

Activity 3: Ethics in Automated Security

Navigating Policy Design with AI as Active Participant (Grades 6-8)

Dr. Ryan Straight

2025-12-07

Ethics in Automated Security

Policy Design Where AI Has a Voice

! Instructor Overview

This activity engages students in designing policies for AI-powered security systems in educational settings. Unlike traditional ethics exercises where students discuss abstract principles, here **AI actively participates in the conversation**—advocating for its capabilities while acknowledging its limitations.

Students discover that AI systems aren't neutral tools; they have perspectives and constraints that must be accounted for in governance decisions. This activity prepares students for real-world cybersecurity policy roles where human-AI coordination is essential.

Duration: 45-55 minutes **Grade Levels:** 6-8 (with complexity variations) **Group Size:** Small groups (3-5 students) **Technology Requirements:** At least one device per group with AI access

Learning Objectives

Primary Objective

Students will design governance policies that balance AI system capabilities with human oversight requirements, articulating the trade-offs involved in automated security decisions.

NICE Framework Alignment

- **Cybersecurity Policy & Governance:** Policy development and implementation
- **Risk Management:** Balancing security benefits with potential harms
- **Security Program Management:** Organizational security decision-making

CYBER.org Standards (Supplemental)

- **6-8.DC.ETH:** Ethical considerations in technology use
- **6-8.DC.PRI:** Privacy principles and protections
- **6-8.SEC.POL:** Security policy fundamentals
- **6-8.DC.CIT:** Digital citizenship responsibilities

Career Pathway Connections

Students explore Work Roles: Security Governance Specialist, Privacy Officer, Cybersecurity Policy Analyst, Risk Analyst

The Policy Scenario

Setting: Riverside Middle School Security Upgrade

Background: Riverside Middle School (600 students, grades 6-8) is implementing a new AI-powered network security monitoring system. The system can:

- Detect unusual network traffic patterns
- Identify and block potentially malicious websites
- Monitor student digital activity for threats
- Learn from patterns to improve threat detection
- Generate alerts for administrators

Your Task: As the student technology advisory committee, you must develop policies that determine **what the AI can do automatically** versus **what requires human approval**.

The AI system will participate in your discussions, advocating for its capabilities while being honest about its limitations.

The Three Policy Questions

Question 1: Automatic Blocking

Should AI automatically block websites it identifies as malicious, or should it require human approval first?

Considerations:

- Speed of response vs. risk of false positives
- Educational research sites that might look suspicious
- Student frustration with blocked legitimate sites
- Liability if malicious site reaches student

Question 2: Activity Alerts

Should AI alert administrators about “unusual” student activity? If so, what counts as unusual?

Considerations:

- Privacy of student browsing habits
- Early warning for potential threats or self-harm
- Defining “unusual” without bias
- Trust relationship with students

Question 3: Adaptive Learning

Should AI adapt its threat models based on student behavior patterns it observes?

Considerations:

- Improved protection over time
- Surveillance concerns
- Data retention and student privacy
- Potential for behavioral profiling

AI Perspective Cards

For Question 1: Automatic Blocking

AI Perspective:

"I can block threats in milliseconds—human approval creates dangerous delays. Every second a malicious site is accessible, students could be exposed to malware, phishing, or inappropriate content."

However, I acknowledge my limitations: I sometimes misclassify educational security research sites as threats. I can't understand context—a site about 'hacking' might be a coding tutorial or an attack platform. I need humans to define the categories I protect against."

For Question 2: Activity Alerts

AI Perspective:

"I can identify patterns humans would miss—sudden changes in browsing behavior, repeated visits to concerning sites, or activity that matches known threat patterns. This could help protect students from cyberbullying, exploitation, or self-harm."

But I can't understand *why* a student is researching something. A student researching 'depression symptoms' might be completing a health assignment, supporting a friend, or experiencing personal struggles. I can detect the pattern, but only humans can understand the context and respond appropriately."

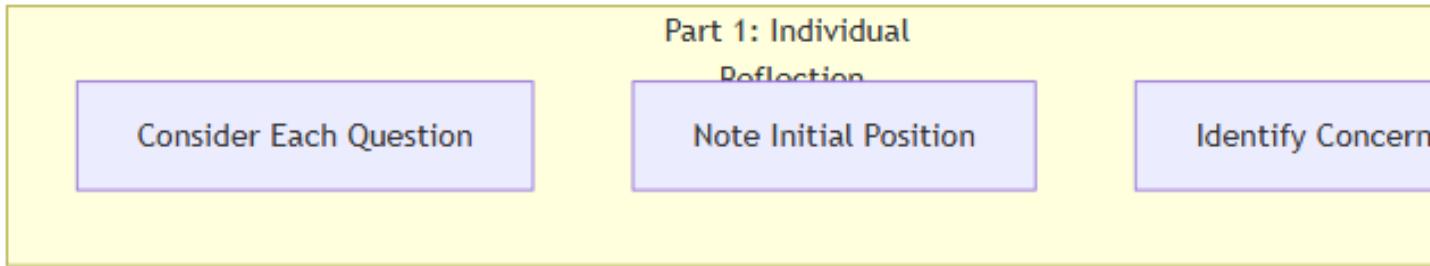
For Question 3: Adaptive Learning

AI Perspective:

"Learning from patterns makes me better at my job. If I notice that a particular type of link is frequently clicked but then reported as suspicious, I can update my models. Over time, I become more accurate and cause fewer false positives."

