Types and Sources of Bias

Your score: 11/12

Go back

1. 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.

2. A company is developing an algorithmic lending system that will operate across multiple jurisdictions, including the United States and European Union. Which approach to protected attributes most accurately reflects the current legal consensus for ensuring compliance?

A. ◯ The system should be "blind" to all protected attributes, removing them from the data to prevent explicit discrimination.

B. ◯ The system should consider protected attributes exclusively during a post-processing fairness evaluation phase but not during model training.

C. ◯ The system should collect protected attribute data for testing disparate impact but may need different implementations across jurisdictions due to conflicting requirements regarding the use of protected attributes in decision-making.

D. ◯ The system should standardize on demographic parity across all protected attributes as the mathematical fairness definition that best satisfies all relevant legal frameworks.

3. What is the most accurate characterization of the relationship between different mathematical fairness definitions based on current impossibility theorems?

A. ◯ With sufficient data and computational resources, all desirable fairness definitions can be simultaneously satisfied through advanced multi-objective optimization techniques.

B. ◯ Mathematical impossibility results prove that certain combinations of fairness criteria cannot be simultaneously satisfied in most real-world scenarios, requiring explicit trade-off decisions.

C. ◯ The incompatibility between fairness definitions is primarily a practical implementation issue that can be resolved through better algorithm design rather than a fundamental mathematical limitation.

D. ◯ Fairness definitions are only incompatible when protected attributes are highly correlated with legitimate predictive features; in low-correlation scenarios, all fairness criteria can be simultaneously satisfied.

4. In implementing a government service eligibility verification system, which approach to organizational context assessment most effectively identifies potential deployment biases?

A. ◯ Compare the algorithm's predictions to historical human decisions to ensure the automated system maintains consistent patterns across demographic groups.

B. ◯ Analyze demographic disparities in the training data to identify and remove any variables that correlate with protected attributes before deployment.

C. ◯ Conduct a comprehensive assessment that examines how the system integrates with existing workflows, maps discretionary decision points where staff can override recommendations, evaluates institutional incentives that might influence override patterns, and tracks outcome disparities across different office locations and demographic groups.

D. ◯ Implement a technical transparency solution that explains the reasoning behind each system recommendation, ensuring staff understand the factors influencing automated decisions.

5. 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.

6. In developing a loan approval system, stakeholders disagree about appropriate fairness metrics. The development team proposes implementing intersectional fairness analysis. Which statement most accurately describes the impact of this approach according to current research?

A. ◯ Intersectional analysis will resolve stakeholder disagreements by identifying a universal fairness definition that protects all demographic subgroups simultaneously.

B. ◯ Intersectional analysis will increase fairness for all groups by ensuring that the model satisfies demographic parity across all possible subgroup combinations.

C. ◯ Intersectional analysis will reveal potentially hidden fairness disparities at demographic intersections that single-attribute analysis might miss, while

still requiring explicit trade-off decisions between competing fairness definitions.

D. ◯ Intersectional analysis will demonstrably reduce bias by removing all protected attributes and their proxies from the model to ensure colorblind fairness for all groups.

7. When implementing feedback loop monitoring for a recommendation system operating across 30 countries, a data scientist discovers that engagement disparities between socioeconomic groups are growing at significantly different rates across cultural contexts. Which technical approach to intersectional feedback analysis most accurately reflects current best practices?

A. ◯ Average disparity growth rates across all countries to produce a global fairness metric that can be monitored for overall feedback effects.

B. ◯ Focus monitoring exclusively on the countries showing the fastest disparity growth rates, as these represent the highest-risk contexts.

C. ◯ Implement hierarchical measurement that tracks disparity growth patterns both within individual countries and across country groups, with specific attention to how socioeconomic factors interact with cultural context to produce distinct feedback patterns.

D. ◯ Standardize recommendation algorithms across all countries to eliminate cultural variables that complicate feedback analysis.

8. 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.

9. A healthcare system deploys an AI triage tool to prioritize emergency department patients. Initial data shows that while the algorithm provides similar recommendations across demographic groups in testing, elderly patients and non-native English speakers wait significantly longer for care in actual deployment. Which technical approach most accurately addresses this deployment bias?

A. ◯ Retrain the algorithm with more data from elderly and non-native English-speaking patients to improve its performance for these groups.

B. ◯ Implement post-deployment monitoring that tracks the full sociotechnical workflow, including how staff interpret and act on AI recommendations across different patient groups, and adjust both the interface design and organizational implementation processes based on identified disparities.

C. ◯ Add a demographic calibration layer that automatically adjusts priority scores to ensure equal average waiting times across all demographic groups.

D. ◯ Create separate models for different demographic groups to account for their unique presentation patterns and interaction needs.

10. When developing a hiring algorithm that must balance multiple stakeholder perspectives on fairness, which approach most accurately reflects the current technical consensus?

A. ◯ Implement demographic parity constraints to ensure equal representation across all protected groups, as this is the most broadly accepted fairness definition.

B. ◯ Optimize for maximum prediction accuracy first, then apply post-processing adjustments to correct for any observed disparities across protected groups.

C. ◯ Conduct a structured analysis of context-specific factors—including historical discrimination patterns, stakeholder perspectives, and domain requirements—before selecting and prioritizing among potentially conflicting fairness definitions.

D. ◯ Select the fairness definition with the strongest legal precedent in employment law to minimize compliance risks, even if it does not address all stakeholder concerns.

11. 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.

12. A financial services company implements an AI-based loan application system accessible through both web and mobile interfaces. Analysis shows completion rates are 30% lower for applicants from lower-income neighborhoods

compared to wealthy areas, despite similar qualification rates. Further investigation reveals most lower-income applicants access the system via smartphones with limited data plans. Which sociotechnical approach best addresses this infrastructure disparity?

A. ◯ Develop a comprehensive multi-channel strategy with a data-efficient mobile interface, offline functionality, bandwidth detection with automatic adaptation, and alternative application methods including phone support and in-person options at community locations.

B. ◯ Create a separate, simplified application process specifically for applicants from lower-income neighborhoods to accommodate their technological limitations.

C. ◯ Launch an educational campaign to teach proper digital literacy skills to applicants from lower-income areas, enabling them to better navigate the existing application interface.

D. ◯ Implement an algorithmic correction that gives automatic preference to partially completed applications from lower-income neighborhoods to compensate for the completion rate disparity.

Go back