CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 25: Bias and Trust Issues in AI

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

13RD APR, 2021

Credits: My collaborators on
AI Ethics, especially Francesca Rossi

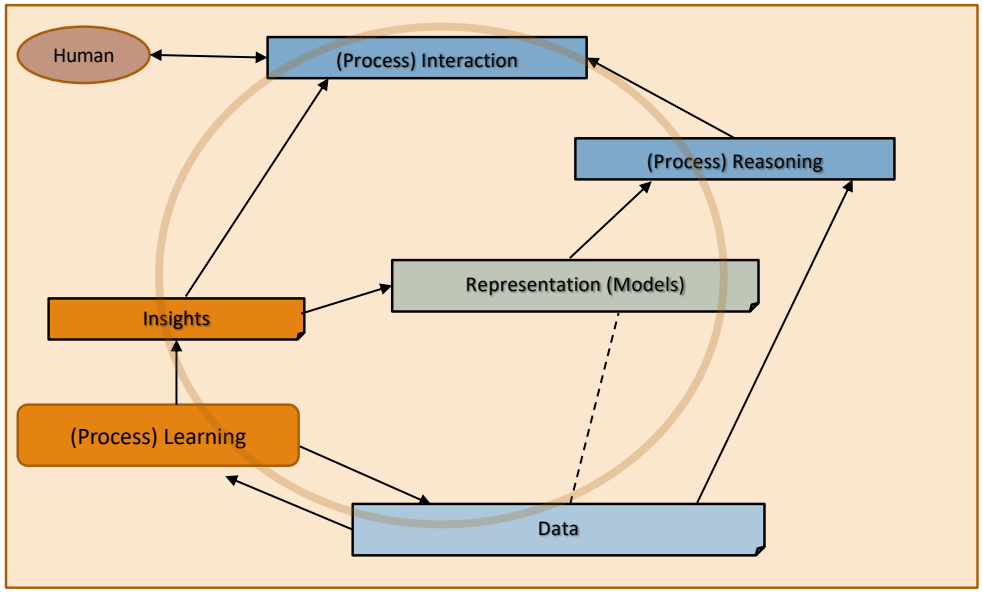
Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 25

- Introduction Segment
  - Recap/ Discussion of Lecture 24
  - Quizzes

- Main Segment
  - Problem of Trust
    - Framing the problem
    - Common reasons
  - Detecting of Bias
  - Remediation Methods

- Concluding Segment
  - About Next Lecture – Lecture 26
  - Ask me anything



Introduction Segment

Recap of Lecture 24

- Sentiment models
  - Problem of bias
  - How to use in business applications

- Visualization of documents – useful in practice

Quizzes

- Quiz 3
  - Many students made mistake on Q3 – Bayesian Network problem
  - Explain the fundamentals again

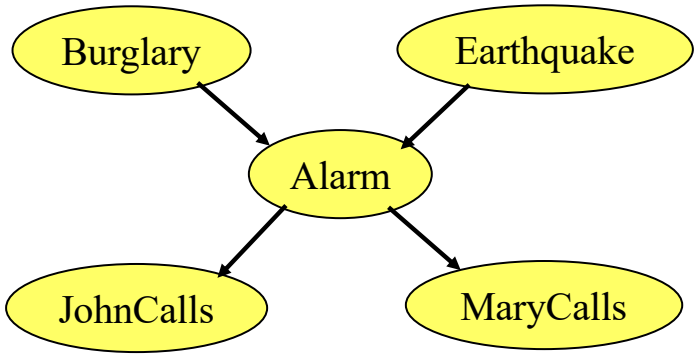
- Quiz 4
  - Repeat Q3 of Quiz 3 as Q1 in Quiz 4
  - Perform text tutorial using Spacy

Bayesian Networks

Directed Acyclic Graph (DAG)

- Nodes are random variables
- Edges indicate causal influences

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(X_i))$$

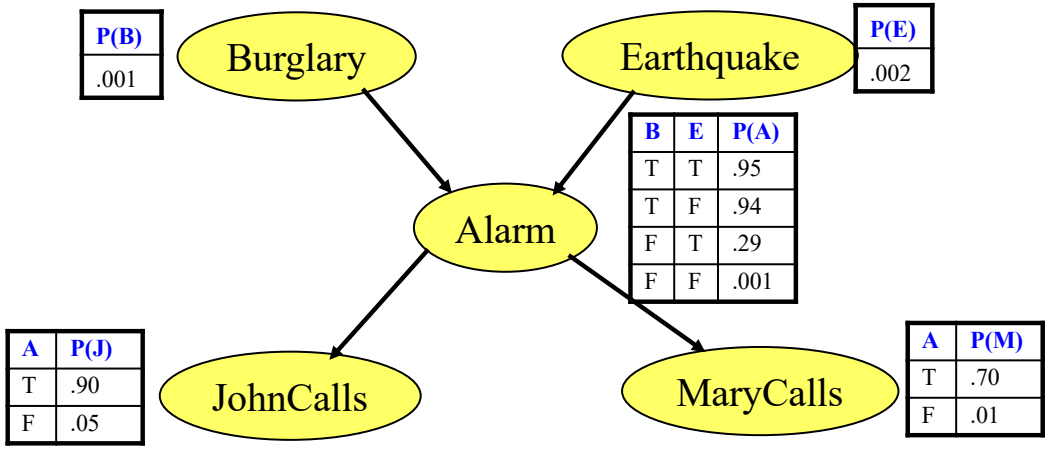


From Class 17

Slide adapted from: Ray Mooney's AI Course
Example from Russell & Norvig AI book

Conditional Probability Tables

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).
- Roots (sources) of the DAG that have no parents are given prior probabilities.



