CIS 350: AI for Cybersecurity

Week 7 Activity: When to Request Manual Labels

CISA-Style Incident Tabletop Exercise (Team Worksheet)

Team members: _____

Roles: Incident Commander _____ Tech Lead _____

Communications Lead _____ Legal/Compliance _____ Scribe _____

Activity Overview

In this tabletop exercise your team acts as the incident-response group at NorthBank, an online bank whose fraud-detection AI has started missing real fraud and flagging good customers. The model needs better training data - better labels. A label is simply a confirmed correct answer: 'this transfer WAS fraud' or 'this WAS legitimate.' You must decide how to get those labels under time pressure, knowing you cannot have cheap, fast, AND perfect all at once. You will play through three timed rounds, record every decision with its justification, and produce an audit artifact a regulator could trust.

Task 1: Set Up the Tabletop

Resource: CISA Tabletop Exercise Packages (cisa.gov/resources-tools/services/cisa-tabletop-exercise-packages)

A tabletop exercise is a facilitated discussion of a realistic but fake incident - no real systems are touched. The value is in the decisions and the discussion. Before you begin: assign the five roles above, choose a Scribe to record decisions, and agree that each round ends with a committed team decision plus a one-sentence justification for RISK, COST, and EVIDENCE.

The Two Paths (your core trade-off)

```
PATH A - Manual labeling: analysts confirm each case by hand.
Cost: about $10,000 and 2 weeks. Result: clean, trustworthy labels.
PATH B - Weak supervision: quick rules auto-label many cases at once.
Cost: about 1 day and little money. Result: fast but noisy, error-prone labels.
```

Warm-Up (2 min): Quick Prediction

Before you play, each person makes a gut call and the team writes down its first instinct. Which path would you pick RIGHT NOW - A, B, or C - and in one word, what worries you most about it? You will compare this to your final Round 3 decision in the debrief, so be honest about your starting instinct.

Our first instinct (circle A / B / C) and the one thing that worries us most:

Round 1: The First Call

Situation: Fraud losses are rising now. The board wants action by end of day. You must choose how to get better labels for the model immediately.

Decision space (weigh these options):
A) Commit to manual labeling - clean data, but fraud runs free for 2 weeks.
B) Commit to weak supervision - live today, but you trust noisy labels.
C) Run both in parallel - weak labels now, manual labels replace them later.

Team decision (circle A / B / C and explain):

Justification - write one sentence for each:
RISK (worst outcome, who gets hurt if we are wrong):

COST (money and time - can the business survive it?):

EVIDENCE (what we know vs. what we are assuming):

Round 2: A Complication Is Injected

New information (mid-incident): Your spot-check finds the weak-supervision rules mislabeled about 30 percent of cases. Many legitimate immigrant customers were wrongly tagged as fraud - a bias problem. A local news reporter is now asking why good customers are being frozen out.

Decision: Do you keep the weak labels running, pause them, or add a human check before any account is frozen? Consider reputation risk, fairness, and the cost of pausing protection entirely.

Revised team decision:

What specifically changed your mind since Round 1? (name the new risk, cost, or evidence)

How will you detect the bad labels? (e.g., random spot-check sample size, rule disagreement, confidence gap, watching whether real fraud still slips through)

Round 3: Final Decision and Documentation

Situation: The 2-week manual labeling effort is now underway and partly done. Decide your final labeling strategy going forward: which cases still get weak labels, which require a human, and when you will re-train the model.

Final labeling strategy:

Audit artifact - write a short record an outsider could trust. Include: (1) what you decided and when, (2) why (risk, cost, evidence), (3) what safeguards protect customers from bad labels.

Stakeholder message. In 3-4 plain-language sentences, write what you would tell customers and the reporter - honest, clear, and without technical jargon.

Task 2: Team Debrief

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

Q2: What would you document so the next team could defend this decision to an auditor?

Q3: What governance gap did this reveal? (For example: no label-quality check existed before, or no rule for when a human must confirm.)

Task 3: Connect to Lab 6 (Semi-Supervised)

In Lab 6 you trained two models on the same 10 labeled samples: a supervised model using only those 10 labels, and a semi-supervised label-propagation model using the 10 labels PLUS many unlabeled samples. The unlabeled data helped - the semi-supervised model was more accurate without buying more labels. You tuned n_neighbors at 3, 5, 10, 20, and 40, and saw that a value too low or too high both hurt accuracy.

Q1: In Lab 6, why did adding unlabeled data improve accuracy even though no new labels were purchased?

Q2: n_neighbors of 3 and of 40 both gave worse accuracy than a middle value. In plain language, why do both extremes hurt?

Q3: NorthBank could only afford a few manual labels. Explain how a semi-supervised approach (a third path) could cut how many expensive manual labels the bank needs, and where it would still need humans.

Grading Rubric (20 points)

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

Total: 20 points
Scoring: Completion (8) + Reasoning (7) + Communication (5) = 20 points.
You are graded on the quality of your reasoning and writing, not on choosing a single 'right' answer.

How to Submit

1. Complete the worksheet as a team through all 3 rounds.
2. Make sure each decision has its risk / cost / evidence justification.
3. Include the Round 3 audit artifact and the stakeholder message.
4. Answer the debrief and Lab 6 questions.
5. Submit one PDF worksheet per team on Canvas by [DUE DATE].

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