Historical Context Assessment Tool (HCAT) with Implementation Frameworks

1. Historical Patterns of Discrimination

Purpose: To examine how previous technologies have reinforced, challenged, or transformed social hierarchies, especially in domains such as lending, healthcare, or education.

Implementation Framework:

1. Historical Pattern Identification:

- Examine past discrimination in the domain (e.g., consumer lending).

- Investigate how earlier technologies reflected existing hierarchies.

- Identify recurring mechanisms of social exclusion.

2. Pattern-to-Risk Mapping:

- Link historical patterns to components of the current ML system.

- Analyze influence on feature definitions and encodings.

- Review accuracy disparities across groups.

3. Prioritization Framework:

- Rate the connection strength between past patterns and the system.

- Assess visibility and harm potential of each bias.

- Use a prioritization matrix to guide fairness evaluations.

2. Data Representation and Social Context

Purpose: To understand the social context of data creation, identify underrepresentation, and evaluate how categories and classification influence fairness.

Implementation Framework:

1. Classification Audit:

- Document protected attributes and proxies.

- Research historical development of classifications.

- Evaluate inclusivity and relevance.

2. Representation Gap Analysis:

- Compare dataset demographics with population benchmarks.

- Identify data deserts and missingness by group.

- Validate proxy variable meaning across groups.

3. Codification Evaluation:

- Examine how features are encoded (e.g., one-hot, distance metrics).

- Assess assumptions embedded in technical encoding.

- Document flattening of complex identities.

4. Power Analysis in the Data Lifecycle:

- Analyze who designed, collected, labeled, and validated the data.

- Investigate the context of data collection (e.g., coercion vs. consent).

- Assess representativeness controls and accountability.

3. Ethical Frameworks for Fairness Evaluation

Purpose: To assess fairness through multiple ethical lenses and identify how different stakeholder values influence judgments.

Implementation Framework:

1. Multi-framework Analysis:

- Use consequentialist (utilitarian, prioritarian), deontological, and virtue ethics.

- Apply them to data sourcing, feature design, model evaluation.

2. Stakeholder-Centered Ethical Mapping:

- Identify which ethics align with which stakeholders.

- Map potential value conflicts across groups.

3. Contextual Adaptation:

- Modify framework usage based on historical harms and cultural context.

- Identify domain-specific ethics gaps.

4. Trade-off Documentation:

- Create transparent documentation of trade-offs.

- Provide rationale for prioritizing certain ethics in each case.

- Establish feedback and review loops for ethical re-assessment.

4. Modern Manifestations of Historical Biases

Purpose: To identify how technical choices today instantiate legacy biases, including proxies and feedback loops.

Implementation Framework:

1. Map Historical Patterns to Technical Mechanisms:

- Identify technical design decisions tied to past discrimination.

- Review architecture, feature sets, or rules with historical links.

2. Conduct Proxy Variable Analysis:

- Measure correlation of variables with protected attributes.

- Identify variables with dual utility and discriminatory potential.

3. Perform Feedback Loop Assessment:

- Investigate how predictions reinforce biased data collection.

- Trace cycles of self-confirming system behavior.

4. Domain-Specific Analysis:

- Use precedents from domain-specific case studies (e.g., FICO scores, COMPAS, etc.).

- Adapt techniques to domain-specific risks and history.