Historical & Societal Foundations of AI Fairness

Your score: 11/12

Go back

1. In developing a resume screening algorithm for technical roles, a team discovers that the mathematical encoding of educational background creates larger vector distances between international universities and elite U.S. institutions than between elite and non-elite U.S. institutions, despite similar educational quality. Applying the concept of codification of social categories, which approach most appropriately addresses this issue?

A. ○ Remove educational institution features entirely from the model to eliminate any potential bias based on educational background.

B. ○ Apply dimensionality reduction techniques like PCA to educational features to create a more compact representation that minimizes differences.

C. ○ Develop an encoding approach informed by educational quality metrics rather than institutional prestige, validate this encoding through outcomes analysis across demographic groups, and incorporate uncertainty measures for institutions with limited representation in the data.

D. ○ Standardize all educational institutions to binary values indicating only whether the candidate has the required degree level, regardless of institution.

2. A financial institution is implementing a machine learning model to predict loan default risk. During testing, they discover the model produces higher average risk scores for applicants from historically redlined neighborhoods despite excluding ZIP codes and racial data from the feature set. Which technical mechanism most accurately explains this finding?

A. ○ Representation bias in the dataset, as applicants from redlined neighborhoods are numerically underrepresented in the training data.

B. ○ Proxy discrimination, where variables like income, debt ratios, and credit history correlate with neighborhood due to historical disinvestment in redlined areas.

C. ○ Technical instantiation of historical bias, where the model explicitly incorporates historical redlining maps into its prediction algorithm.

D. ○ Feedback loop amplification, where previous loan denial decisions have created progressively stronger correlations between neighborhood and default risk.

3. When implementing fairness interventions in a lending algorithm deployed across multiple cultural contexts, which approach to ethical frameworks most accurately reflects current best practices?

A. ○ Apply universal ethical principles consistently across all contexts to ensure standardized fairness implementations regardless of cultural differences.

B. ○ Conduct an analysis of local ethical traditions in each deployment context, adapt fairness definitions to align with these traditions where they differ from Western frameworks, and document how these adaptations address context-specific historical patterns.

C. ○ Allow each local implementation team to select whatever ethical framework seems most appropriate for their context, without requiring documentation or comparative analysis.

D. ○ Identify which ethical framework has the strongest philosophical foundation according to academic consensus, then apply it consistently across all contexts.

4. When implementing a historical pattern analysis for an automated resume screening system in hiring, which technical approach most effectively addresses intersectional considerations according to current research?

A. ○ Analyze performance disparities separately for each protected attribute, then combine these independent analyses to understand the system's overall fairness properties across demographic groups.

B. ○ Select the single most historically significant demographic attribute for the specific job sector based on documented discrimination patterns, and focus the analysis on disparities along this primary dimension.

C. ○ Design a comprehensive analysis that examines performance patterns across all possible combinations of protected attributes, treating each unique combination as an independent classification requiring equal analytical attention.

D. ○ Conduct targeted analysis of specific demographic intersections where research indicates unique historical discrimination patterns have operated, examining how these specific patterns might manifest differently in the current system from what would be predicted by single-attribute analysis.

5. Which of the following most accurately characterizes how historical discrimination patterns persist across technological transitions according to current scholarship?

A. ○ Historical biases typically diminish as technologies advance, with newer computational systems inherently reducing discrimination through increased precision and objectivity

B. ○ Discrimination patterns transform but persist across technological transitions, often becoming encoded in new technologies through problem formulations, data practices, and optimization choices that reflect existing social hierarchies

C. ○ Historical patterns are primarily relevant to legacy systems but largely eliminated in modern AI through technical advances like feature selection algorithms and regularization techniques

D. ○ Technological transitions create disruptions that typically reset discrimination patterns, with biases in new technologies emerging independently rather than inheriting historical patterns

6. When developing the ethical evaluation component of a Historical Context Assessment Tool, which approach most effectively incorporates intersectional considerations?

A. ○ Analyze historical patterns separately for each protected attribute (race, gender, etc.), then combine these analyses to create a comprehensive evaluation.

B. ○ Focus on the most historically marginalized single demographic group for each application domain to ensure the most significant historical patterns receive attention.

C. ○ Design the evaluation framework to examine how ethical implications differ at demographic intersections, prioritizing impacts on multiply-marginalized groups and documenting how discrimination operates differently across intersecting identities.

D. ○ Apply a mathematical weighting system that assigns higher importance to attributes with stronger historical discrimination patterns to ensure appropriate prioritization.

7. A healthcare system is implementing an algorithm to identify patients for preventive interventions. During ethical evaluation, stakeholders disagree about whether resources should maximize aggregate health improvements across the population or prioritize historically underserved communities with greater health needs. Which approach most effectively addresses this ethical tension?

