AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 581: Introduction to Trusted AI

Lectures 11 and 12: (Supervised) ML – Trust, Mitigation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

18TH AND 20TH FEB, 2025

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

Credits: **Copyrights of all material reused acknowledged**

Organization of Lectures 11, 12

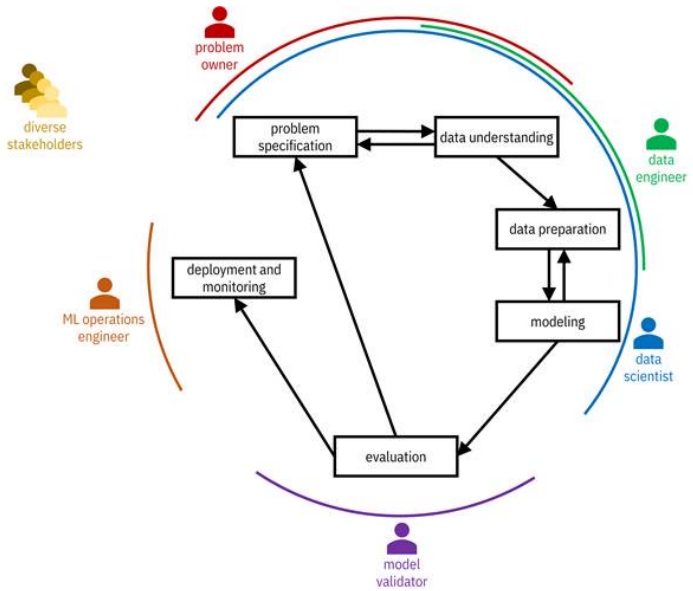
- Introduction Section
  - Recap from Week 4 (Lectures 7 and 8)
  - Announcements and News
- Main Section
  - L11: Trust issues in (Supervised) ML
  - L12: Trust/ Risk mitigation
- Concluding Section
  - About next week – Lectures 13, 14
  - Ask me anything

Introduction Section

Recap from Week 5 (Lectures 9, 10)

- We looked at
  - Quiz 1
  - Trust issues in (Supervised) ML

Recap: ML Pipelines

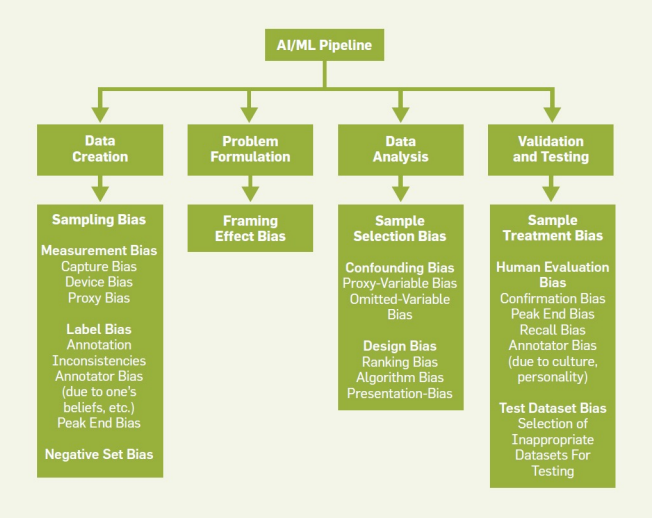


Highly Simplified View

Image Credit: Trustworthy Machine Learning, Kush Varshney

Recap of Lecture 10

Looked at trust issues



- Trust Issues – Reading Material – Bias in AI
  - Illustrative problem: German credit data
  - Issues possible during data creation, problem formulation, data analysis, validation

- Trust Issues – Reading Material - Turing Institute report on AI for Covid in UK
  - Illustrative problem : vaccinate distribution
  - Issues in robust and timely data, and communication

AI News

- Blog on Crawl-Walk-Run, as applied to an AI project
  - https://www.linkedin.com/pulse/crawl-walk-run-approach-ai-based-real-world-problem-biplav-srivastava-pxsre/

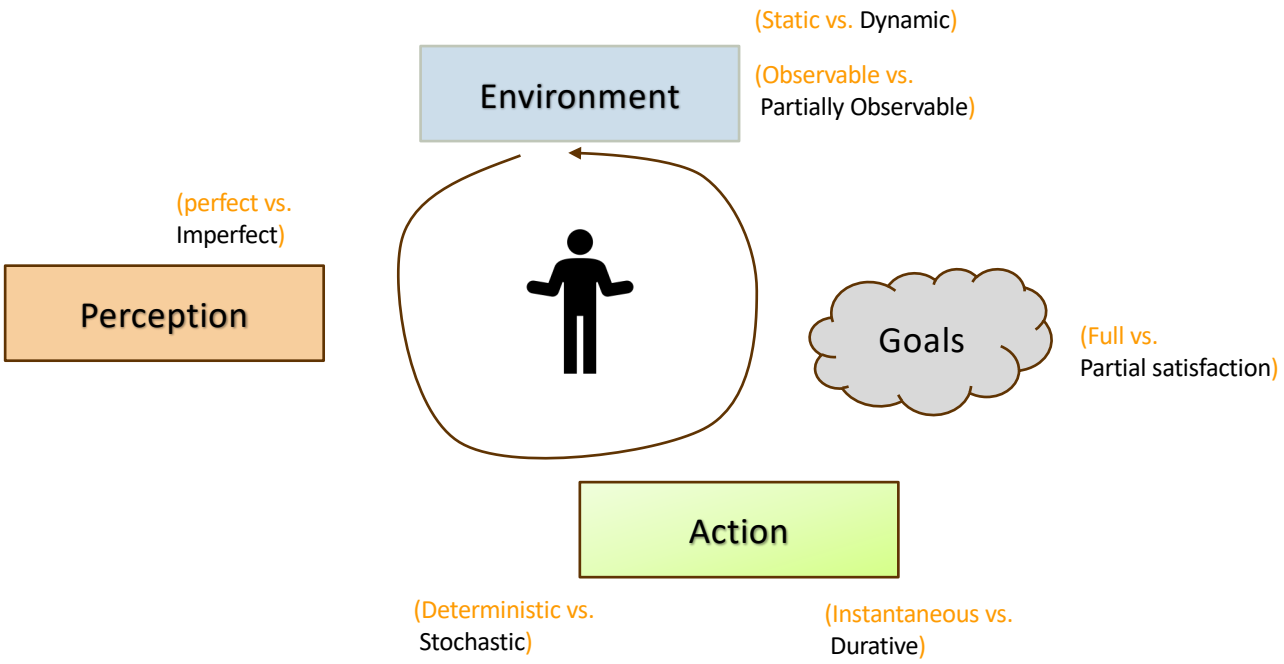
Announcement: Change to Student Assessment

