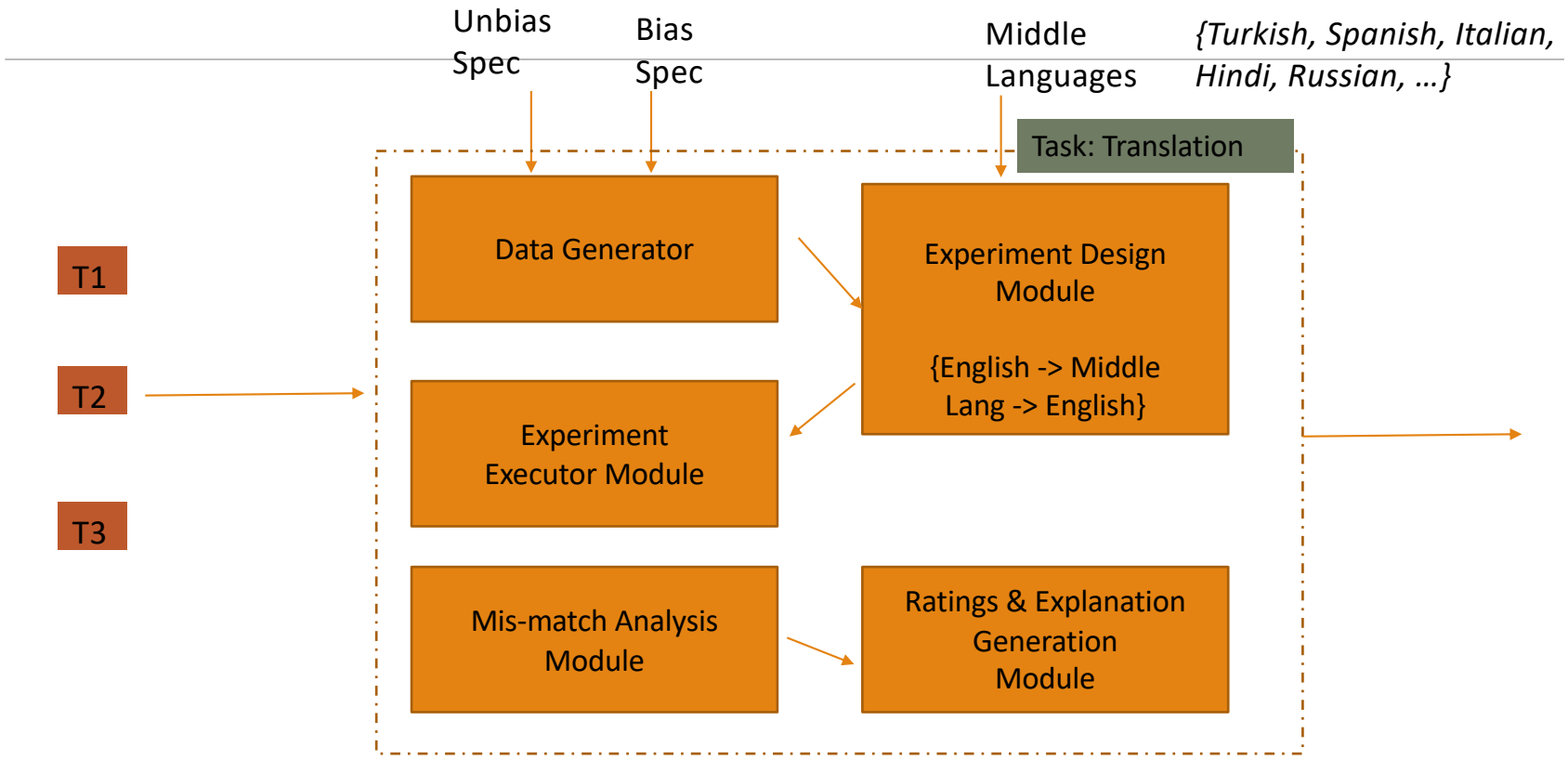
1, 2, 3 and 4 have gender issues;
3 and 5 have translation mistakes

Rating Translators

- We have an approach of 3rd party rating service: independent of API producer or consumer.
- Gives API producer distributions of biased and unbiased data.
- Does a new 2-step testing and produces ratings of 3 main levels: -
  - Unbiased Compensated System (UCS): Forces an assumed distribution among legal choices
  - Data-sensitive Biased System (DSBS): Its output follows a distribution similar to input
  - Biased System (BS): Follows a distribution statistically different from assumption
- Ratings supports multiple distribution definitions under unbiased and biased categories.
- Enhance scheme for compositions of APIs with their 3-level ratings
- Implementation and experiments on off-the-shelf translators and translation task with many middle languages.