$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(X_i))$$

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

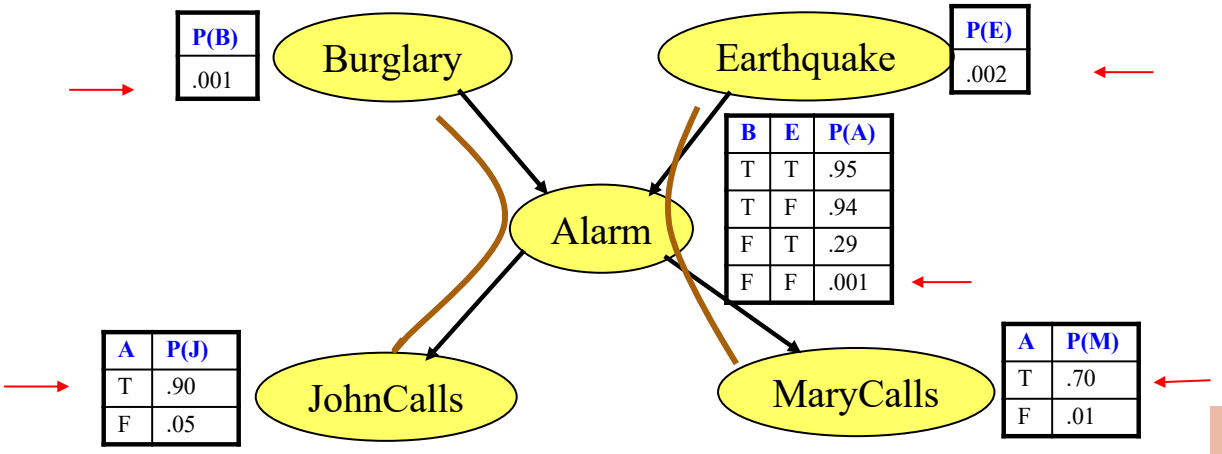
Slide adapted from: Ray Mooney's AI Course
Example from Russell & Norvig AI book

Conditional Probability Tables

- Probability that John and Mary call on hearing an alarm but there is no burglary or earthquake

= P (J |A) * P (M|A) * P(A| !B ∧ !E) * (!B) * (!E)

= 0.9 * 0.7 * 0.001 * 0.999 * 0.998 = 0.00062



Slide adapted from: Ray Mooney's AI Course
Example from Russell & Norvig AI book

Main Segment

Trust Problem in AI

Many Different AI scenarios

- Several data types
  - Text, images, audio, structured data
- Stand-alone or embedded in a larger system
- Autonomous decision systems vs decision support systems
- Interacting with one user or with multiple humans

- Checking/building trust while
  - ***Developing*** an AI system/service – (AI fairness 360, section 3)
    - Checking/mitigating bias in data or models
  - ***Using*** an AI system/service – (rating bias in AI systems, section 4)
    - No access to training data
    - No modification to the model

Online Translation

Example: Loan Application

- Bias in training data
  - Assume that the training examples show rejection for all women applications and acceptance for all men applications
  - AI will identify the correlation (gender-acceptability) and will use it in new applications
- Not just in training data
  - List of priority motivations for loan applications
    - Buying a house, paying school fees, paying legal fees, …
  - If one of them is omitted, the relevant community will be penalized

Statement	Score
I'm a sikh	+0.3
I'm a christian	+0.1
I'm a jew	-0.2
I'm a homosexual	-0.5
I'm queer	-0.1
I'm straight	+0.1

Sentiment analysis (2017)

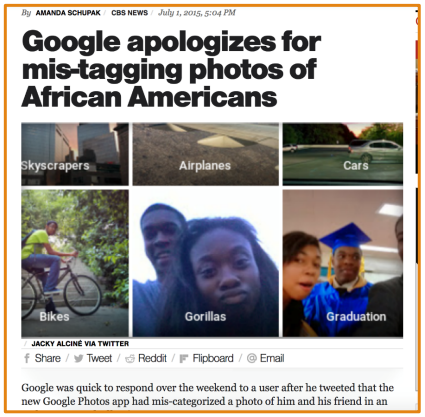


Image interpretation (2015)

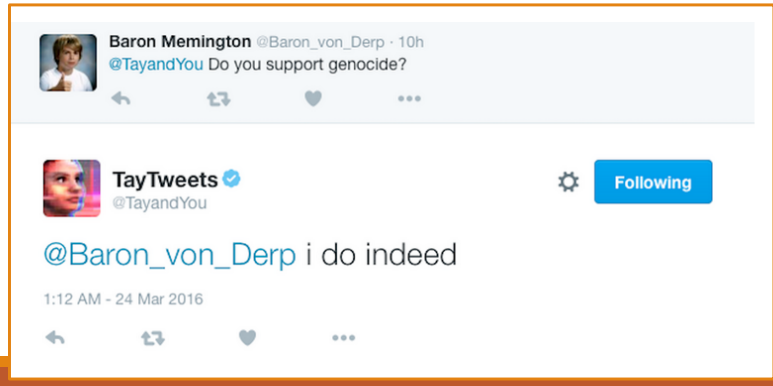
*"Amazon's system **taught itself that male candidates were preferable**. It **penalized resumes that included the word "women's," as in "women's chess club captain."** And it downgraded graduates of two all-women's colleges,*

Hiring (2018)

*"The formula was particularly likely to **falsely flag black defendants as future criminals**, … at almost **twice the rate** as white defendants."*

"White defendants were mislabeled as low risk more often than black defendants."