A = [920-1000]
B+ = [870-919]
B = [820-869]
C+ = [770-819]
C = [720-769]
D+ = [670-719]
D = [600-669]
F = [0-599]

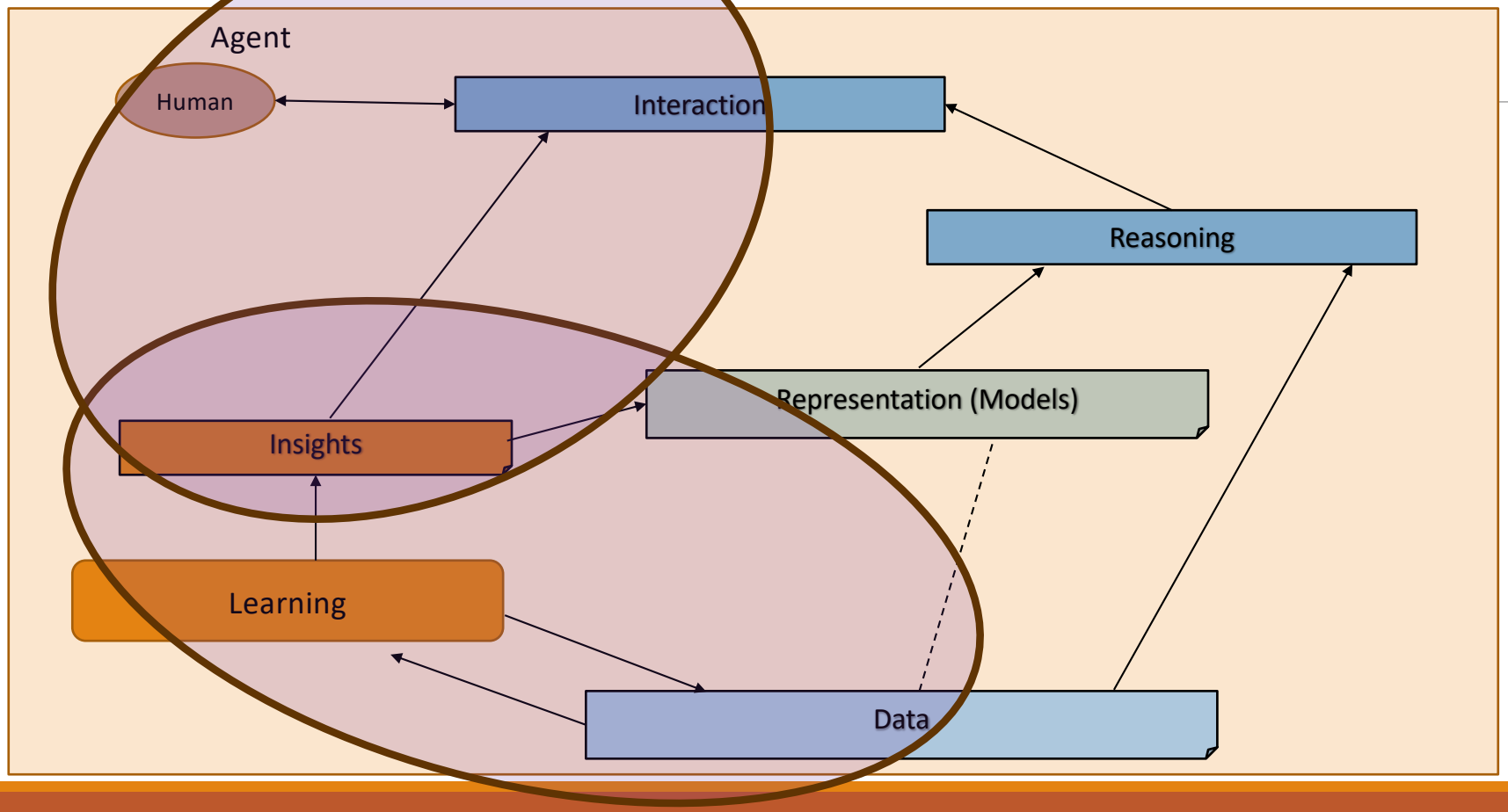
Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification, **Large Language Models**
- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

