Lesson 2.4: Algorithmic Fairness and Bias – Quiz

Module 2: The Access Problem

Prof. Dr. Joerg Osterrieder

Question 1

A credit scoring model does not use race as an input feature. However, it uses zip code, which is highly correlated with race due to historical residential segregation. What type of discrimination does this represent?

A. Disparate treatment – the model explicitly uses a protected attribute

B. Disparate impact – a facially neutral feature produces disproportionate outcomes

C. Neither – since race is not used, there is no discrimination

D. Disparate treatment through proxy – which is the same as using race directly

Question 1

A credit scoring model does not use race as an input feature. However, it uses zip code, which is highly correlated with race due to historical residential segregation. What type of discrimination does this represent?

A. Disparate treatment – the model explicitly uses a protected attribute

B. Disparate impact – a facially neutral feature produces disproportionate outcomes

C. Neither – since race is not used, there is no discrimination

D. Disparate treatment through proxy – which is the same as using race directly

Answer: B

Zip code is facially neutral (not a protected attribute), but produces disproportionate outcomes because it correlates with race. This is classic disparate impact. Disparate treatment requires explicit use of the protected attribute.

Question 2

A lending model approves 60% of white applicants and 42% of Black applicants. What is the disparate impact ratio, and does it satisfy the four-fifths rule?

A. DI = 0.70, does not satisfy the four-fifths rule (below 0.80)

B. DI = 0.70, satisfies the four-fifths rule (above 0.60)

C. DI = 1.43, satisfies the four-fifths rule

D. DI = 0.42, does not satisfy the four-fifths rule

Question 2

A lending model approves 60% of white applicants and 42% of Black applicants. What is the disparate impact ratio, and does it satisfy the four-fifths rule?

A. DI $= 0.70$, does not satisfy the four-fifths rule (below 0.80)

B. DI $= 0.70$, satisfies the four-fifths rule (above 0.60)

C. DI $= 1.43$, satisfies the four-fifths rule

D. DI $= 0.42$, does not satisfy the four-fifths rule

Answer: A

DI $= 42\% \,/\, 60\% = 0.70$. The four-fifths rule requires DI ≥ 0.80. Since $0.70 < 0.80$, there is evidence of disparate impact.

Question 3

Which fairness metric requires that the approval rate be equal across all demographic groups, regardless of their qualifications?

A. Calibration
B. Equalized odds
C. Demographic parity
D. Counterfactual fairness

Question 3

Which fairness metric requires that the approval rate be equal across all demographic groups, regardless of their qualifications?

A. Calibration
B. Equalized odds
C. Demographic parity
D. Counterfactual fairness

Answer: C
Demographic parity requires $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$—equal positive prediction rates regardless of qualifications. Equalized odds conditions on the true label; calibration requires equal meaning of scores.

Question 4

The Chouldechova impossibility theorem states that it is impossible to simultaneously satisfy demographic parity, equalized odds, and calibration, except when:

A. The model is perfectly accurate (AUC $= 1.0$)

B. Base rates (e.g., default rates) are equal across groups

C. The model uses no proxy variables

D. Post-processing fairness corrections are applied

Question 4

The Chouldechova impossibility theorem states that it is impossible to simultaneously satisfy demographic parity, equalized odds, and calibration, except when:

A. The model is perfectly accurate (AUC = 1.0)

B. Base rates (e.g., default rates) are equal across groups

C. The model uses no proxy variables

D. Post-processing fairness corrections are applied

Answer: B

The impossibility theorem holds when base rates differ between groups. If Group A has a 5% default rate and Group B has a 12% default rate, no model can simultaneously achieve all three metrics. Only when base rates are equal is simultaneous satisfaction possible.

Question 5

A SHAP beeswarm plot shows that "zip code" is the third most important feature in a credit scoring model. The SHAP values for zip codes in predominantly minority neighborhoods are consistently negative (pushing toward denial). What should the fairness auditor conclude?

A. The model is working correctly – zip code reflects legitimate risk differences

B. Zip code may be acting as a proxy for race, and further disparate impact analysis is needed

C. The SHAP values are meaningless because SHAP cannot detect proxy variables

D. Remove zip code and retrain – this conclusively proves racial bias

Question 5

A SHAP beeswarm plot shows that "zip code" is the third most important feature in a credit scoring model. The SHAP values for zip codes in predominantly minority neighborhoods are consistently negative (pushing toward denial). What should the fairness auditor conclude?

