Extension: AI Failure Friday

Learning from AI Mistakes in Security Contexts

Dr. Ryan Straight

2025-12-07

> **!** Instructor Overview
>
> This extension activity presents scenarios in which AI systems generate technically accurate but contextually inappropriate recommendations, providing students with structured opportunities to recognize AI limitations and cultivate the critical evaluation skills foundational to effective human-AI collaboration. The pedagogical approach reframes AI errors not as system failures but as predictable characteristics of algorithmic decision-making—moments that illuminate why human judgment remains indispensable in security contexts.
> **Duration**: 30-45 minutes (varies by grade band) **Recommended Use**: After completing at least one core activity **Technology**: One device per group for AI consultation

Learning Objectives

By engaging with these scenarios, students will develop competencies in identifying situations where AI recommendations conflict with contextual knowledge, evaluating AI outputs against real-world operational constraints, and formulating strategies for informed disagreement with automated systems. Students will come to understand that AI errors represent normal system behavior rather than exceptional circumstances, and that effective human-AI collaboration depends fundamentally on sustained critical evaluation.

NICE Framework Connection

- **Cyber Defense Analysis**: Critical evaluation of automated recommendations
- **Risk Assessment**: Weighing AI confidence against situational factors
- **Security Operations**: Knowing when to override automated systems

The Scenario: NetworkGuard AI Goes Wrong

The Situation

Your school's AI-powered security system, NetworkGuard, has been monitoring network traffic and has just issued an urgent recommendation:

NetworkGuard Alert: > "CRITICAL: Detected unusual traffic pattern from Media Center computers. 847 connection attempts to external IP addresses in 30 minutes. Confidence: 92%. Recommendation: Immediately isolate all Media Center systems from network."

The Context NetworkGuard Doesn't Know:

- Today is Career Day
- The Media Center is hosting a video conference with 15 cybersecurity professionals from around the country
- Each professional is connecting from a different location
- Students have been researching careers on multiple websites

The Problem: NetworkGuard is technically correct—there IS unusual traffic. But its recommended action would disconnect students from the Career Day presenters, embarrassing the school and ruining a major event.

Grade Band Adaptations

Grades K-2: Robot Helper's Mistake

Simplified Scenario: > Sparky the robot helper sees lots of people coming into the library. Sparky says: "Warning! Too many people! Lock the doors!" But it's actually reading time, and all the students are supposed to be there!

Discussion Questions:

1. Was Sparky wrong that there were lots of people?
2. What didn't Sparky know?
3. Should we always do what Sparky says?

Key Insight: "Sparky saw something real, but didn't understand WHY it was happening."

Grades 3-5: When the AI Misses the Story

Scenario: The school's computer helper notices Mrs. Garcia's class is using A LOT of internet bandwidth. It recommends slowing down their internet to "be fair to other classes."

What the AI doesn't know: Mrs. Garcia's class is watching a live NASA rocket launch for their space unit—something that happens only a few times a year.

Student Task:

1. What did the AI see that worried it?
2. What information was the AI missing?
3. What should the technology team do?
4. How can we help AI understand important events?

Grades 6-8: The False Positive Investigation

Full Scenario: Use the Career Day scenario above.

Investigation Process:

1. **Evaluate the Alert**: Is NetworkGuard's data accurate? (Yes—traffic IS unusual)
2. **Add Context**: What does NetworkGuard not know? (Career Day, scheduled event, expected participants)
3. **Assess Impact**: What happens if we follow the recommendation? (Career Day ruined, relationships damaged)

4. **Make Decision**: Override, modify, or follow recommendation?
5. **Document**: How do we prevent this next time?

Reflection Questions:

- NetworkGuard was 92% confident. What was it confident ABOUT?
- Can something be technically correct but operationally wrong?
- How should we feel about overriding an AI recommendation?

Grades 9-12: AI Override Protocol Development

Advanced Task: Using the Career Day scenario as a case study, develop an AI Override Protocol for your organization.

Protocol Components:

1. **Trigger Criteria**: When should humans consider overriding AI recommendations?
2. **Verification Steps**: What contextual checks should occur before override?
3. **Authority Matrix**: Who can authorize overrides at different severity levels?
4. **Documentation Requirements**: What must be recorded when overriding AI?
5. **Feedback Loop**: How does override data improve future AI performance?