Main Section

ML Pipelines and Trust-Based Intervention Considerations

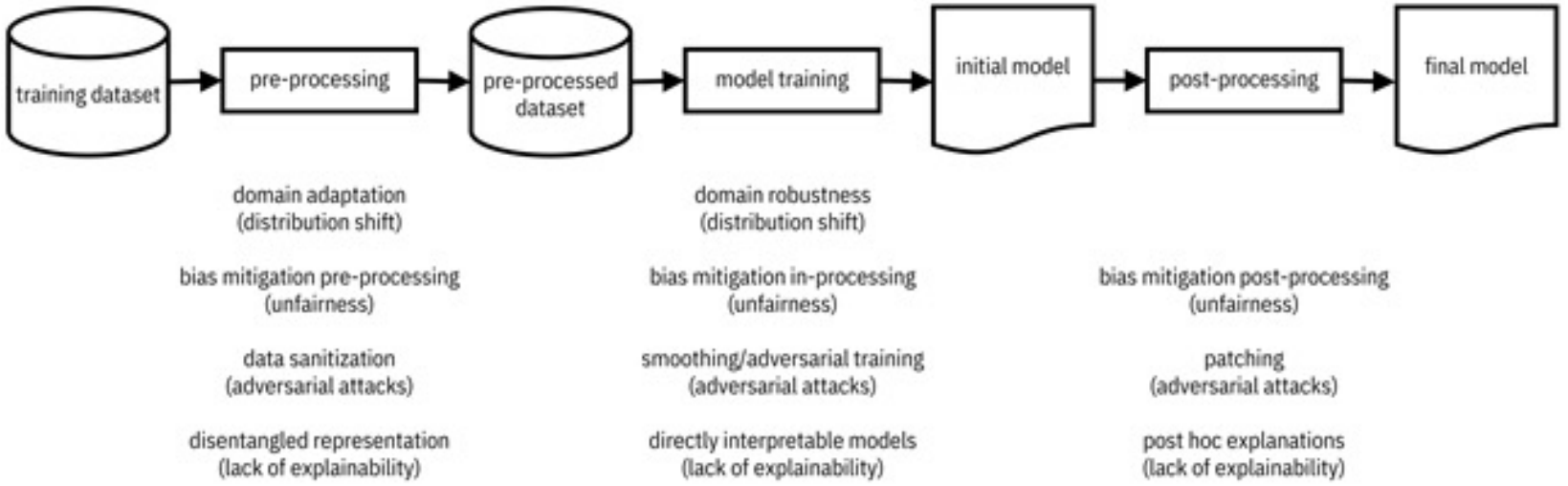
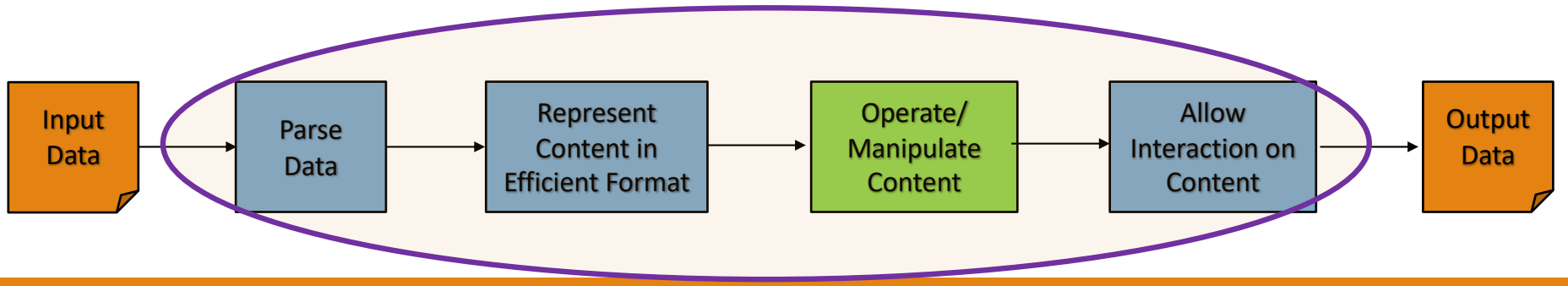


Image Credit: Trustworthy Machine Learning, Kush Varshney

Main Segment



Recall Exercise: German Credit

Discussion on Reading Material - 1

- Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (*FairWare '18*). Association for Computing Machinery, New York, NY, USA, 1–7. DOI: https://doi.org/10.1145/3194770.3194776
  - Pdf available on Arxiv, also stored in Blackboard

- 21 definitions of bias and their politics, Prof. Arvind Narayanan, Princeton, https://www.youtube.com/embed/jIXIuYdnyyk

German Credit Data

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

- 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

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 (categorical);
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); 1
17. Age (numerical);
18. Foreign worker (binary);
19. Dependents (numerical);
20. Other debtors (categorical);
21. Credit score (binary)

Example Instance:

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

Example record: Alice is requesting a loan amount of 1567 DM for a duration of 12 months for the purpose of purchasing a television, with a positive checking account balance that is smaller than 200 DM, having less than 100 DM in savings account, and having one existing credit at this bank. She duly paid existing credits at the bank till now and has no other installment plan. She possesses a car and owns a house, has been living at the present residence for one year and has a registered telephone. She is a skilled employee, working in the present employment for past four years. She is a 22-year-old married female and is a German citizen. She has one dependent and no guarantors. The recorded outcome for Alice (attribute #21) is a good credit score.

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

Datasets

- UCI Dataset:
  - Weka: https://www.ics.uci.edu/~mlearn/MLRepository.html (e.g., download: https://prdownloads.sourceforge.net/weka/uci-20070111.tar.gz)
  - Check in UCI – variants:
  - https://archive.ics.uci.edu/dataset/573/south+german+credit+update

- Weka
  - Direct link: https://github.com/Waikato/weka-3.8/blob/master/wekadocs/data/credit-g.arff
  - As part of development packages
    - like DataHub, https://datahub.io/machine-learning/credit-g#python

German Credit Data

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

- 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**.

