

# AI/ML – Dangers of Feedback Loops

Group members: \_\_\_\_\_ Date: \_\_\_\_\_

Chosen domain (circle one): Music • News • Shopping • Short-video • Jobs

1) What is a feedback loop? (2–3 sentences)

In your own words:

2) Mini-Simulation: “RecoNarrow v1”

| Day | Belief A (%) | Belief B (%) | #A shown (of 10) | #B shown (of 10) | Clicks A (20%) | Clicks B (15%) |
|-----|--------------|--------------|------------------|------------------|----------------|----------------|
| 0   | 60           | 40           | 6                | 4                | 1              | 1              |
| 1   |              |              |                  |                  |                |                |
| 2   |              |              |                  |                  |                |                |
| 3   |              |              |                  |                  |                |                |

What happened? Why?

3) Where can harm occur? (tick all that apply and explain briefly)

☐ Loss of diversity (users see less variety) → \_\_\_\_\_

☐ Unfair exposure (Category B creators never get surfaced) → \_\_\_\_\_

☐ Stereotype reinforcement (system doubles down on past patterns) → \_\_\_\_\_

☐ User disengagement (boredom/echo chamber) → \_\_\_\_\_

#### 4) Stress test: cold start or trend shift

Imagine a new Category C appears on Day 2 (22% click rate if shown, but the model doesn't know that).

If your system only shows A/B based on past clicks, how likely is C to be discovered?

What simple rule could ensure exploration so C gets a fair chance?

#### 5) Break the loop — Design fixes (choose at least three)

- Exploration quota (10–20% of slots reserved for under-shown items) → How/Trade-off: \_\_\_\_\_

- Diversity metric (no more than 4/10 from one category) → How/Trade-off: \_\_\_\_\_

- Debiasing the signal (normalise clicks by impressions) → How/Trade-off: \_\_\_\_\_

- Fair exposure constraints (per-creator/category floors) → How/Trade-off: \_\_\_\_\_

- User controls (toggle 'wider variety', reset) → How/Trade-off: \_\_\_\_\_

- Fresh-start windows (periodic re-seeding with broader data) → How/Trade-off: \_\_\_\_\_

#### 6) Your group's recommendation (pitch)

Write a 5–7 sentence product note to your Head of Recommenders explaining the risks and your fixes.

#### 7) Success metrics (pick at least two and define a target)

- Category diversity (entropy  $\uparrow$  by \_\_\_\_% within 14 days)

- Creator exposure fairness (Gini  $\downarrow$  to  $\leq$  \_\_\_\_)

- User satisfaction (survey score  $\uparrow$  to  $\geq$  \_\_\_\_/5)

- Retention (7-day return rate  $\uparrow$  to \_\_\_\_%)

- Novel discovery rate (share of clicks on 'new to user'  $\uparrow$  to \_\_\_\_%)

#### 8) Reflection (individual, 2–3 sentences)

What did you learn about feedback loops that surprised you?