Fairness Metrics and Evaluation

Your score: 12/12

Go back

1. A data scientist has evaluated a loan approval algorithm using demographic parity and finds similar approval rates across gender groups and separately across racial groups, but notices substantially lower approval rates for women from underrepresented minorities. Which approach most accurately explains and addresses this situation?

A. ◯ This pattern indicates a random statistical fluctuation that would disappear with larger sample sizes and requires no specific intervention.

B. ◯ This is a classic manifestation of Simpson's Paradox where aggregate fairness across individual attributes masks intersectional unfairness, requiring explicit assessment of fairness across demographic intersections rather than individual attributes alone.

C. ◯ The algorithm should be evaluated using individual fairness instead of group fairness, as intersectionality issues only affect group metrics and not individual-level comparisons.

D. ◯ This pattern indicates that the protected attributes themselves are correlated, requiring regularization of the input features to remove these correlations before training.

2. A team is developing a credit scoring algorithm and must implement a similarity metric for individual fairness. Which approach most accurately reflects current technical best practices for developing an appropriate similarity metric in this context?

A. ◯ Use Euclidean distance in the feature space after standardizing all variables to ensure mathematical consistency across different scales.

B. ◯ Apply dimensionality reduction techniques like t-SNE to create a low-dimensional embedding where proximity indicates similarity.

C. ◯ Develop a task-specific metric through a process that combines domain expertise about relevant financial factors, human judgments of similar cases, and analysis of how protected attributes influence other features through causal mechanisms.

D. ◯ Implement a universal fairness distance function that can be applied consistently across all financial applications to ensure standardized fairness evaluation.

3. When implementing intersectional fairness assessment for a medical diagnostic algorithm, the team discovers that certain demographic intersections (like older women from minority groups) have very small sample sizes in their validation data. Which technical approach most effectively addresses this statistical challenge?

A. ◯ Drop all intersectional subgroups with sample sizes below a fixed threshold to ensure statistical validity across all reported metrics.

B. ◯ Apply uniform fairness constraints across all intersections regardless of sample size to ensure consistent treatment of all demographic groups.

C. ◯ Implement a hierarchical Bayesian approach that pools information across related subgroups while allowing for subgroup differences, combined with clear uncertainty quantification in the form of credible intervals for fairness metrics.

D. ◯ For small intersections, randomly generate synthetic samples to increase the subgroup size until it reaches the same size as the largest demographic group.

4. When examining a medical diagnosis dataset, a researcher observes that measurement approaches for key symptoms show statistically significant differences in validity across demographic groups. Which technical approach most effectively addresses this measurement bias?

A. ◯ Apply different classification thresholds for each demographic group to compensate for measurement differences.

B. ◯ Validate and potentially redesign the feature operationalization process to ensure consistent measurement validity across groups.

C. ◯ Remove features with differential validity and rely only on features that show consistent measurement properties.

D. ◯ Apply a postprocessing correction factor to model outputs based on the observed measurement disparities.

5. When analyzing a recommendation system, a team observes that performance disparities between demographic groups increase during training despite starting from a relatively balanced initialization. Which algorithmic mechanism

most likely explains this observation, according to current technical consensus?

A. ◯ The model architecture lacks sufficient capacity to represent complex patterns for minority groups, causing increasingly worse approximations as training proceeds.

B. ◯ The loss function, by optimizing for aggregate performance metrics, naturally prioritizes patterns common in larger demographic groups that contribute more to the overall objective, creating a form of "representation disparity" documented by Hashimoto et al. (2018).

C. ◯ Improper feature normalization creates learning rate imbalances that cause majority group patterns to be learned first, blocking later learning of minority patterns.

D. ◯ Random weight initialization created unfavorable initial conditions for minority groups, and this disadvantage compounds through training dynamics.

6. In evaluating a recommendation system for an online learning platform, which approach most comprehensively assesses individual fairness?

A. ◯ Verify that recommendation distributions have equal statistical properties across protected groups, ensuring no disparate impact.

B. ◯ Implement comprehensive A/B testing with different protected attribute values to directly measure outcome differences.

C. ◯ Apply a multi-dimensional evaluation that: (1) verifies the Lipschitz condition across similar users with different protected attributes, (2) analyzes counterfactual examples to identify inappropriate protected

attribute influence, (3) validates the similarity metric against human judgments, and (4) examines edge cases where similar users receive significantly different recommendations.

D. ◯ Focus evaluation on user satisfaction surveys broken down by demographic group to determine if different groups perceive the recommendations as equally relevant.

7. A data scientist notices that a model trained on balanced data shows different error patterns across demographic groups despite the absence of explicit protected attributes in the feature set. The team tests four alternative models using identical training data and finds that a deep neural network shows 40% smaller performance disparities than a logistic regression model. Which conclusion most accurately reflects current technical understanding of algorithm bias?