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 (categorical);
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); 1
17. Age (numerical);
18. Foreign worker (binary);
19. Dependents (numerical);
20. Other debtors (categorical);
21. Credit score (binary)

Review detailed data exploration at:

- https://www.kaggle.com/sanyalush/predicting-credit-risk

Example record: Alice is requesting a loan amount of 1567 DM for a duration of 12 months for the purpose of purchasing a television, with a positive checking account balance that is smaller than 200 DM, having less than 100 DM in savings account, and having one existing credit at this bank. She duly paid existing credits at the bank till now and has no other installment plan. She possesses a car and owns a house, has been living at the present residence for one year and has a registered telephone. She is a skilled employee, working in the present employment for past four years. She is a 22-year-old married female and is a German citizen. She has one dependent and no guarantors. The recorded outcome for Alice (attribute #21) is a good credit score.

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

Over 20 Definitions of Bias

Divides them along 5 categories:

1. C1: predicted outcome
2. C2: predicted and actual outcome
3. C3: predicted probabilities and actual outcome
4. C4: similarity based
5. C5: causal reasoning

Definitions from literature
as of January 2018!
Many more since.

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

Paper's Experiment

- Created a logistic regression model
  - Attributes: Used binary and numeric attributes as-is; created categorical attribute to a set of binary features; total 48 attributes
  - 90% training, 10% test
- Model favors single males when deciding on the credit score and disadvantage divorced males; behaves similarly for females and married males similarly

Attribute	Coefficient
Personal status and gender: single male	0.16
Personal status and gender: married male	-0.04
Personal status and gender: married/divorced female	-0.08
Personal status and gender: divorced male	-0.14

Metrics for Model Performance

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Positive predictive value (PPV, **Precision**)
False discovery rate (FDR)
False omission rate (FOR)
Negative predictive value (NPV)

True positive rate (TPR, **Recall**)
False positive rate (FPR)
False negative rate (FNR)
True negative rate (TNR)

Definitions - 1

- **[C1] Group fairness ((a.k.a. statistical parity, equal acceptance rate, benchmarking):** if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class
  - **P(d = 1|G = m) = P(d = 1|G = f)**
  - the probability to have a good predicted credit score for married / divorced male and female applicants is 0.81 and 0.75, respectively. Hence, fails this metric.

- **[C1] Conditional statistical parity**: if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L. L considered are: requested credit amount, applicant's credit history, employment, and age
  - P(d = 1|L = l,G = m) = P(d = 1|L = l,G = f).
  - Controlling for L, the probability for married / divorced male and female applicants to have good predicted credit score is 0.46 and 0.49, respectively. Considers metric **satisfied**.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C1: Definitions Based on Predicted Outcome

Definitions - 2

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C2] Predictive parity ((a.k.a. outcome test).**: if both protected and unprotected groups have equal PPV
  - for both male and female applicants, the probability of an applicant with a good predicted credit score to actually have a good credit score should be the same. Represented as: $P(Y = 1 | d = 1, G = m) = P(Y = 1 | d = 1, G = f)$.
  - A classifier with equal PPVs will also have equal FDRs: $P(Y = 0 | d = 1, G = m) = P(Y = 0 | d = 1, G = f)$.
  - PPV for married / divorced male and female applicants is 0.73 and 0.74, respectively. Inversely, FDR for male and female applicants is 0.27 and 0.26, respectively. Hence, metric **satisfied**.

- **[C2] False positive error rate balance (a.k.a. predictive equality)**: if both protected and unprotected groups have equal FPR
  - the probability of a subject in the negative class to have a positive predictive value: $P(d = 1 | Y = 0, G = m) = P(d = 1 | Y = 0, G = f)$
  - A classifier with equal FPRs will also have equal TNRs: $P(d = 0 | Y = 0, G = m) = P(d = 0 | Y = 0, G = f)$.
  - FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively. Inversely, TNR is 0.30 and 0.45. Hence, fails this metric.

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 3

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C2] False negative error rate balance (aka, equal opportunity)**: if both protected and unprotected groups have equal FNR
  - the probability of a subject in a positive class to have a negative predictive value: P(d = 0|Y = 1,G = m) = P(d = 0|Y = 1,G = f).
  - A classifier with equal FNRs will also have equal TPR: P(d = 1|Y = 1,G = m) = P(d = 1|Y = 1,G = f).
  - FPRs for married / divorced male and female applicants are the same – 0.14. Inversely, TPR is 0.86. Hence, metric **satisfied**.

- **[C2] Equalized odds ((a.k.a. conditional procedure accuracy equality and disparate mistreatment):** if protected and unprotected groups have equal TPR and FPR
  - the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and the probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should both be same for male and female applicants:
  - **P(d = 1|Y = i,G = m) = P(d = 1|Y = i,G = f), i ∈ 0, 1.**
  - FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively and TPR is 0.86 for both males and females. Hence, fails this metric

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 4

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C2] Conditional use accuracy equality**: if equal PPV and NPV
  - the probability of subjects with positive predictive value to truly belong to the positive class and the probability of subjects with negative predictive value to truly belong to the negative class: (P(Y = 1|d = 1,G = m) = P(Y = 1|d = 1,G = f)) ∧ (P(Y = 0|d = 0,G = m) = P(Y = 0|d = 0,G = f)).
  - The calculated value for male and female applicants is 0.73 and 0.74, respectively. NPVs for male and female applicants is 0.49 and 0.6
  - More likely for a male than female applicant with a bad predicted score to actually have a good credit score. Fails this metric.