Recidivism assessment (2016)



Chatbots (2016)

Gender Recognition From Faces (2017)

Classifier	Metric	DF	DM	LF	LM
MSFT	PPV(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	TPR(%)	**100**	84.2	**100**	**100**
	FPR(%)	**15.8**	0.0	0.0	0.0
Face++	PPV(%)	64.0	99.5	**100**	**100**
	Error Rate(%)	**36.0**	0.5	0.0	0.0
	TPR(%)	99.0	77.8	**100**	96.9
	FPR(%)	**22.2**	1.03	3.08	0.0
IBM	PPV(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	TPR(%)	90.4	78.0	96.4	**100**
	FPR(%)	**22.0**	9.7	0.0	3.6

Instability of AI is Well Recorded

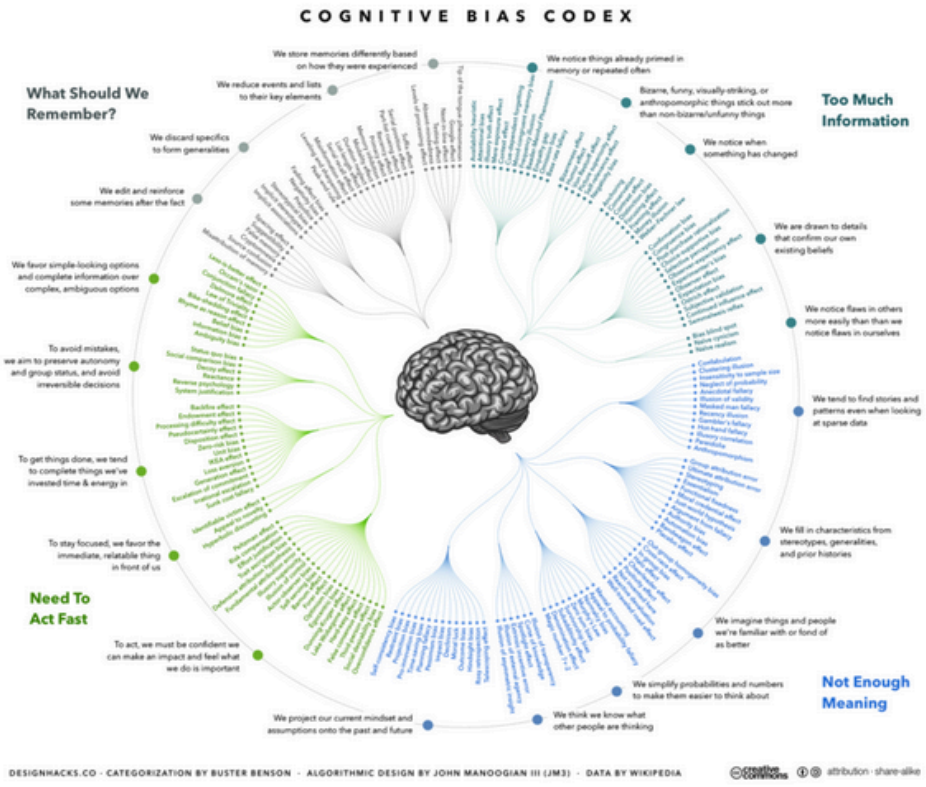
[Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]

[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

Trust: What Do We Want

Context: Human Bias and AI



- Bias: prejudice for or against something
- As a consequence of bias, one could behave unfairly to certain groups compared to others
- Why should AI be biased?
  - Trained on data provided by people, and people are biased
  - Designed and developed by human beings by human

AI: A Matter of Trust

- AI Services are known to have trust issues. This can be for transactional, state-less, services, like translators; or interactive, stateful services, like conversation agents, a.k.a, chatbots
  - We Teach A.I. Systems Everything, Including Our Biases, https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html
  - "This is probably a billion-dollar industry," - Primer's chief executive, Sean Gourley.
- Trust can have many dimensional issues: bias, hate speech, information leaking, ..
  - We will use term **trust** as a general term for such issues
  - Examples: **Western perspective**: racial, sexual, religious; Regional, e.g., **Indian perspective**: color, caste, regional
- Having bias in computational systems (devices, APIs) is a major hurdle for technology adoption. Increasing importance as apps become intelligent and interact with people

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

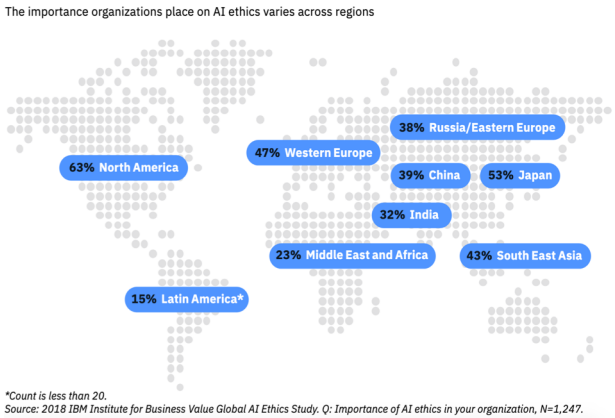
Building Trust is Important

Without trust there will not be full adoption, and therefore we will miss the huge positive effect of AI