A. The model is working correctly – zip code reflects legitimate risk differences

B. Zip code may be acting as a proxy for race, and further disparate impact analysis is needed

C. The SHAP values are meaningless because SHAP cannot detect proxy variables

D. Remove zip code and retrain – this conclusively proves racial bias

Answer: B

SHAP identifies *what* drives predictions, not *why*. Zip code may reflect legitimate risk (local economic conditions) or may proxy for race (redlining legacy). Further analysis (e.g., testing disparate impact with and without zip code) is needed before concluding bias.

Question 6

What is the key difference between SHAP and LIME for model explainability?

A. SHAP provides global explanations; LIME provides only local explanations

B. SHAP is based on cooperative game theory and provides both global and local explanations; LIME approximates the model locally using perturbations

C. LIME is more theoretically grounded than SHAP

D. SHAP works only with tree-based models; LIME works with any model

Question 6

What is the key difference between SHAP and LIME for model explainability?

A. SHAP provides global explanations; LIME provides only local explanations

B. SHAP is based on cooperative game theory and provides both global and local explanations; LIME approximates the model locally using perturbations

C. LIME is more theoretically grounded than SHAP

D. SHAP works only with tree-based models; LIME works with any model

Answer: B

SHAP uses Shapley values (cooperative game theory) and provides both global and local explanations. LIME perturbs inputs locally and fits a simple surrogate model. Both are model-agnostic, but SHAP has stronger theoretical guarantees.

Question 7

In the Apple Card case (2019), Goldman Sachs stated that gender was not used as a model input. Yet women consistently received lower credit limits. This failure illustrates which concept?

A. Disparate treatment – gender was explicitly used
B. Fairness through unawareness – removing the protected attribute did not remove the bias
C. Calibration failure – the model's probabilities were wrong
D. Equalized odds – the model had equal error rates

Question 7

In the Apple Card case (2019), Goldman Sachs stated that gender was not used as a model input. Yet women consistently received lower credit limits. This failure illustrates which concept?

A. Disparate treatment – gender was explicitly used

B. Fairness through unawareness – removing the protected attribute did not remove the bias

C. Calibration failure – the model's probabilities were wrong

D. Equalized odds – the model had equal error rates

Answer: B

"Fairness through unawareness" is the strategy of simply excluding protected attributes. The Apple Card case showed this fails because other features (spending patterns, income sources) act as proxies for gender.

Question 8

A model has the following performance across two groups:

Metric	Group A	Group B
Accuracy	92%	78%
False Positive Rate	5%	18%
False Negative Rate	3%	4%

Which fairness criterion does this model most clearly violate?

A. Calibration – predicted probabilities differ in meaning

B. Equalized odds – the false positive rate differs dramatically (5% vs. 18%)

C. Demographic parity – the approval rates are unequal

D. The model is fair because false negative rates are similar

Question 8

A model has the following performance across two groups:

Metric	Group A	Group B
Accuracy	92%	78%
False Positive Rate	5%	18%
False Negative Rate	3%	4%

Which fairness criterion does this model most clearly violate?

A. Calibration – predicted probabilities differ in meaning

B. Equalized odds – the false positive rate differs dramatically (5% vs. 18%)

C. Demographic parity – the approval rates are unequal

D. The model is fair because false negative rates are similar

Answer: B

Equalized odds requires equal TPR and FPR across groups. The FPR difference (5% vs. 18%) is dramatic – Group B members are 3.6x more likely to be falsely flagged. This is exactly the pattern ProPublica found in COMPAS.

Question 9

Which bias mitigation approach modifies the model's decision thresholds per group after training to equalize outcomes?

A. Pre-processing: reweighting
B. In-processing: adversarial debiasing
C. Post-processing: threshold adjustment
D. Pre-processing: disparate impact remover

Question 9

Which bias mitigation approach modifies the model's decision thresholds per group after training to equalize outcomes?

A. Pre-processing: reweighting
B. In-processing: adversarial debiasing
C. Post-processing: threshold adjustment
D. Pre-processing: disparate impact remover

Answer: C
Post-processing threshold adjustment sets different classification thresholds per group (e.g., approve if score > 0.55 for Group A, > 0.48 for Group B) to equalize outcomes. It is model-agnostic but may raise legal concerns about explicit differential treatment.

Question 10

Counterfactual fairness asks: "Would the decision have been different if the individual belonged to a different group?" What does this approach fundamentally require?

A. A large test dataset with balanced group representation

B. A causal model specifying how protected attributes influence other features

C. A perfectly accurate model with AUC $= 1.0$

D. Protected attributes to be included as direct model inputs

Question 10

Counterfactual fairness asks: "Would the decision have been different if the individual belonged to a different group?" What does this approach fundamentally require?

A. A large test dataset with balanced group representation

B. A causal model specifying how protected attributes influence other features

C. A perfectly accurate model with AUC $= 1.0$

D. Protected attributes to be included as direct model inputs

Answer: B

Counterfactual fairness requires a causal graph: flipping race means adjusting all causally downstream features (income, education, neighborhood). Without specifying these causal relationships, the counterfactual is undefined.