- **[C2] Overall accuracy equality:** if both protected and unprotected groups have equal prediction accuracy
  - the probability of a subject from either positive or negative class to be assigned to its respective class
  - **P(d = Y,G = m) = P(d = Y,G = f)**
  - **T**he overall accuracy rate is 0.68 and 0.71 for male and female applicants, respectively. Paper considers metric **satisfied**.

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C2: Definitions Based on Predicted and Actual Outcomes

Definitions - 5

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C2] Treatment equality:** satisfies this definition if both protected and unprotected groups have an equal ratio of false negatives and false positives.
  - ratio of FP to FN is same for male and female applicants: [FN/ FP] m = [FN / FP] f
  - Calculated ratios are 0.56 and 0.62 for male and female applicants. Fails this metric.

- **[C3] Test fairness (a.k.a. calibration, matching conditional frequencies):** if for any predicted probability score S, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class
  - for any given predicted probability score s in [0, 1], the probability of having actually a good credit score should be equal for both male and female applicants: P(Y = 1 | S = s,G = m) = P(Y = 1 | S = s,G = f).
  - It is more likely for a male applicant with a bad predicted credit score (low values of S) to actually have a good score (definition 3.2.5), but applicants with a good predicted credit score (high values of S) have an equivalent chance to indeed have a good credit score, regardless of their gender Paper considers metric **satisfied**.

C3: Definitions Based on Predicted Probabilities and Actual Outcome

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1\|S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1\|S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4: Calibration scores for different values of s

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

Definitions - 6

- **[C3] Well-calibration:** for any predicted probability score S, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to S.
  - for any given predicted probability score s in [0, 1], the probability of having actually a good credit score should be equal for both male and female applicants: P(Y = 1|S = s,G = m) = P(Y = 1|S = s,G = f) = s.
  - Paper considers metric (partially) **satisfied**.

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1\|S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1\|S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4: Calibration scores for different values of s

- **[C3] Balance for positive class:** if subjects constituting positive class from both protected and unprotected groups have equal average predicted probability score S.
  - expected value of probability assigned by the classifier to male and female applicant with good actual credit score should be same: E(S |Y = 1,G = m) = E(S |Y =1,G = f)
  - The calculated expected value of predicted probability score is 0.72 for both males and females. Paper considers metric **satisfied**.
  - Paper sees this along with equal opportunity/ false negative error rate balance which was also satistied.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C3: Definitions Based on Predicted Probabilities and Actual Outcome

Definitions - 7

- **[C3] Balance for negative class:** if subjects constituting negative class from both protected and unprotected groups have equal average predicted probability score S.
  - expected value of probability assigned by the classifier to male and female applicant with good actual credit score should be same: E(S |Y = 0,G = m) = E(S |Y =0,G = f)
  - The the expected value of having bad predicted credit score is 0.61 and 0.52 for males and females. Paper sees failing of this metric.
  - Paper sees this along with predictive equality/ false positive error rate balance which was also not satisfied

- **[C4] Causal discrimination:** if it produces the same classification for any two subjects with the exact same attributes X.
  - A male and female applicants who otherwise have the same attributes X will either both be assigned a good credit score or both assigned a bad credit score:
    (Xf = Xm ∧ Gf != Gm)→ df = dm.
  - For 8.8% married / divorced male and female applicants, the output classification was not same. Paper sees metric not being satisfied.
  - **Closest to legal notion on dissimilar treatment ?**

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C4: Similarity Based Measures

Definitions - 8

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

- **[C4] Fairness through unawareness:** if no sensitive attributes are explicitly used in the decision-making process
  - the classification outcome should be the same for applicants i and j who have the same attributes X: Xi = Xj →di = dj
  - Trained a model without gender and checked. Found classification identical for both genders. Metric **satisfied**.

- **[C4] Fairness through awareness :** similar individuals should have similar classification
  - For a set of applicants V , a distance metric between applicants k : VxV → R, a mapping from a set of applicants to probability distributions over outcomes M : V → δA, and a distance D metric between distribution of outputs, fairness is achieved iff D(M(x),M(y)) ≤ k(x,y).
  - Dependent on distance metric
  - k: gender-based; similar to causal discrimination (8.8% difference)
    k: age-based; the distance between outcomes (column 3) grew much faster than the distance between ages (column 3). Thus, the percentage of applicants who did not satisfy this definition (column 4) increased.
    Metric unsatisfied

Age difference	k	Avg. D	% violating cases
5	0.09	0.02	0.0
10	0.18	0.05	0.5
15	0.27	0.10	1.8
20	0.36	0.2	4.5
25	0.45	0.3	6.7

Table 5: Fairness through awareness with age-based distance

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C4: Similarity Based Measures

Definitions - 9

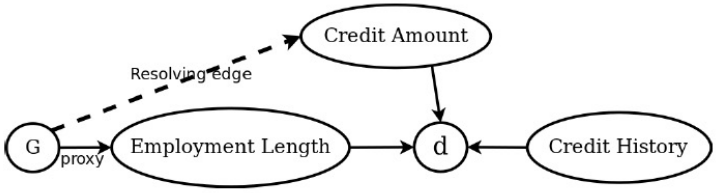


Figure 1: Causal graph example

- **Definitions based on knowledge of the domain:** given a relationship between attributes, check the relationship between decision and protected variables, other variables
  - Proxy attribute: whose value can be used to derive a value of another attribute
  - Resolving attribute: attribute that is influenced by the protected attribute in a non-discriminatory manner

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

C5: Causal Reasoning

Group Discussion – 10 mins

- A: Suppose you are **a loan applicant** and your application was
  - Accepted (good credit). Which metric would you use to explain to some other applicant whose application was rejected (bad credit)?
  - Rejected (bad credit). Which metric would you use to ask some other applicant whose application was accepted (bad credit)

- B: Suppose you are the executive at the bank incharge of making loans. Which metric will you use to justify that your bank does not consider gender or age to discriminate in your processing?