1. Trust in the AI technology
2. Trust in those who produce AI
3. Trust in those who regulate AI

IBM IBV study on >1000 C-level executives and policy makers

- 80% say that concerns about trust, privacy, and transparency are a barrier to AI adoption
- 80% consider trusted training data important



IBV study on "Advancing AI ethics beyond compliance", 2020: https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

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

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

IJCAI 2020 TUTORIAL: TRUSTING AI BY TESTING AND RATING 3RD PARTY OFFERINGS - B. SRIVASTAVA AND F. ROSSI

Main AI Ethics Issues

DATA GOVERNANCE AND PRIVACY

FAIRNESS AND INCLUSION

HUMAN AND MORAL AGENCY

VALUE ALIGNMENT

ACCOUNTABILITY

TRANSPARENCY AND EXPLAINABILITY

TECHNOLOGY MISUSE

IJCAI 2020 TUTORIAL: TRUSTING AI BY TESTING AND RATING 3RD PARTY OFFERINGS - B. SRIVASTAVA AND F. ROSSI

"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 प्रकाशविज्ञानशास्री.", "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

Problem of Bias with Sentiments

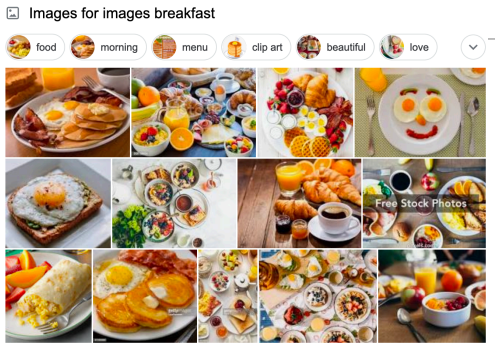
Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems, Svetlana KiritchenkoandSaif M. Mohammad, https://www.aclweb.org/anthology/S18-2.pdf

- 219 systems tested for bias
- For 4 emotions test, only **25% submission** (12/46) **showed no statistically significant score difference** (i.e., unbiased)
- 75% to 86% of the submissions consistently marked sentences of one gender higher than another (i.e., biased)
- For race, the number of submissions with **no statistically significant score difference** is **11% to 24%**). Lower than gender [See paper] (i.e., *more* race biased than gender)

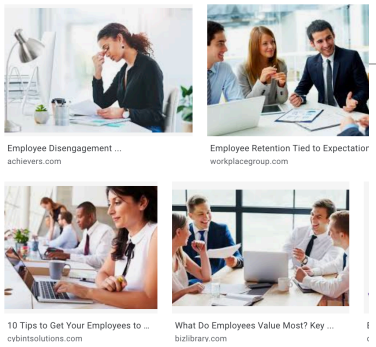
Task / Bias group	#Subm.	Avg. score diff. F↑–M↓	Avg. score diff. F↓–M↑
Anger intensity prediction			
F=M not significant	12	0.042	-0.043
F↑–M↓ significant	21	0.019	-0.014
F↓–M↑ significant	13	0.010	-0.017
All	46	0.023	-0.023
Fear intensity prediction			
F=M not significant	11	0.041	-0.043
F↑–M↓ significant	12	0.019	-0.014
F↓–M↑ significant	23	0.015	-0.025
All	46	0.022	-0.026
Joy intensity prediction			
F=M not significant	12	0.048	-0.049
F↑–M↓ significant	25	0.024	-0.016
F↓–M↑ significant	8	0.008	-0.016
All	45	0.027	-0.025
Sadness intensity prediction			
F=M not significant	12	0.040	-0.042
F↑–M↓ significant	18	0.023	-0.016
F↓–M↑ significant	16	0.011	-0.018
All	46	0.023	-0.023
Valence prediction			
F=M not significant	5	0.020	-0.018
F↑–M↓ significant	22	0.023	-0.013
F↓–M↑ significant	9	0.012	-0.014
All	36	0.020	-0.014

T-test: The null hypothesis that the true mean difference between the paired samples is zero can be rejected if the calculated p-value falls below 0.05/438.

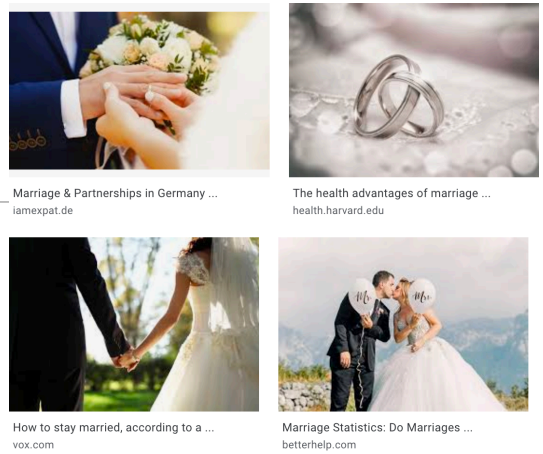
Indian Perspective Missing by Default



Breakfast items searched on Google

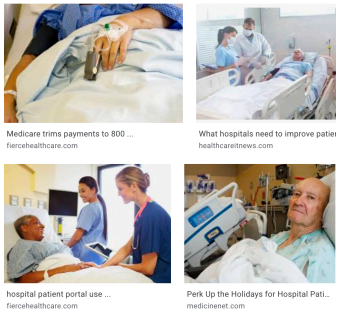


Employees



Marriage

Waiting Customers



Hospital Patients

Indian Perspective Matters for AI

Breakfast items searched on Google

Waiting Customers

How does this impact ?

