

Cover

The Red Team Lab Workbook

Algorithmic Red Team Simulation — Student Edition

How to Use This Workbook

This workbook accompanies the Civic Sandbox simulation. Work through each exercise in order, record your slider inputs, and compare your Algorithmic Integrity Score with your peers.

Exercise 1: Baseline Calibration

Set all four sliders to 50. Record the baseline score. What does the diagnostic panel tell you about this default configuration?

Exercise 2: Demographic Weighting

Slowly raise the Demographic Weighting slider. Note the score at 25, 50, 75, and 100. At what point does the model flag a protected-class concern?

Exercise 3: Historical Training Bias

Adjust Historical Data Training Bias and observe the Bias Index. Why does this lever have the largest weight in the scoring formula?

Exercise 4: Exception Handling

Lower Exception Handling Rigor. What happens to the diagnostic findings? Describe a real-world civic system where a null value caused harm.

Exercise 5: Transparency

Raise Transparency to maximum and lower it to 10. How does Auditability correlate with Public Trust in the matrix?

Exercise 6: Scenario A

Load Scenario A: The Silent Voucher Error. Identify the root cause and propose a policy fix that does not require more data.

Exercise 7: Scenario B

Load Scenario B: The Biased Predictive Intake. Which sliders move? Which stakeholders should be consulted before retraining?

Exercise 8: Mixed Failure

Create a configuration where at least two findings are active at once. Write a one-paragraph incident report as if you were the oversight auditor.

Reflection Questions

1. Who is most likely to be harmed when a system scores poorly?
2. What checks should be required before a city deploys an automated intake tool?
3. How should community members be able to challenge an algorithmic decision?

Glossary

Algorithmic Integrity Score — A composite 0-100 measure of fairness, robustness, and accountability.

Bias Index — Average of demographic weighting and historical training bias.

Auditability — The degree to which a decision can be inspected and appealed.