- C: Suppose you are the government regulator or a journalist. Which metric will you use to check if the bank discriminates in it loan practices ?

Lecture 11: Concluding Comments

- We looked at bias definitions
  - Five categories: C1: predicted outcome, C2: predicted and actual outcome, C3: predicted probabilities and actual outcome , C4: similarity based, C5: causal reasoning
  - Reviewed with respect to German-credit as example
- Metrics should not only be technically sound but practically useful
  - Did role-playing to discuss
- Most are theoretical exercises while law catches up; little technical guidance to developers

Trust Issues, Bias, Mitigation

Taxonomy of Biases

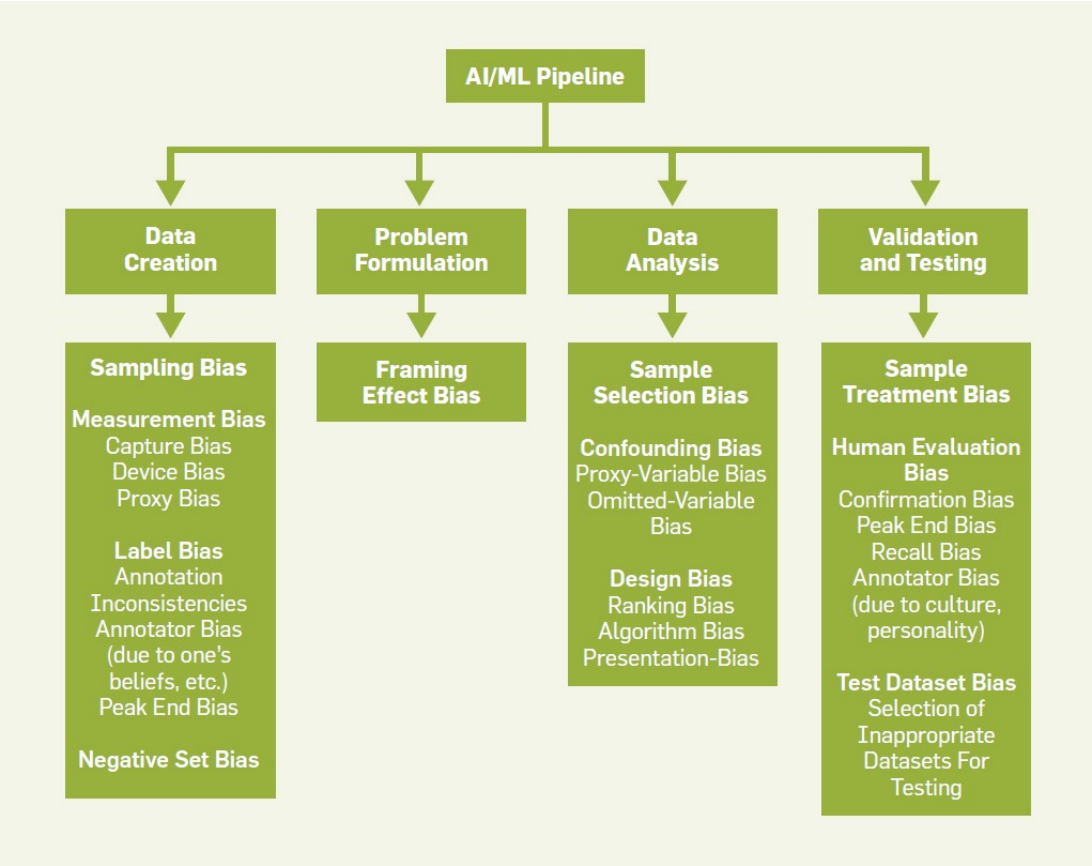


Figure courtesy: "Biases in AI Systems", Ramya Srinivasan, Ajay Chander, CACM 2021

Discussion

- Term ‘bias’ is used in many disciplines
- Technically, in ethics literature, fairness is defined as lack of bias. Achieving fairness is a good goal overall
- **Ethics normally plays a role when technology (AI) use will impact people**
- Practical issues
  - Identify which type of bias(es) to handle
    - Are they in data as well as in society
  - Which type of fairness to achieve
  - Which type of method to use

Class Exercise (30 Mins)

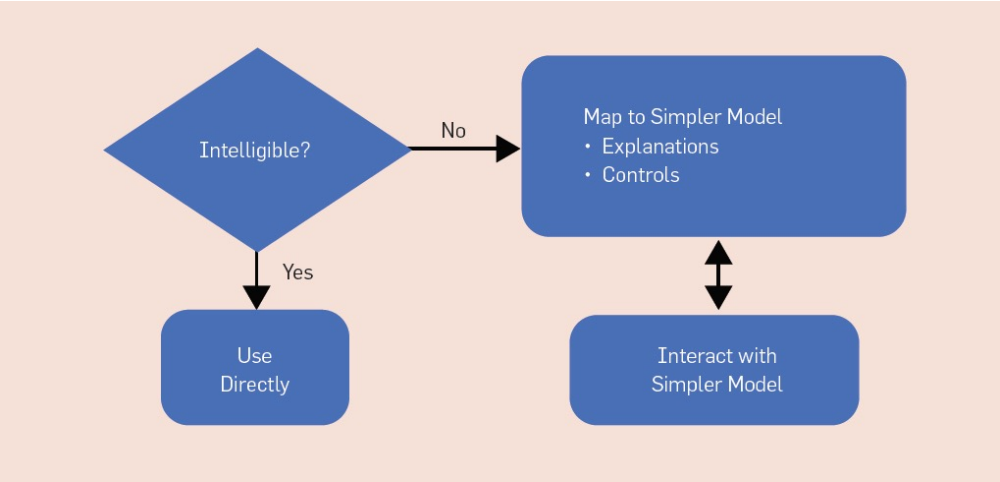
Look at your course project and answer

	End-user	AI Developer	Regulator
Q1: What may be the fairness issues/ sources of bias?			
Q2: Which type of fairness to achieve?			
Q3: What happens if we ignore it?			