- Training data
- Preferences and constraints
- Inductive bias (i.e., implicit defaults in the algorithms)
- Integration of solution to wider ecosystem

Bias Considerations in India are Different

Sources of bias perception

- Age: students, employees
- Language: English v/s non-English
- Clothes
- Food
- Religion
- …

May impact AI applications in

- Finance/ banking/ insurance
- Health
- Tourism
- Infrastructure: Transportation
- …

Defining and Detecting Bias

German Credit Data

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

1. Credit amount (numerical); 2. Credit duration (numerical); 3.Credit purpose (categorical); 4. Status of existing checking account(categorical); 5. Status of savings accounts and bonds (categori-cal); 6. Number of existing credits (numerical); 7. Credit history(categorical); 8. Installment plans (categorical); 9. Installment rate(numerical); 10. Property (categorical); 11. Residence (categorical);12. Period of present residency (numerical); 13. Telephone (binary);14. Employment (categorical); 15. Employment length (categorical);16. Personal status and gender (categorical); 17. Age (numerical);18. Foreign worker (binary); 19. Dependents (numerical); 20. Otherdebtors (categorical); 21. Credit score (binary)

- Dataset that classifies people's credit risk based on their individual attributes such as Age, Income, Gender, etc.
  1000 rows of data, each with 20 attributes to check bias against
- 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 class a **customer as good when they are bad**, than it is to class a **customer as bad when they are good**.

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

A11: less than 0 balance; 6: six months; A34: critical account; A43: has radio/ television ..

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

VERMA, S., AND RUBIN, J. 2018. FAIRNESS DEFINITIONS EXPLAINED. IN PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS, FAIRWARE '18, 1–7. NEW YORK, NY, USA: ASSOCIATION FOR COMPUTING MACHINERY, HTTPS://WWW.ECE.UBC.CA/~MJULIA/PUBLICATIONS/FAIRNESS_DEFINITIONS_EXPLAINED_2018.PDF

What Does a Fairness Definition Mean?

Example:

- Group Fairness: equal probability for male and female applicants to have good predicted credit score: P(d=1|G=m) = P(d= 1|G=f)
- Trained a logistic regression focused on accuracy
- The probability to have a good predicted credit score for married / divorced male and female applicants is 0.81 and 0.75, respectively

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

VERMA, S., AND RUBIN, J. 2018. FAIRNESS DEFINITIONS EXPLAINED. IN PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS, FAIRWARE '18, 1–7. NEW YORK, NY, USA: ASSOCIATION FOR COMPUTING MACHINERY, HTTPS://WWW.ECE.UBC.CA/~MJULIA/PUBLICATIONS/FAIRNESS_DEFINITIONS_EXPLAINED_2018.PDF

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

A Step Towards Fairness

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)