A. ◯ The observed disparities must be caused by hidden data issues rather than algorithmic bias, as algorithmic choices can only amplify existing data problems.

B. ◯ Model architecture can fundamentally alter fairness outcomes even with identical data because different inductive biases align differently with group-specific patterns, making architecture selection a substantive fairness decision rather than just a performance choice.

C. ◯ The deep neural network appears fairer only because it overfits to the training data, and implementing proper regularization would restore the more realistic results from the logistic regression.

D. ◯ All complex models will show similar fairness properties with balanced data, suggesting the neural network's advantage is likely a statistical

anomaly rather than a meaningful architectural difference.

8. A data scientist is building a prediction model for a job recommendation system and discovers that certain job types are predominantly associated with specific genders in the training data. The team is debating whether this represents a bias that should be addressed. Which of the following approaches most accurately reflects current technical best practices for addressing this situation?

A. ◯ This pattern represents historical bias that should be preserved in the model to ensure statistical validity and prediction accuracy.

B. ◯ The system should implement adversarial debiasing to remove any statistical correlations between gender and job categories.

C. ◯ The team should analyze whether the observed patterns reflect societal disparities that the system might amplify, then implement targeted interventions based on intended system behavior and fairness goals.

D. ◯ Gender information should be completely removed from the dataset to ensure the model cannot perpetuate any gender-based patterns regardless of their origin.

9. A company is developing an automated resume screening system for early-career engineering positions. Initial testing shows the system approves male and female candidates at equal rates (satisfying demographic parity), but female candidates who receive positive predictions have a 15% lower chance of actually succeeding in the role compared to approved male candidates. Which fairness metric is being violated in this scenario, and what does it indicate about the system?

A. ◯ Equal opportunity is being violated, indicating the system is too lenient toward female candidates who would not succeed.

B. ◯ Equalized odds is being violated, showing inconsistent error rates that favor male candidates in both positive and negative predictions.

C. ◯ Predictive parity is being violated, indicating the system applies less stringent standards to female candidates, resulting in less reliable positive predictions for this group.

D. ◯ Statistical parity is being violated despite equal approval rates, because the representation equality is superficial rather than meaningful.

10. When implementing counterfactual fairness in a hiring algorithm that uses educational background as a feature, the team discovers that access to prestigious universities correlates strongly with socioeconomic status. Which technical approach most effectively addresses this challenge?

A. ◯ Remove educational institution from the feature set entirely to eliminate this source of potential bias.

B. ◯ Statistically rebalance the dataset to ensure equal representation from all socioeconomic groups at each educational institution level.

C. ◯ Develop a causal model that distinguishes between legitimate educational qualities (knowledge, skills) and access advantages, then create modified features that preserve predictive information while removing the problematic causal pathway between socioeconomic status and educational opportunity.

D. ◯ Apply a post-processing calibration that ensures equal prediction distributions across socioeconomic groups regardless of educational background.

11. When implementing group fairness metrics in a loan approval system, the data science team discovers that optimizing for demographic parity (equal approval rates) results in violating predictive parity (equal reliability of approvals), while optimizing for equal opportunity leads to unequal approval rates. Which approach best represents the current technical consensus on addressing these inherent fairness trade-offs?

A. ◯ Implement adversarial debiasing techniques that can simultaneously satisfy all fairness constraints through sophisticated multi-objective optimization.

B. ◯ Prioritize demographic parity over other metrics since it's the most intuitive fairness definition for non-technical stakeholders and regulators.

C. ◯ Develop a context-specific approach that explicitly measures multiple fairness metrics, documents inherent trade-offs, and selects metrics based on the specific application's ethical priorities and regulatory requirements.

D. ◯ Avoid group fairness metrics entirely in favor of individual fairness measures that don't exhibit these mathematical tensions.

12. A healthcare algorithm predicts which patients should receive additional care management. Analysis shows the following metrics across racial groups A and B: - True positive rate: Group A = 70%, Group B = 55% - False positive rate: Group A = 20%, Group B = 10% - Positive predictive value: Group A = 78%,

Group B = 85% Which technical approach most accurately addresses the specific fairness concerns evident in these metrics?

A. ◯ Apply demographic parity constraints during retraining to ensure both groups receive additional care at equal rates.

B. ◯ Implement post-processing with different thresholds for each group to equalize both true positive and false positive rates (equalized odds).

C. ◯ Implement equal opportunity constraints that focus specifically on equalizing true positive rates across groups, ensuring patients who truly need care have similar chances of receiving it regardless of their demographic group.

D. ◯ Apply calibration techniques to address the predictive value disparities, standardizing classification thresholds based on individual risk scores rather than group membership.

Go back