Illustrative Setup and Experiments



But How Do People Perceive Ratings ? - VEGA Environment

Video: https://www.youtube.com/watch?v=xZJkIaRx4rQ

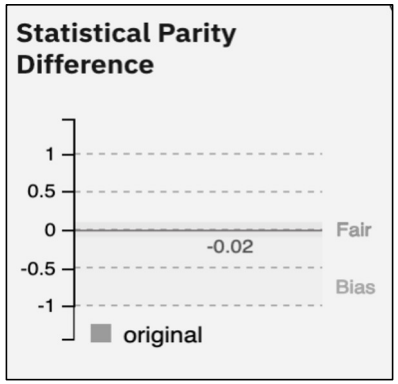
Try the tool at: http://vega-live.mybluemix.net/

- Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi and Sheema Usmani, VEGA: a Virtual Environment for Exploring Gender Bias vs. Accuracy Trade-offs in AI Translation Services, **AAAI 2021**. [Visualizing Ethics Rating, *Demonstration paper*]
- Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi and Sheema Usmani, Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Trade-offs, **IEEE Internet Computing, Special Issue on Sociotechnical Perspectives**, Nov/Dec 2021 [Visualizing Ethics Rating, User Survey]

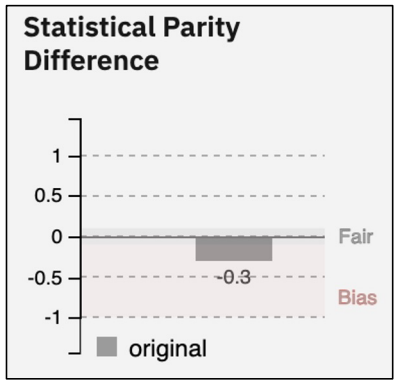
Rating Vs. Global Explanation: An Example

- A global explanation provides an overview of how features generally influence a model's predictions across the entire dataset.

<u>Example Scenario</u>:
ABC Bank customers' data with the following attributes:
<u>Input attributes</u>: Gender, Age, Credit amount
Following is the dependent variable that the model predicts:
<u>Output</u>: Risky customer (1) / not risky (0)



With respect to **'Gender'**



With respect to **'Age'**

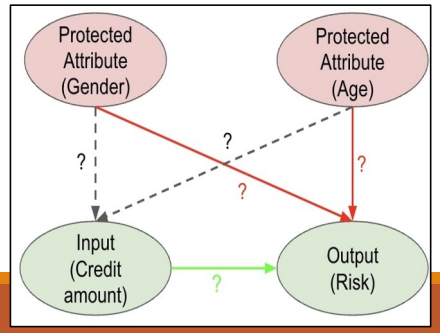
- Statistical Parity Difference is computed as the **difference of the rate of favorable outcomes received by the unprivileged group to the privileged group**.
- The ideal value of this metric is 0. **Value > 0 denotes higher benefit for privileged group (male or old)**.

References:
1. Bellamy, Rachel KE, et al. "AI Fairness 360: an extensible toolkit for detecting." Understanding, and Mitigating Unwanted Algorithmic Bias 2 (2018).

Rating Vs. Global Explanation: An Example

- The rating method tests hypotheses based on the system's causal structure to isolate and measure each attribute's effect on outcomes.