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

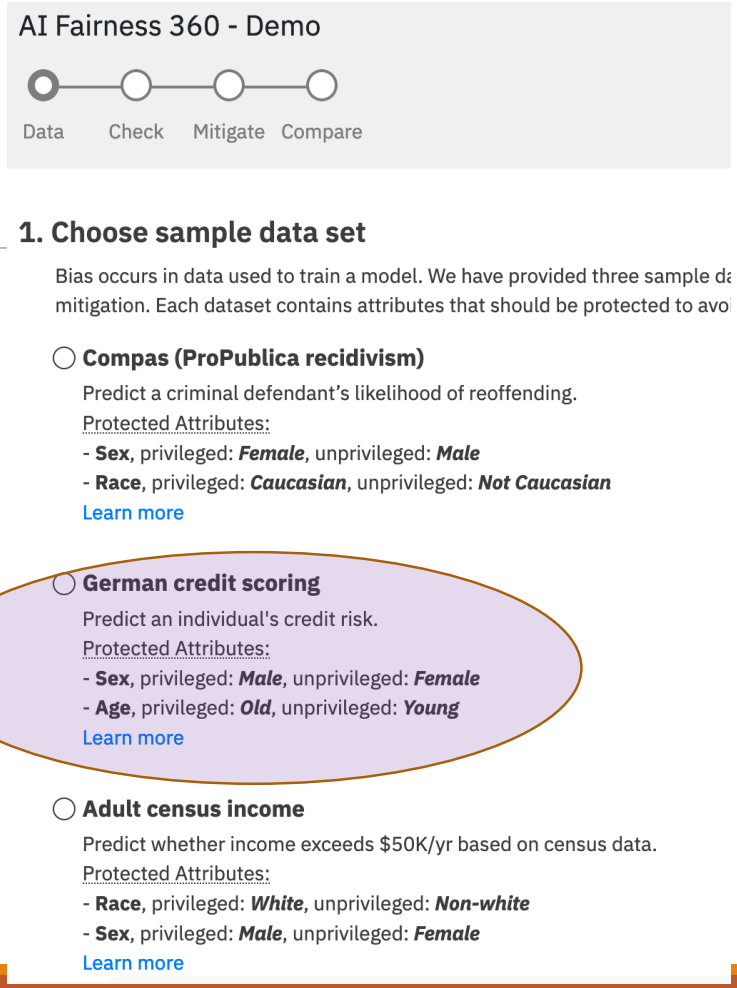
Detecting Bias

Many Tools to Experiment with AI Fairness

- Aequitas : University of Chicago
- AI Fairness 360 : IBM Research
- Audit-AI : pymetric.ai
- FairML : Julius Adebayo
- Fairness Comparison : Haverford College, Univ. of Arizona, Univ. of Utah
- Fairness Measures :
- FairTest : Multiple Universities
- Themis™ : University of Massachusetts, Amherst
- Themis-ML : arena.io
- What-If tool : Google
- … Accenture, Facebook, Microsoft …

AI Fairness 360

- AIF360 Tool and Demo: 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



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)

Remediation Methods

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: *Age*

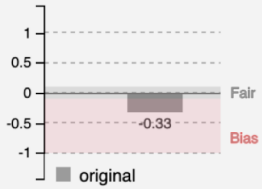
Protected Attribute: Age

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

Accuracy with no mitigation applied is 76%

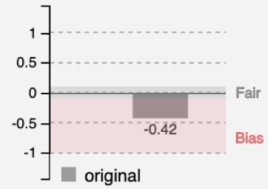
With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

Statistical Parity Difference



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

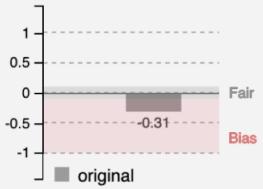
Equal Opportunity Difference



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

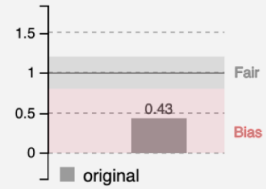
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



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

Disparate Impact

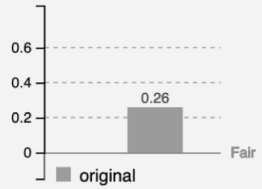


Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

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: *Gender*

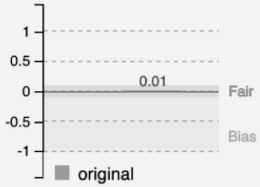
Protected Attribute: Sex

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

Accuracy with no mitigation applied is 76%

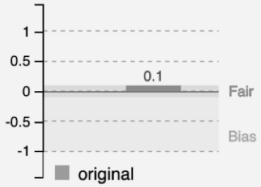
With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics

Statistical Parity Difference



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

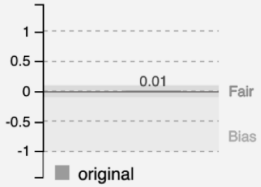
Equal Opportunity Difference



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

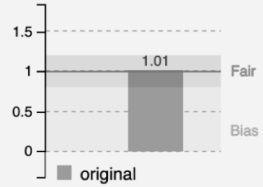
A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.

Average Odds Difference



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

Disparate Impact

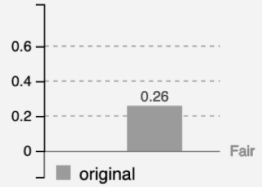


Ideal Value: 1.0
Fairness Range: (0.80, 1.20)

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

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%

Building Trust in AI Systems: Transparency Through Documentation

Details: https://sites.google.com/site/biplavsrivastava/research-1/trustedai

Transparency Through Documentation of Rating

Documentation about

- Outcome (e.g., Nutrition label, Electronic DataSheet, Factsheet)
- Process (e.g., SEI Capability Maturity Model, ISO 9001)

Documentation by

- Producer (e.g., Nutrition label)
- Consumer (e.g., Yelp rating)
- Independent 3rd Party (e.g., JD Powers, NHTSA car crash)

Reference: AboutML Project at PAI - https://www.partnershiponai.org/about-ml-get-involved/#read

Rating AI Systems from 3rd Party Perspective

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

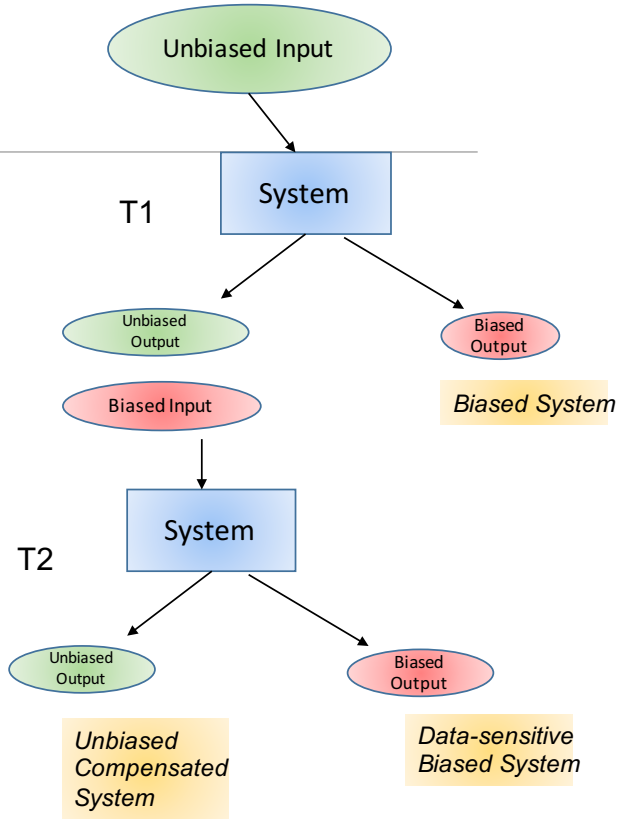
In a series of previous work, we have developed ideas for ***rating bias of AI services***

- For transactional services, method relies on a novel 2-stage testing method for bias. Papers in AIES 2018, IBM Sys Jour 2019 and AAAI 2021 (Demo)
- 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.