A. ○ Select the consensus ethical framework that most stakeholders can agree on and implement it consistently.

B. ○ Determine which ethical framework is most closely aligned with scientific evidence about health outcomes and prioritize that perspective.

C. ○ Conduct a multi-framework analysis that documents how different ethical perspectives would evaluate the system, then develop a blended approach that explicitly acknowledges trade-offs and gives priority to addressing the most severe historical health disparities.

D. ○ Implement a technical solution that optimizes for the mathematical mean between competing fairness metrics, allowing the algorithm to find a statistically optimal balance.

8. A data scientist is developing a natural language processing system for criminal justice risk assessment and notes that the training data contains offense descriptions using different terminology patterns for defendants from different racial backgrounds, with some actions described as "aggressive" versus "assertive" in ways that correlate with race. Which approach best applies the concept of classification politics to address this issue?

A. ○ Remove all subjective terminology from the dataset, restricting the model to objective numerical features like prior offense counts.

B. ○ Apply uniform terminology transformations that standardize all language regardless of context to ensure computational consistency.

C. ○ Examine the historical context of these terminological differences, analyze the implicit classification systems they represent, and either reframe descriptions to eliminate bias-laden terms or develop representations that explicitly account for these documented patterns.

D. ○ Average the language patterns across groups to create a balanced representation that incorporates all perspectives equally.

9. A content recommendation system for educational materials significantly increases engagement metrics after deployment. However, analysis reveals that users from underrepresented groups are progressively receiving less diverse content recommendations over time, with their recommendations increasingly differing from those of majority users with similar stated interests. Which technical mechanism best explains this emergent pattern?

A. ○ Technical instantiation of historical bias, where historical patterns in educational material development influence initial recommendation quality across groups.

B. ○ Proxy discrimination, where seemingly neutral engagement metrics actually correlate with user demographics through historical usage patterns.

C. ○ Feedback loop amplification, where initial small differences in recommendations lead to divergent engagement patterns that progressively intensify through the recommendation algorithm.

D. ○ Domain-specific manifestation in content systems, where educational materials inherently resonate differently with different demographic groups.

10. A healthcare company is developing an AI system to prioritize patients for specialized care, trained on historical treatment data. Which approach most effectively applies historical pattern analysis to this scenario according to current research?

A. ○ Conduct a technical analysis of model accuracy across demographic groups while avoiding historical considerations that might introduce subjective biases into the assessment process.

B. ○ Research whether the training data includes diverse demographic representation and ensure protected attributes are removed from model inputs to prevent historical biases from influencing predictions.

C. ○ Conduct a focused historical analysis of how medical technologies and classification systems have historically classified and measured disease across demographic groups, examining how these patterns might manifest in the current system through problem formulation, feature selection, and outcome definition.

D. ○ Apply industry-standard fairness metrics like demographic parity to ensure the system meets established benchmarks, as these metrics already incorporate relevant historical considerations.

11. When analyzing a healthcare dataset for development of a diagnostic support system, a data scientist discovers that certain symptoms are systematically documented in less detail for elderly patients from lower socioeconomic backgrounds compared to other patients with similar conditions. Which technical approach most effectively addresses this strategic ignorance in the dataset?

A. ○ Drop all records with lower documentation detail to ensure consistent data quality across all patients in the training data.

B. ○ Treat missing or less detailed information as statistically random events that can be addressed through standard imputation methods.

C. ○ Develop a model that explicitly accounts for documentation patterns as informative features rather than random noise, uses uncertainty quantification for predictions involving underrepresented patterns, and incorporates validation approaches that test model performance specifically on underrepresented groups.

D. ○ Generate synthetic data to replace all records for elderly patients from lower socioeconomic backgrounds to eliminate documentation disparities.

12. A hospital implements an algorithm to identify patients who would benefit from enhanced care management programs. The system uses historical healthcare costs as a proxy for medical need, as this data is readily available in billing records. When analyzed by race, the algorithm assigns significantly fewer Black patients to enhanced care programs compared to white patients with the same medical conditions. What domain-specific historical pattern best explains this disparity?

A. ○ Racial bias among doctors leading to intentional undertreatment of Black patients, with this discriminatory intent now embedded in the algorithm.

B. ○ Historical differences in disease prevalence between racial groups, causing the algorithm to correctly identify different care needs.

C. ○ Historical underrepresentation of Black patients in electronic health record systems, creating data quality issues that affect prediction accuracy.

D. ○ Structural inequities in healthcare access and utilization resulting in lower historical healthcare spending for Black patients with the same medical needs as white patients.

Go back