Question 11

Under the EU AI Act, credit scoring algorithms are classified as:

A. Minimal risk – no requirements beyond general transparency
B. Limited risk – transparency obligations only
C. High risk – requiring risk management, data governance, transparency, and human oversight
D. Unacceptable risk – prohibited entirely

Question 11

Under the EU AI Act, credit scoring algorithms are classified as:

A. Minimal risk – no requirements beyond general transparency

B. Limited risk – transparency obligations only

C. High risk – requiring risk management, data governance, transparency, and human oversight

D. Unacceptable risk – prohibited entirely

Answer: C

The EU AI Act classifies AI systems used for creditworthiness assessment and credit scoring as high-risk, requiring conformity assessments, technical documentation, data governance, transparency, human oversight, and accuracy/robustness standards.

Question 12

A bank's model development team reports that their new credit model has "no bias" because overall accuracy is 91%. What is the fundamental flaw in this claim?

A. 91% accuracy is too low for production deployment

B. Overall accuracy can mask significant disparities in error rates across demographic groups

C. Accuracy is not a valid metric for credit scoring

D. The team should have used AUC-ROC instead of accuracy

Question 12

A bank's model development team reports that their new credit model has "no bias" because overall accuracy is 91%. What is the fundamental flaw in this claim?

A. 91% accuracy is too low for production deployment

B. Overall accuracy can mask significant disparities in error rates across demographic groups

C. Accuracy is not a valid metric for credit scoring

D. The team should have used AUC-ROC instead of accuracy

Answer: B

A model with 91% overall accuracy could have 96% accuracy for Group A and 72% for Group B. Disaggregated metrics are essential. Fairness cannot be assessed from aggregate performance alone.

Question 13

In the COMPAS case, ProPublica showed that Black defendants had a higher false positive rate, while Northpointe showed the model was calibrated. Both were correct. How is this possible?

A. One of them made a statistical error

B. They measured different aspects of fairness, and the impossibility theorem proves both cannot be satisfied simultaneously when base rates differ

C. COMPAS was only used for sentencing, not prediction

D. The false positive rates were actually equal after adjusting for income

Question 13

In the COMPAS case, ProPublica showed that Black defendants had a higher false positive rate, while Northpointe showed the model was calibrated. Both were correct. How is this possible?

A. One of them made a statistical error

B. They measured different aspects of fairness, and the impossibility theorem proves both cannot be satisfied simultaneously when base rates differ

C. COMPAS was only used for sentencing, not prediction

D. The false positive rates were actually equal after adjusting for income

Answer: B

ProPublica measured equalized odds (equal error rates); Northpointe measured calibration (equal score meaning). The impossibility theorem proves both cannot hold when base rates differ – and recidivism base rates differed. This is the theorem's most famous real-world demonstration.

Question 14

Which of the following is a proxy variable for race in US credit scoring?

A. Annual income
B. Zip code
C. Employer name
D. All of the above

Question 14

Which of the following is a proxy variable for race in US credit scoring?

A. Annual income
B. Zip code
C. Employer name
D. All of the above

Answer: D
All three can act as proxies: zip code correlates with race due to residential segregation; income reflects historical wage gaps; employer name correlates with industry concentration by race. Any feature correlated with a protected attribute is a potential proxy.

Question 15

A credit model's fairness audit reveals that "years at current address" disproportionately penalizes minority applicants. Which bias source does this represent?

A. Label bias – the target variable is incorrectly defined

B. Historical bias – systemic housing instability for minorities is encoded in this feature

C. Measurement bias – the feature is measured incorrectly for minority applicants

D. Aggregation bias – the model treats all groups identically when it should not

Question 15

A credit model's fairness audit reveals that "years at current address" disproportionately penalizes minority applicants. Which bias source does this represent?

A. Label bias – the target variable is incorrectly defined

B. Historical bias – systemic housing instability for minorities is encoded in this feature

C. Measurement bias – the feature is measured incorrectly for minority applicants

D. Aggregation bias – the model treats all groups identically when it should not

Answer: B

Historical discrimination in housing (redlining, discriminatory lending, exclusion from homeownership) means minority individuals are more likely to be renters and to move more frequently. The feature encodes this historical inequality, creating systematic disadvantage.

Question 16

Adversarial debiasing is an in-processing bias mitigation technique. How does it work?

A. It trains a secondary model to predict group membership from the main model's output, then penalizes the main model when group membership is predictable

B. It removes all features correlated with protected attributes before training

C. It augments the training data with synthetic samples from underrepresented groups

D. It adjusts decision thresholds per group after training

Question 16

Adversarial debiasing is an in-processing bias mitigation technique. How does it work?