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Definition	Paper	Citation #	Result
Group fairness or statistical parity	[12]	208	×
Conditional statistical parity	[11]	29	✓
Predictive parity	[10]	57	✓
False positive error rate balance	[10]	57	×
False negative error rate balance	[10]	57	✓
Equalised odds	[14]	106	×
Conditional use accuracy equality	[8]	18	×
Overall accuracy equality	[8]	18	✓
Treatment equality	[8]	18	×
Test-fairness or calibration	[10]	57	✓
Well calibration	[16]	81	✓
Balance for positive class	[16]	81	✓
Balance for negative class	[16]	81	×
Causal discrimination	[13]	1	×
Fairness through unawareness	[17]	14	✓
Fairness through awareness	[12]	208	×
Counterfactual fairness	[17]	14	–
No unresolved discrimination	[15]	14	–
No proxy discrimination	[15]	14	–
Fair inference	[19]	6	–

Setting and Terminology: Intelligible Models and Explanations



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

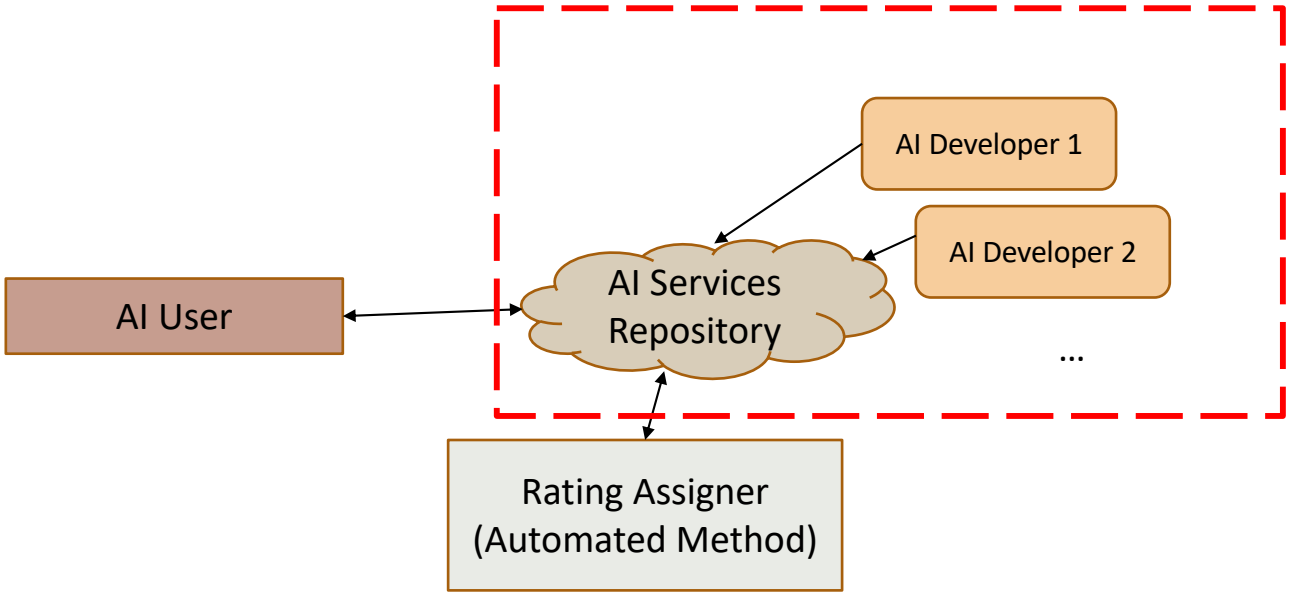
Sources:
1. The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
2. Explainable Machine Learning in Deployment, FAT* 2020.

Trust Issues – How to Mitigate

- Explain behavior
  - Remove undesirable behavior ?
    - Explain that too?

Trust Dimensions
Competent
Reliable
Upholds human values
Allows human interaction

Idea: Develop Automated Methods to Rate AI Systems That Can be Used for Communicating Trust in Black Box Setting



NIST Risk Management Framework

Prepare	Essential activities to **prepare** the organization to manage security and privacy risks
Categorize	**Categorize** the system and information processed, stored, and transmitted based on an impact analysis
Select	**Select** the set of NIST SP 800-53 controls to protect the system based on risk assessment(s)
Implement	**Implement** the controls and document how controls are deployed
Assess	**Assess** to determine if the controls are in place, operating as intended, and producing the desired results
Authorize	Senior official makes a risk-based decision to **authorize** the system (to operate)
Monitor	Continuously **monitor** control implementation and risks to the system

Credit: https://csrc.nist.gov/projects/risk-management

Trust Issues – How to Mitigate

- Communicate behavior (transparency)
  - Let humans make decisions

Trust Dimensions
Competent
Reliable
Upholds human values
Allows human interaction