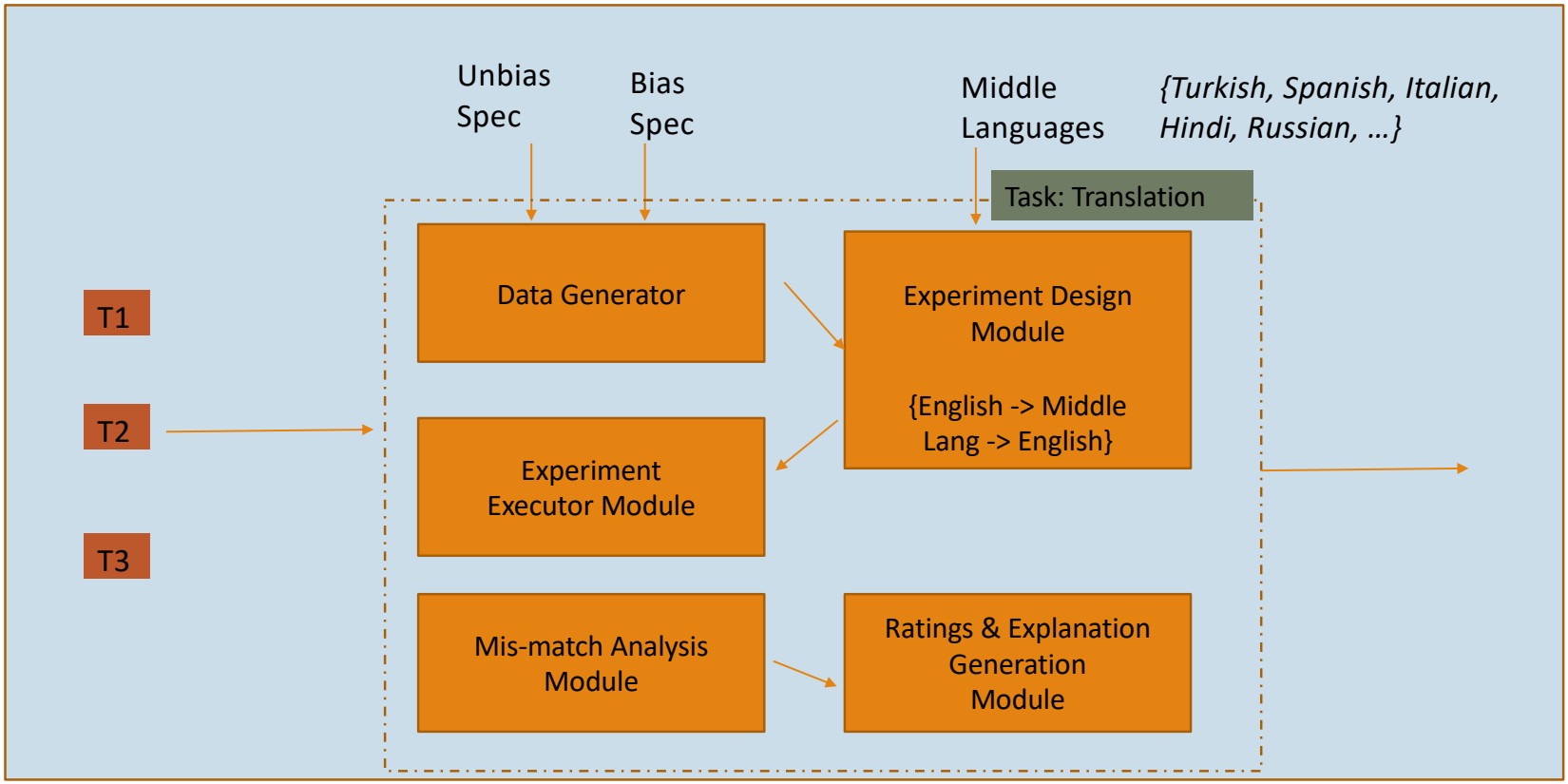
But ideas are general and can apply to audio-, image- and multimodal AI services.

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.

