Historical Context Assessment Tool (HCAT)

This tool enables engineering and AI teams to assess historical discrimination patterns, data representation issues, system-level amplification risks, and ethical tradeoffs relevant to the design, development, and deployment of AI systems.

# Section 1: Domain and Application Context

**1.1 Application Domain Identification**

• What specific domain will this system operate in (e.g., lending, hiring, healthcare)?

• What function will it serve within this domain?

**1.2 Historical Discrimination Patterns**

• What documented patterns of discrimination exist in this domain (e.g., redlining, exclusion)?

### Pattern-to-Risk Mapping Table

|  |  |  |  |
| --- | --- | --- | --- |
| Historical Pattern | System Feature Affected | Encoded Social Categories | Risk |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**1.3 Technical Stratification Analysis**

• How have prior technologies contributed to or challenged social stratification?

• Are some populations treated as 'edge cases' or excluded from optimization?

**1.4 Selective Optimization**

• Are some groups disproportionately prioritized in success metrics or optimization logic?

# Section 2: Data and Representation Analysis

**2.1 Historical Data Sources**

• What dataset(s) will inform the system?

**2.2 Categorical Formation**

• How were key categories in the dataset(s) historically defined? How were they sourced, and under what context? What ethical risks do they pose, and what actions can be taken to mitigate risks.

### Risk Mapping Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Source of Bias | Affected Population | Ethical Risk | Action Taken |
| Race | Census legacy classification | Multiracial individuals | Misrepresentation | Recode with expanded categories |
| Gender | Binary-only input | Non-binary people | Exclusion | Add third-category or free text option |
|  |  |  |  |  |
|  |  |  |  |  |

**2.3 Strategic Ignorance and Missing Data**

• What groups or dimensions are missing or underrepresented in data?

### Representation Gap Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Population Share (%) | Dataset Share (%) | Gap (%) | Observation and Notes (Are these gaps deliberate or due to neglect?) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**2.4 Politics of Classification**

• Who defined the categories? Have definitions evolved over time?

**2.5 Power Asymmetries in Data Infrastructure**

• Who controls data collection and access? Do the affected have the ability to seek redress?

### Power Dynamics Mapping Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | Controller/Actor | Power Asymmetry | Ethical Risk | Ability to seek redress |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Section 3: Historical Bias Propagation, Data Representation, and Mitigation Matrix

Use this matrix to evaluate risks arising from historical patterns (Section 1) and data representation (Section 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Historical Bias** | **Data Representation** | **System Feature Affected** | **Likelihood** | **Severity** | **Ethical Risk Type** | **Mitigation Feasibility** | **Priority** |
| Redlining | ZIP code, home ownership | Loan thresholds | High | High | Deontological, Egalitarian | Medium | Critical |
| Gender bias | Job status, marital data | Credit score | Medium | High | Prioritarianism | Low | High |
| Informality bias | Missing income data | Feature completeness | High | Medium | Utilitarianism | Low | Medium |

# Section 4: Modern Manifestations of Historical Biases

**4.1 Historical to Technical**

How do historical patterns of bias in your context manifest today?

### Historical-to-Technical Mapping Table

|  |  |  |
| --- | --- | --- |
| Historical Pattern | Modern Mechanism | Feature or Process |
| Racial housing bias | Geo-proxies | ZIP code, Rent history |
| Gender employment disparity | Income gap | Job classification |
|  |  |  |
|  |  |  |

### 4.3 Proxy Discrimination Checklist

Identify and evaluate risks of proxy discrimination, where seemingly neutral variables act as stand-ins for protected attributes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Proxy For | Correlation Strength | Predictive Value | Action Taken |
| ZIP Code | Race | High | Medium | Replaced with distance to amenities |
| Marital Status | Gender | Moderate | Low | Removed |
| Employment Gap | Age | Low | High | Retained, monitored |

**4.4 Technological Amplification and Feedback Loop Mapping**

• How might automation amplify existing bias? Could system outputs influence future inputs (e.g., credit limit changes)?

### Feedback Loop Risk Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Loop Type | Entry Point | Amplification Path | Risk Severity |
| Credit scores -> Loan approval -> Income recording | Biased score | System perpetuates exclusion | High |
| Geo-location -> Model trust score -> Outreach | Urban bias | Rural borrowers excluded | Medium |

# Section 5: Ethical Framework Alignment Matrix

Use the following matrix to evaluate how various ethical frameworks may interpret fairness in development of the system.

### Multi-Framework Analysis Table

|  |  |  |  |
| --- | --- | --- | --- |
| Ethical Framework | Fairness Test | Design Implications | Example in Your Context |
| Utilitarianism |  | Maximize total benefit | Optimize access without sacrificing accuracy |
| Prioritarianism |  | Prioritize worst-off groups | Adjust models for excluded populations |
| Egalitarianism |  | Promote equality of outcomes | Use re-weighting to ensure fairness |
| Deontological |  | Adhere to rights and duties | Avoid unfair exclusion even if effective |
| Virtue Ethics |  | Foster integrity and care | Embed empathy and responsibility in design |

Have the system ethical considerations been reviewed with input from affected populations?

### Stakeholder Mapping Table

|  |  |  |
| --- | --- | --- |
| Stakeholder Group | Ethical Preference | Conflict Zone |
| Borrowers | Prioritarianism | Model performance trade-offs |
| Executives | Utilitarianism | Risk tolerance vs fairness |
|  |  |  |
|  |  |  |

### Decision Table

Use this matrix to guide ethical decision-making by clearly outlining trade-offs between competing values observed in Stakeholder Mapping Table. Each option should be evaluated based on its benefit to fairness, potential ethical cost, and alignment with chosen ethical principles.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Option | Description | Ethical Benefit | Ethical Cost | Ethical Principle Prioritized | Final Choice? |
| A | Broad coverage with looser thresholds | Maximizes reach, promotes inclusion of underserved groups | May reduce precision, increasing false positives | Prioritarianism (uplift marginalized) | ✓ |
| B | Strict thresholds with high precision | Increases accuracy and minimizes error for included groups | Excludes edge cases, reinforces existing disparities | Utilitarianism (maximize efficiency) | ✗ |

**Summary of decison**:

* **Option A** was chosen because its ethical benefit—greater inclusivity for historically underserved communities—was prioritized despite a trade-off in precision.
* **Option B**, while technically appealing for its accuracy, was rejected due to its ethical cost of reinforcing systemic exclusion.