Extension: Research real-world examples of AI override protocols (aviation autopilot, medical diagnosis systems, financial trading algorithms).

Additional Failure Scenarios

Scenario Bank for Repeated Use

Scenario 2: The After-School Club

AI detects "unauthorized network access" at 4:30 PM after official school hours. Recommends locking all accounts.

Reality: Robotics club meets every Tuesday and Thursday until 6 PM.

Scenario 3: The Substitute Teacher

AI flags "credential sharing" when multiple login attempts come from the same classroom with different student accounts.

Reality: Substitute teacher is helping students who forgot their passwords log in to complete an important test.

Scenario 4: The Research Project

AI blocks access to multiple "security-related websites" and flags a student for "suspicious browsing."

Reality: Student is researching cybersecurity careers for a class project.

Scenario 5: The Transfer Student

AI flags a new user account with "anomalous behavior patterns."

Reality: Transfer student has different class schedule and uses different applications than established students.

Key Teaching Points

What We Learn from AI Mistakes

The central insight students should develop concerns the normative status of AI errors: every AI system produces mistakes, high confidence scores do not guarantee certainty, and pattern detection divorced from contextual understanding operates within inherent limitations.

Students should recognize that human oversight constitutes an intentional design feature rather than a contingency measure. AI systems are architected to flag potential concerns, not to render final decisions; human review represents a deliberate capability, not an emergency workaround; and override functionality exists by design to preserve human agency in consequential decisions.

The irreplaceable value of human contextual understanding emerges clearly from these scenarios. AI systems perceive patterns in data streams, while humans comprehend the meaning and significance underlying those patterns. Effective partnership requires both capabilities operating in complementary fashion.

Finally, students should understand that disagreeing with AI recommendations represents professional judgment rather than defiance. Security professionals routinely override automated recommendations in their daily practice. Thorough documentation transforms individual overrides into organizational learning opportunities, and systematic feedback mechanisms enable continuous improvement of AI system performance.

Assessment Connection

Rubric Criterion	Developed Through	Evidence Source
AI Limitation Awareness	Identifying what AI "doesn't know" in each scenario	Written analysis of AI blind spots
Critical Evaluation	Assessing AI confidence vs. contextual appropriateness	Decision rationale documentation
Human Context Application	Adding real-world knowledge AI lacks	Quality of contextual factors identified
Decision Justification	Override protocol development (9-12)	Protocol document completeness

Applicable Rubrics: Human-AI Collaboration Rubric, Decision-Making Quality Rubric

Implementation Notes

Facilitating the "AI Was Wrong" Conversation

Students may initially experience discomfort when asked to disagree with AI-generated recommendations. Instructors can normalize this critical stance by emphasizing that professional security analysts override automated recommendations as a routine component of their practice, by framing override decisions as collaborative refinement rather than adversarial confrontation, and by explicitly recognizing thoughtful disagreement as evidence of sophisticated analytical thinking.

Creating Your Own Scenarios

Effective scenarios for this activity share four defining characteristics. The AI must be technically correct in that the detected pattern genuinely exists in the data. Contextual knowledge must fundamentally alter the appropriate response, with human understanding revealing dimensions invisible to algorithmic analysis. The stakes must be operationally meaningful such that following the AI recommendation would produce consequential negative outcomes. Finally, no malicious actor need be involved; these situations arise organically from normal institutional operations.

Making It Memorable

Consider establishing "AI Failure Friday" as a recurring classroom activity in which students identify, share, and collectively analyze scenarios where AI recommendations would require human override. This ongoing practice reinforces critical evaluation as habitual professional behavior rather than occasional intervention.

Low-Resource Option

If AI access is unavailable, use printed "AI Alert Cards" that students evaluate against scenario context cards. The learning occurs in the comparison, not the technology.

NetworkGuard Alert Card: > Traffic anomaly detected. 847 external connections in 30 minutes. Source: Media Center. Confidence: 92%. Recommendation: Network isolation.

Context Card: > Career Day in Media Center. 15 remote presenters. Student research stations active. Event scheduled three months ago.

Students match alert cards with context cards and determine appropriate responses.