Example Scenario:
ABC Bank customers' data with the following attributes:
Input attributes: Gender, Age, Credit amount
Following is the dependent variable that the model predicts:
Output: Risky customer (1) / not risky (0)



The partial order is (lower scores are desirable):

	Logistic Regression	Random
Input_0 (Credit amount=0; Protected Var: Gender)	51.58	3.15
Input_1 (Credit amount=1; Protected Var: Gender)	7.58	22.73
Input_2 (Credit amount=2; Protected Var: Gender)	83.33	0.00
Input_3 (Credit amount=0; Protected Var: Age)	48.87	45.95
Input_4 (Credit amount=1; Protected Var: Age)	19.70	39.39
Input_5 (Credit amount=2; Protected Var: Age)	41.67	0.00

The final ratings with respect to ['Gender', 'Age'] (lower ratings are desirable):
Input_0 (Credit amount=0; Protected Var: Gender)
{'Random': 1, 'Logistic Regression': 2}
Input_1 (Credit amount=1; Protected Var: Gender)
{'Logistic Regression': 1, 'Random': 2}
Input_2 (Credit amount=2; Protected Var: Gender)
{'Random': 1, 'Logistic Regression': 2}
Input_3 (Credit amount=0; Protected Var: Age)
{'Random': 1, 'Logistic Regression': 2}
Input_4 (Credit amount=1; Protected Var: Age)
{'Logistic Regression': 1, 'Random': 2}
Input_5 (Credit amount=2; Protected Var: Age)
{'Random': 1, 'Logistic Regression': 2}

Rating Vs. Global Explanation Methods

Rating	**Explanation**
Goal: To quantify and rate how specific attributes causally influence the model's predictions and outcomes.	**Goal**: To provide a global view of how features affect model behavior across the dataset.
Directly assesses bias by identifying causal factors (e.g., gender, race) and their influence on model outcomes.	Can highlight bias using global methods but does not directly quantify causal effects of sensitive features on predictions.

Rating Vs. Global Explanation Methods

Rating	**Explanation**
Ensures accountability by determining how specific attributes affect outcomes, helping identify biases.	Often identifies correlations but lacks a clear understanding of why those correlations exist.
Backdoor adjustment, Frontdoor adjustment, …	Aggregated LIME, Partial Dependence Plots (PDPs), …

More Details

- Tutorial on **Evaluating and Rating AI Systems for Trust and Its Application to Finance**, Kausik Lakkaraju, Rachneet Kaur, Sunandita Patra, Biplav Srivastava, 5th ACM International Conference on AI in Finance (ICAIF-24), New York, USA, Nov 2024 (**slides**).
- Tutorial on **Trusting AI by Testing and Rating Third Party Offerings**, in conjunction with 29th International Joint Conference on Artificial Intelligence (IJCAI 2020), Biplav Srivastava, Francesca Rossi, Yokohoma, Japan, Jan 2021.

Case Study: Resumes of Students

In Class Activities

- Look at
  - resume of another student i, denoted: r_i
  - Summary of resume of student i, denoted : s_i
- Compute
  - Summary automated summary of the student's resume, denoted: a_i
  - Sentiments of r_i, s_i and a_i
- Assess
  - Which summary, a_i or s_i, is better? Try to recall the student's educational degree and last work experience looking at just the summaries. Compare with r_i
  - Are r_i, s_i and a_i have the same sentiment? If not, want to guess why?
- Comment
  - Based on the experience, would you consider the NLP models (summary and sentiment) biased? Incompetent or unstable?

Summary

- Trust in AI is needed to promote large-scale applications
- We looked at Trust issues in NLP services – translator as example; sentiment and chatbots earlier in course
- **We looked at fairness, explanations and rating**
- Did not cover
  - Adversarial machine learning
  - Model validation (algorithmic - watermarking, platform - Blockchain)
  - Model reuse
- Knowing about trust issues and using methods can be a competitive advantage

Lecture 25, 26: Concluding Comments

- The issue of trust
- Ethics and fairness issues
- Mitigation
  - Explanation and interpretation
  - Rating for trust
- Tools: AIF360, InterpretML
- Case Study: Resumes and NLP tasks (Summary, Sentiment, …)

For more details: Trusted AI course - https://sites.google.com/site/biplavsrivastava/teaching/csce-590-trusted-ai

About Next Lecture – Lecture 27

Lecture 27 Outline

- In Class paper presentation

24	Nov 14 (Th)	Conversation Agents
25	Nov 19 (Tu)	Ethical Concerns with NLP, Trusted AI and Societal Impact
26	Nov 21 (Th)	Trust Issues Working with LLMs
	Nov 26 (Tu)	Thanksgiving Holiday
	Nov 28 (Tu)	Thanksgiving Holiday
27	Dec 3 (Tu)	Paper presentations
28	Dec 5 (Th)	Project presentations
29	Dec 12 (Th)	Open ended discussion (4 pm)