# Case Study:

# “SafeGuard AI” Risk-Mitigation Platform

**Background:**

FinHealth Analytics, a mid-market insurtech firm, has developed **SafeGuard AI**, an AI-driven tool that continuously monitors transactional and compliance data for community banks. Its goal is to detect emerging operational, credit, and regulatory risks—and surface them to executives, relationship managers, and regulators in real time.

---

## 1. Model Transparency

- 🪪 **Model Passport**
  • A living document recording data sources, preprocessing steps, feature sets and model versions.
- 🏷️ **Access-Controlled Model Cards**
  • Three tiers: Board summary (impact & high-level metrics), Regulator dossier (compliance KPIs), Engineering spec (hyperparameters & lineage).
- 🔍 **Open-Box Audits**
  • Internal white-box code reviews every sprint.
  • Quarterly third-party “red team” stress tests against adversarial scenarios (e.g., data drift, spoofed inputs).

## 2. Hallucination Management

- 📚 **Retrieval-Augmented Generation (RAG)**
  • All regulatory interpretations are anchored to an internal “Reg-Code” knowledge base; free-text advice outside that base is flagged with an uncertainty score.
- ⚙️ **Dynamic Calibration Layers**
  • A lightweight fact-checker vets any “novel” compliance recommendation against structured rule engines—either auto-correcting or annotating doubtful outputs.
- 🔄 **Continuous Fine-Tuning**
  • Monthly retraining on a curated set of past hallucinations plus expert-approved corrections to shrink error rates over time.

## 3. Trust-Building with Stakeholders

- **Board of Directors**
  • **Risk-Quantified Dashboards** showing “Most-Likely” vs. “Worst-Case” loss scenarios.

• **Governance Playbook** detailing escalation paths, approval rights, and audit cadence.

- **Customers (Community Banks)**
  • Quarterly **Transparency Reports** on model accuracy, uptime, and known limitations.
  • **Consent & Control Portal** allowing banks to opt into data-sharing tiers and view how their data is used.
- **Regulators**
  • **Compliance Blueprints** mapping each SafeGuard AI component to specific supervisory guidelines (e.g., Basel III, FFIEC).
  • **Standardized Disclosure Templates** (IEEE P7001-compatible) for rapid audit and licensing.

# 4. Explainability Techniques for Non-Technical Leaders

- ✏️ **Counterfactual Narratives**
  “If a borrower’s liquidity ratio had been 5 % lower, SafeGuard AI would have flagged the loan for manual review.”
- 📊 **Feature-Impact Charts**
  Simple bar/waterfall visuals showing the top three drivers of each risk score (e.g., transaction volatility ↑40 % → risk ↑0.2).
- 🗂️ **Layered Dashboards**
  • **Tier 1:** Plain-English summary (“Why we flagged this account”).
  • **Tier 2:** Drill-down technical details (model weights, data snapshots).
- 🤝 **Analogy-Driven Briefs**
  “Think of SafeGuard AI as a virtual risk committee: each feature votes, and the model aggregates to decide ‘raise hand’ or ‘stay silent.’”
- 🎛️ **Interactive “What-If” Simulators**
  Sliders for key inputs (e.g., credit utilization) let execs see real-time shifts in the risk score.

---

# Participant Exercises

## A. Discussion Questions

1. **Gap Analysis:** Which Responsible AI pillar is most at risk in SafeGuard AI’s current design—and why?
2. **Enhancement Proposal:** Suggest an additional hallucination-management control that could further reduce false positives.
3. **Stakeholder Strategy:** How would you tailor communications to a skeptical regulator versus a growth-focused bank CEO?

## B. Key Deliverables

1. **Draft Model Passport Template** covering data lineage, algorithm versions, and key performance metrics.
2. **Risk-Quantified Dashboard Mock-up** for the Board, with “Most Likely”/“Worst-Case” scenario panels.
3. **Governance Playbook Outline** specifying roles, sign-off workflows, and audit schedules.
4. **Explainability Brief** (one-page) using counterfactuals and feature-impact charts designed for non-technical executives.

---

Please work in small teams to map each deliverable to the corresponding branches of the Responsible AI mind-map and be prepared to present your solutions.