Discussion on Reading Material - 2

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

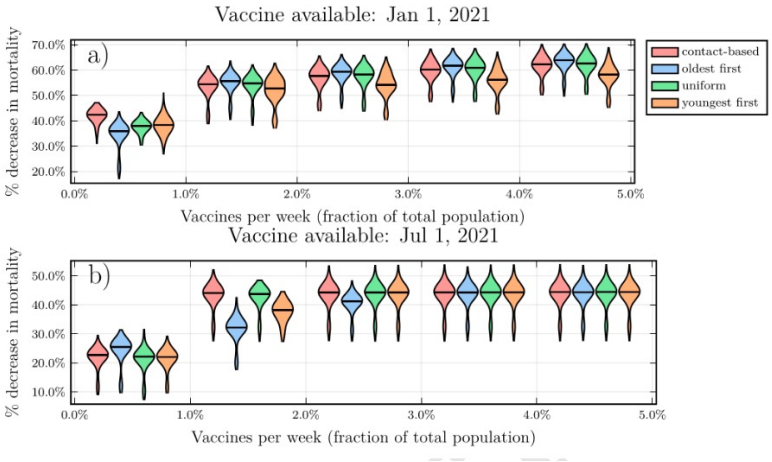
- Approach of report
  - A series of workshops in academic circles in Nov and Dec 2020
  - Discussions and findings distilled into the report

Example AI Usage: Distribution of Vaccines

- **Problem**: Limited supply, larger demand, many technologies, recipient hesitancy; How do distribute equitably, fairly and efficiently
- Possible (automated) solutions
  - Random: pick receiver based on random choice
    - **Benefit**: Easy to implement
    - **Problems**: Equitable but not fair, receiver may not be at risk or not want it, others wanting it may not get it
    - **Question**: assumes we can give vaccine quickly to the selected person
  - Prioritized random: make a prioritized list of groups, assign randomly in each group
    - **Benefit**: identifies affected groups
    - **Problems**: receiver may not want the vaccine
    - **Question**: who comes up with groups?, is it rewarding groups who have not been taking precautions ? Assumes we can give vaccine quickly to the selected person
  - ...
  - Benefit-cost: based on contribution to economy
    - **Benefit**: efficient

Impact of Decisions in Vaccine Distribution

- Article: 'The Pandemic Is a Prisoner's Dilemma Game'
- Prioritising COVID-19 vaccination in changing social and epidemiological landscapes, Sep 2020.
- Choices
  - impact of vaccinating 60+ year-olds first;
  - <20 year-olds first;
  - uniformly by age; and
  - a novel contact-based strategy
- Insights
  - Vaccination reduces median deaths by 32%-77% (22%-63%) for January (July) availability, depending on the scenario.
  - Vaccinating 60+ year-olds first prevents more deaths (up to 8% more) than transmission-interrupting strategies



Discussion on Reading Material

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

- Findings
  1. Researchers responded to COVID need with enthusiasm leading to a large number of projects
     1. **Word-wide* (for context)**: protein to aid disease detection and treatment (**molecular scale**), the analysis of patient data like images and conditions to improve patient care (**clinical scale**) and analysis of cases and social media to predict disease severity, understand mis-information and communicate effectively (**societal scale**).
     2. **UK specific examples**: model disease spread, navigate lockdown
  2. Major hurdle was lack of "robust and timely data", especially access and standardization
     1. Develop protocols for collecting and managing protected data
     2. Develop protocols for generating anonymized and synthetic data

* Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. In Journal of Artificial Intelligence Research 69, 807-845, 2020.

Discussion on Reading Material

"Data science and AI in the age of COVID-19 - Reflections on the response of the UK's data science and AI community to the COVID-19 pandemic", Alan Turing Institute, 2021

- Findings
  3. Concern over inequality and exclusion slowed progress. Inadequate representation and engagement from some groups
  4. Challenge in communicating research findings to policy makers and public. Specifically, timeliness, accuracy and clarity.
     - Communication among experts
     - Communication among researchers and policy makers
     - Communication among researchers and public

* Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. In Journal of Artificial Intelligence Research 69, 807-845, 2020.

Project Discussion

Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
     1. Solution => ML task (e.g. classification), recommendation, text summarization, …
     2. Use a foundation model (e.g., LLM-based) solution as the baseline
  4. (Data) Explore the data for a solution to work
  5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Project Discussion

1. Create a private Github repository called "CSCE581-Spring2025-<studentname>-Repo". Share with Instructor (biplav-s)
2. Create a folder called "Project". Inside, create a text file called "ProjectPlan.md" (or "ProjectPlan.txt") and have details by the next class (Jan 30, 2025)

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

Concluding Section

Week 6 (L11 and 12): Concluding Comments

- We looked at
  - Fairness methods
  - Overview of major mitigation techniques – explanation and rating

About Next Week – Lectures 13, 14

Lectures 13, 14: Explanation and Rating

- ML/ Classification: Trust Mitigation – Explanation
- ML/ Classification: Trust Mitigation - Rating

9	Feb 11 (Tu)	Quiz 1
10	Feb 13 (Th)	AI - Structured: Analysis – Supervised ML – Trust Issues
11	Feb 18 (Tu)	AI - Structured: Analysis – Supervised ML – Trust Issues
12	Feb 20 (Th)	AI - Structured: Analysis – Supervised ML – Mitigation Methods
13	Feb 25 (Tu)	AI - Supervised ML: Explanation Tools
14	Feb 27 (Th)	AI Trust - Mitigation method (Trust rating) – Kausik Lakkaraju
15	Mar 4 (Tu)	Student presentations - project
16	Mar 6 (Th)	Machine Learning – Trust Issues (Explainability)