The AI Reliability & Availability Risk Assessment
Quantifying Algorithmic Bias as a Single Point of Failure
Prepared For: Executive Leadership / Risk Management / MLOps Teams
Objective: To calculate "Functional Downtime" caused by demographic bias and assess the risk of Algorithmic Disgorgement (Asset Deletion).

[Page 1: Executive Summary & The Core Concept]

The 99.9% Lie

Most organizations measure AI availability based on server uptime. If the API is responding, the system is considered "Up."

However, if an AI model systematically rejects valid inputs from a specific demographic due to algorithmic skew (bias), the system is effectively **offline** for that segment of your user base. Furthermore, regulatory bodies (FTC, EU) are increasingly using "Algorithmic Disgorgement"—forced model deletion—as a penalty for biased data practices.

Bias is not just an ethical issue. It is an availability risk.

Use this assessment to:

1. Calculate your **True Demographic Uptime**.
2. Score your risk of **Regulatory Asset Deletion**.
3. Identify early warning signs of **Model Collapse**.

[Page 2: The Calculator]

Tool 1: The Demographic Uptime Calculator

Instructions: Use current performance metrics to determine if your model meets functional availability standards for all user groups.

Step A: Input Metrics

1. **Standard Server Uptime:** _____ % (e.g., 99.9%)
2. **Targeted Demographic (Group A):** _____ (e.g., Users in Zip Code 90210)
3. **Traffic Share of Group A:** _____ % (Percentage of total volume)
4. **Disparate Rejection Rate:** _____ % (How much higher is the rejection/failure rate

for Group A compared to the baseline?)

Step B: The Formula

Functional Availability is lower than Server Availability.

$$True Availability = Server Uptime - (Traffic Share \times Disparate Rejection Rate)$$

Step C: Worked Example

- **Server Uptime:** 99.9%
- **Group:** Mobile Android Users (30% of traffic)
- **Bias Error:** Model fails/rejects Android users 20% more often than iOS users.

Calculation:
$99.9 - (0.30 \times 0.20)$
$99.9 - 0.06 = 93.9\%$
Result: Your "99.9%" system actually has **93.9% Availability**. This likely breaches your Master Service Agreement (MSA) or SLA.

Step D: Your Calculation

[_____] - ([_____] x [_____]) = _____ % True Availability

[Page 3: The Risk Audit]

Tool 2: The Disgorgement & Deletion Risk Audit

Instructions: Rate your current model on a scale of 1 (Low Risk) to 5 (High Risk). A score above 15 indicates a high probability of regulatory intervention or "Kill Switch" necessity.

1. Data Provenance

- (1) Fully licensed, consented data.
- (5) Web-scraped data without explicit consent (e.g., "Everalbum" Scenario).
- **Score:** [___]

2. Model Interpretability

- (1) Fully explainable (Decision Trees, Linear Regression).
- (5) Deep Neural Net "Black Box" with no post-hoc explainability tools (SHAP/LIME) implemented.
- **Score:** [___]

3. Proxy Variable Reliance

- (1) Protected attributes (Race/Gender) explicitly excluded and tested for proxies.
- (5) Model relies on high-correlation proxies (Zip Code, Surname, Device Type) without debiasing.
- **Score:** [___]

4. Synthetic Data Loop (Model Collapse Risk)

- (1) Trained on 100% human-generated, verified ground truth.
- (5) Trained on >40% synthetic data generated by previous AI versions.
- **Score:** [___]

5. Insurance Status

- (1) AI Liability policy active with "Bias" coverage included.
- (5) Uninsured or standard Cyber policy with "AI/Algorithm" exclusions.
- **Score:** [___]

TOTAL RISK SCORE: [_____ / 25]

- **0-10:** Low Risk (Monitor Annually)
- **11-15:** Moderate Risk (Implement Drift Detection immediately)
- **16-25:** Critical Risk (High probability of Asset Deletion/Disgorgement)

[Page 4: The Mitigation Matrix]

From "Unfair" to "Unavailable": A Reference Guide

Use this matrix to categorize bias bugs in your Incident Response (IR) platform (e.g., Jira, ServiceNow).

Risk Category	The Bias Indicator	The Availability Consequence	Recommended Action
Regulatory	Disparate Impact Ratio < 0.8 (4/5ths Rule)	**Asset Deletion:** Regulator orders model destruction (Disgorgement).	**Immediate Freeze:** Stop deployment. Conduct Legal Privilege Audit.
Operational	High False Positive Rate for specific	**Functional Outage:** Service is	**Circuit Breaker:** Route high-risk

	demographic	effectively "down" for that group.	demographic traffic to human review.
Technical	Decreasing output variance / High Perplexity	**Model Collapse:** Model quality degrades to usability zero.	**Data Refresh:** Retrain immediately on pure human-generated data.
Financial	High correlation of error rate to protected class	**Uninsurability:** Carriers deny liability coverage.	**Red Teaming:** Commission 3rd party adversarial stress test.

[Page 5: Next Steps]

Moving Forward

If your Risk Score is above 15 or your True Availability is below your SLA:

1. **Update your Risk Register:** Move "Algorithmic Bias" from the "Ethics" column to the "Availability/Continuity" column.
2. **Implement Observability:** Install drift detection tools (e.g., Arize, Fiddler, or open-source equivalents) to alert on demographic performance drops in real-time.
3. **Red Team Your Model:** Don't just audit for compliance; stress-test the model to see if adversarial inputs cause it to fail.