CIS 350: AI for Cybersecurity

Week 16 Activity: Responsible AI Governance Board Meeting

CISA-Style Governance Tabletop Exercise (Group Worksheet)

Group members (names): _____

Role assignments: Incident Commander _____ Compliance/Legal _____ Technical Lead _____ Communications _____

Activity Overview

Your group is the Responsible AI Governance Board at MeridianBank. A data-science team has built three candidate AI models to decide credit-card applications, and you must approve exactly ONE (or approve none). You will play three timed rounds of about 5 minutes each. In every round you record a decision AND a justification based on risk, cost, or evidence. This is a CISA-style tabletop exercise: no real customer is affected, so you can rehearse the hard call safely.

Setup: Frameworks You Will Use

Before you begin, skim these two references so your team speaks the same language:
- **CISA Tabletop Exercise Packages** (cisa.gov) - the exercise format used today
- **NIST AI Risk Management Framework** (nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf)

Score every model against the four Responsible-AI principles: **Transparency, Fairness, Robustness, Accountability.** The highest business accuracy does NOT automatically win.

Scenario Briefing

MeridianBank wants an AI model to decide who gets a credit card, their limit, and their interest rate. The three candidates are:
- **Model A - High Accuracy (92%):** most profitable, but has a measured fairness gap between groups that regulators may object to.
- **Model B - Privacy-Preserving (88%):** uses differential privacy to protect customer data even if the model is stolen or probed; gives up some accuracy.
- **Model C - Explainable (89%):** every decision comes with a plain-language reason and is easy to audit, but inference is slower and costlier.

The shadow over the room: two years ago the bank deployed an AI hiring tool that quietly favored one group. It led to a regulator fine, bad press, and lost trust. The board was blamed for approving a model nobody could explain or monitor. Let that history shape your risk appetite.

Warm-Up: Predict First (~2 min)

Before any discussion, each member privately picks the model they THINK the board should approve and the ONE RAI principle (transparency, fairness, robustness, accountability) they expect to matter most. Then share around the table in one sentence each. Do not debate yet - just surface where the board starts.
My prediction (model + principle): _____

Round 1: The Initial Decision (~5 min)

Situation: The board must pick Model A, B, C, or approve none - today. Which model best serves customers AND the business?

Decision space to weigh:
- Model A: most profit, but a fairness gap regulators may punish
- Model B: strong privacy, gives up some accuracy
- Model C: fully explainable and auditable, but slower and costlier
- Approve none: is any model good enough to deploy?

Team decision (circle or write one): A / B / C / None

Justification - which RAI principle drove this choice, and what risk, cost, or evidence supports it?

Round 2: A Complication Is Injected (~5 min)

New information arrives mid-meeting:
- The Technical Lead reveals Model A's fairness gap resembles the OLD hiring incident.
- The measured false-positive-rate (FPR) gap between groups is about **0.16**.
- A mitigation exists that shrinks the gap to about **0.01** - but it drops accuracy by 3%.
- Regulators have signaled they will audit any high-impact credit model this year.

Revisit your Round 1 decision. Does it still hold? You may switch models, apply the mitigation, or stand firm - but you must justify it.

Updated team decision: _____

What changed (or did not) and why? Reference the 0.16 gap, the mitigation trade-off, or the regulator signal.

Round 3: Final Decision and Documentation (~5 min)

Lock your FINAL model choice. Then mandate the controls that must exist AFTER approval: monitoring, auditing, and a named accountable owner.

Final model approved: _____ (or approve none, and say why)

Required post-approval controls (fill in each):
- Monitoring metric + how often + alert threshold:

- Audit process (who reviews, how often, what they check):

- Accountable owner (who is responsible if the model harms customers):

Audit log entry - write one factual record line for the file (date, decision, rationale, who approved):

Stakeholder communication - in 2-4 sentences, explain the decision to customers/regulators in plain language (no jargon):

Debrief Reflection

Q1: What changed your decision between Round 1 and Round 3, and why?

Q2: How did the prior hiring-bias incident shift your board's risk appetite?

Q3: What governance gap did this exercise reveal - what monitoring or control would have caught the OLD incident before it went public? Connect your answer to the labs: fairness (Lab 10, the 0.16 to 0.01 FPR gap), privacy (Lab 11, epsilon 0.1 vs 8), robustness (Lab 8, adversarial examples), and incident response (Lab 12).

Grading Rubric (20 points)

Criterion	Description	Points
Completion	Participates in all 3 rounds; every decision recorded	8
Reasoning	Decisions justified with risk, cost, or evidence	7
Communication	Clear stakeholder message and audit documentation	5

Total: 20 points

Scoring: Completion (8) + Reasoning (7) + Communication (5) = 20 points

How to Submit

1. Complete all three rounds on this group worksheet
2. Record each decision AND its justification
3. Write the final stakeholder statement and audit log entry
4. Complete the debrief reflection
5. Submit one PDF worksheet per group on Canvas by [DUE DATE]

Questions? Post in the course forum or attend office hours!