A. It trains a secondary model to predict group membership from the main model's output, then penalizes the main model when group membership is predictable

B. It removes all features correlated with protected attributes before training

C. It augments the training data with synthetic samples from underrepresented groups

D. It adjusts decision thresholds per group after training

Answer: A

Adversarial debiasing trains a predictor and an adversary simultaneously. The adversary tries to predict the protected attribute from the predictor's output. The predictor is penalized when the adversary succeeds, forcing it to learn representations that do not encode group membership.

Question 17

Under the US Equal Credit Opportunity Act (ECOA), what must a lender provide when denying a credit application?

A. Nothing – lenders are not required to explain denials
B. A general statement that the applicant did not meet lending criteria
C. Specific reasons for the denial (adverse action notice) that the applicant can understand and act upon
D. The full model source code and training data

Question 17

Under the US Equal Credit Opportunity Act (ECOA), what must a lender provide when denying a credit application?

A. Nothing – lenders are not required to explain denials
B. A general statement that the applicant did not meet lending criteria
C. Specific reasons for the denial (adverse action notice) that the applicant can understand and act upon
D. The full model source code and training data

Answer: C
ECOA requires specific, actionable adverse action reasons (e.g., "high debt-to-income ratio," "insufficient credit history"). The CFPB has clarified (2022) that using complex ML models does not excuse vague explanations. SHAP/LIME can generate these reasons.

Question 18

A fairness audit finds that a lending model's overall disparate impact ratio is 0.83 (above the 0.80 threshold). However, for applicants aged 55+, the DI ratio is 0.62. What concept does this illustrate?

A. Simpson's paradox – aggregate fairness masks subgroup unfairness

B. The impossibility theorem – you cannot satisfy multiple fairness criteria

C. Intersectionality – unfairness appears at the intersection of protected attributes

D. Calibration failure – the model is poorly calibrated for older applicants

Question 18

A fairness audit finds that a lending model's overall disparate impact ratio is 0.83 (above the 0.80 threshold). However, for applicants aged 55+, the DI ratio is 0.62. What concept does this illustrate?

A. Simpson's paradox – aggregate fairness masks subgroup unfairness

B. The impossibility theorem – you cannot satisfy multiple fairness criteria

C. Intersectionality – unfairness appears at the intersection of protected attributes

D. Calibration failure – the model is poorly calibrated for older applicants

Answer: C

Intersectionality means that unfairness may be hidden when examining single protected attributes in isolation. The model may be "fair" for race overall, but unfair for older minority applicants specifically. Audits must test intersectional subgroups.

Question 19

Which statement about the fairness–accuracy tradeoff is most accurate?

A. Enforcing fairness always destroys model accuracy – fairness and accuracy are incompatible
B. Fairness constraints never affect accuracy – they only change the threshold
C. Moderate fairness improvements often require small accuracy sacrifices; the tradeoff follows a Pareto frontier
D. The tradeoff only exists for linear models, not for deep learning

Question 19

Which statement about the fairness–accuracy tradeoff is most accurate?

A. Enforcing fairness always destroys model accuracy – fairness and accuracy are incompatible

B. Fairness constraints never affect accuracy – they only change the threshold

C. Moderate fairness improvements often require small accuracy sacrifices; the tradeoff follows a Pareto frontier

D. The tradeoff only exists for linear models, not for deep learning

Answer: C

Research shows the fairness–accuracy Pareto frontier often has a "knee" where significant fairness gains come at modest accuracy cost. The tradeoff is real but often smaller than feared, and it applies to all model types.

Question 20

A financial institution is deploying a credit scoring model in both the EU and the US. Which combination of regulatory requirements must it satisfy?

A. EU AI Act only – it supersedes all national regulations

B. ECOA and Fair Lending (US) plus EU AI Act (EU) – both jurisdictions' requirements apply to their respective markets

C. Neither – credit scoring is exempt from AI regulation

D. US regulations only – the EU AI Act has not yet been enforced

Question 20

A financial institution is deploying a credit scoring model in both the EU and the US. Which combination of regulatory requirements must it satisfy?

A. EU AI Act only – it supersedes all national regulations

B. ECOA and Fair Lending (US) plus EU AI Act (EU) – both jurisdictions' requirements apply to their respective markets

C. Neither – credit scoring is exempt from AI regulation

D. US regulations only – the EU AI Act has not yet been enforced

Answer: B

Multinational institutions must satisfy each jurisdiction's requirements for operations in that jurisdiction. The US focuses on outcome-based testing (disparate impact), while the EU mandates process-based compliance (risk management system, documentation, human oversight).