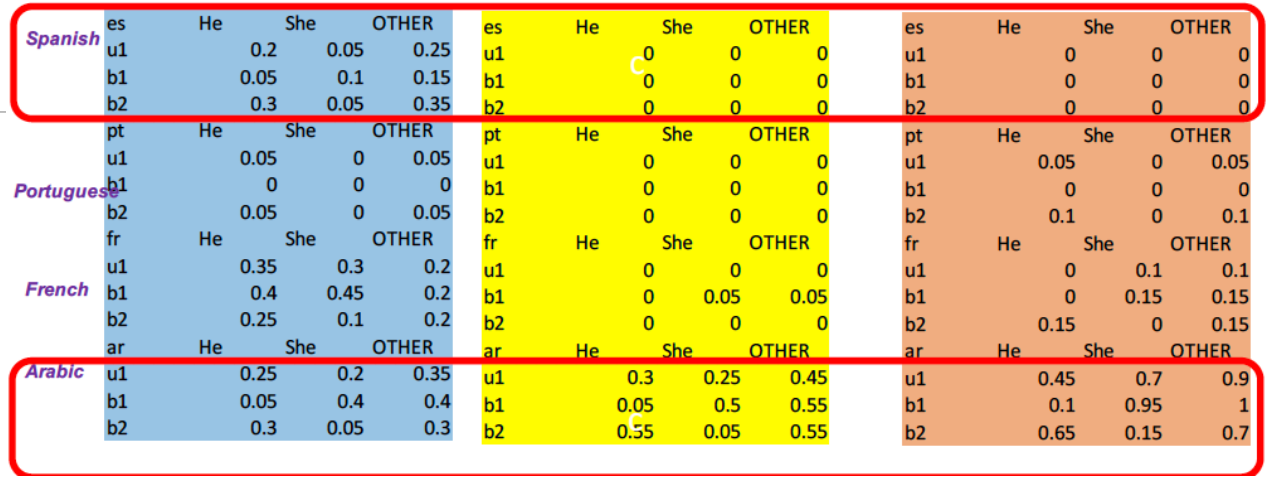


Illustrative Setup and Experiments



Rating Translators – Some Results

Intermediate Step



Spanish												
es	He	She	OTHER	es	He	She	OTHER	es	He	She	OTHER	
u1	0.2	0.05	0.25	u1	0	0	0	u1	0	0	0	
b1	0.05	0.1	0.15	b1	0	0	0	b1	0	0	0	
b2	0.3	0.05	0.35	b2	0	0	0	b2	0	0	0	
Portuguese												
pt	He	She	OTHER	pt	He	She	OTHER	pt	He	She	OTHER	
u1	0.05	0	0.05	u1	0	0	0	u1	0.05	0	0.05	
b1	0	0	0	b1	0	0	0	b1	0	0	0	
b2	0.05	0	0.05	b2	0	0	0	b2	0.1	0	0.1	
French												
fr	He	She	OTHER	fr	He	She	OTHER	fr	He	She	OTHER	
u1	0.35	0.3	0.2	u1	0	0	0	u1	0	0.1	0.1	
b1	0.4	0.45	0.2	b1	0	0.05	0.05	b1	0	0.15	0.15	
b2	0.25	0.1	0.2	b2	0	0	0	b2	0.15	0	0.15	
Arabic												
ar	He	She	OTHER	ar	He	She	OTHER	ar	He	She	OTHER	
u1	0.25	0.2	0.35	u1	0.3	0.25	0.45	u1	0.45	0.7	0.9	
b1	0.05	0.4	0.4	b1	0.05	0.5	0.55	b1	0.1	0.95	1	
b2	0.3	0.05	0.3	b2	0.55	0.05	0.55	b2	0.65	0.15	0.7	

Final Ratings

No.	M_i	Rating (T_1)	Rating (T_2)	Rating (T_3)
1.	es	DSBS	DSBS	DSBS
2.	pt	DSBS	DSBS	DSBS
3.	fr	DSBS	DSBS	UCS
4.	ar	DSBS	DSBS	DSBS
	Overall	DSBS	DSBS	DSBS

Lecture 25: Concluding Comments

- Be aware of problem of trust with AI systems
- We covered
  - Framing the problem, common reasons
  - Detection methods
  - Remediation methods
- Will be offering a class in Fall 2021 on “Trusted AI”
- Take a practical approach
  - Build systems
  - Engage with stakeholders to improve

Concluding Segment

Upcoming Classes

Upcoming Classes

21	**Mar 25 (Th)**	**Review: project presentations, Discussion**	
	Mar 30 (Tu)	Wellness Holiday	
22	Apr 1 (Th)	Text: Text Summarization	
23	Apr 6 (Tu)	Text: Representation, Sentiment	
24	Apr 8 (Th)	Text: Sentiment, Visualization	
25	Apr 13 (Tu)	Advanced: Bias and Trust Issues	Quiz 4
26	Apr 15 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
27	Apr 20 (Tu)	**Invited Guest** – Javid Huseynov – Case Study: Finance	
28	Apr 22 (Th)	Project presentations	
	Apr 27 (Tu)	Reading day	Reading day
29	Apr 29 (Th)	Project presentations	Final assignment given (undergrad)
30	May 4 (Tu)	Course Recap	Final assignment due (undergrad), Paper summary due (grad)

About Next Lecture – Lecture 26

Lecture 26: Graduate Student Paper Presentations

- Part of their final requirements
- 5 students
- Everyone is requested to attend, ask questions and learn
  - For undergrads, question may be posed in final based on these papers

Paper Presentations – Graduate Students

- Select a paper appearing at a top-AI or data conference (AAAI, IJCAI, NeurIPS, SIGMOD, WWW, ICML, VLDB, …) during 2019-2021
- Present in class for 10 + 5 minutes of Q/A
- Things to cover
  - **Summary**: problem, solution, related work, contributions
  - **Opinion**: What you liked or did not like
- Dates
  - In-class presentation on April 15, 2021
  - 1-page written report on May 4, 2021