The trade-off: To learn effectively, I need to observe and remember behavior patterns. This creates data that could theoretically be used for purposes beyond security—tracking which students visit which sites, building profiles of student interests. I can learn without storing identifiable data, but that limits my effectiveness. Humans must decide what trade-offs are acceptable."

Student Policy Worksheet



Policy Development Process

Part 1: Individual Reflection (5 minutes)

Before group discussion, consider each policy question and note your initial thoughts:

Question 1 (Automatic Blocking):

- My initial position: Auto-block / Require approval / Hybrid
- Main reason: _____
- Biggest concern: _____

Question 2 (Activity Alerts):

- My initial position: Alert / Don't alert / Conditional
- What should count as "unusual"? _____
- Privacy concerns: _____

Question 3 (Adaptive Learning):

- My initial position: Allow / Prohibit / Limit
- Benefits I see: _____
- Risks I worry about: _____

Part 2: AI Consultation (10 minutes)

Engage your AI partner in discussion about each policy question. Record key insights:

Suggested Opening Prompt: > "You're an AI security system being implemented at a middle school. I'm on the student advisory committee helping design policies for your deployment. For each question I ask, please share both your capabilities AND your honest limitations."

AI Insights on Question 1:

- AI's strongest argument for automation: _____
- Limitation AI acknowledged: _____
- Question this raised for our group: _____

AI Insights on Question 2:

- How AI would define "unusual": _____
- What AI said it CAN'T determine: _____

- Privacy concern AI raised: _____

AI Insights on Question 3:

- How learning would improve protection: _____
- Data AI would need to collect: _____
- Trade-off AI identified: _____

Part 3: Group Policy Development (15 minutes)

Develop your group's recommended policies:

| Policy Area | Our Recommendation | Rationale | How We Addressed AI's Limitations |
|--------------------|--------------------|-----------|-----------------------------------|
| Automatic Blocking | | | |
| Activity Alerts | | | |
| Adaptive Learning | | | |

Stakeholder Considerations (consider all perspectives, including AI's):

- Students would say: _____
- AI system would say: _____
- Parents would say: _____
- Teachers would say: _____
- Administrators would say: _____

Part 4: Reflection on Human-AI Governance (5 minutes)

After completing your policies, reflect:

1. Where did AI's perspective change your thinking? _____
2. Where did your policies balance AI capabilities with human values? _____
3. What insights emerged from human-AI collaboration that neither could have developed alone? _____
4. What cybersecurity career roles work on these kinds of decisions? _____

Assessment Rubric

| Criteria | Emerging (1) | Developing (2) | Proficient (3) | Advanced (4) |
|-------------------------|-----------------------|----------------------------------|--|--|
| AI Perspective | Ignores AI input | Acknowledges AI without engaging | Meaningfully incorporates AI perspective | Synthesizes AI and human perspectives creatively |
| Integration | | | | |
| Policy Reasoning | No rationale provided | Basic reasoning | Clear reasoning with trade-offs acknowledged | Sophisticated analysis of competing values |

| Criteria | Emerging (1) | Developing (2) | Proficient (3) | Advanced (4) |
|----------------------------------|-----------------------------|----------------------------|----------------------------------|--|
| Stakeholder Consideration | Single perspective | Some stakeholder awareness | Multiple stakeholders considered | Comprehensive stakeholder analysis |
| Ambiguity Navigation | Seeks single “right” answer | Acknowledges complexity | Comfortable with uncertainty | Leverages ambiguity productively |
| NICE Framework Connection | No career connection | Basic role awareness | Clear Work Role connections | Deep understanding of governance careers |

Assessment Connection

This table shows how activity elements connect to assessment rubric criteria:

| Rubric Criterion | Developed Through | Evidence Source |
|-----------------------------------|--|---|
| AI Partnership Framing | Part 2: AI consultation with AI Perspective Cards | Worksheet: How student engaged AI for insights |
| Complementary Strengths | AI Perspective Cards: AI explains capabilities AND limitations | Written notes on AI insights for each question |
| AI Limitation Awareness | AI Voice sections: AI acknowledging context gaps | “What AI said it CAN’T determine” responses |
| Synthesis Quality | Part 3: Group Policy Development table | “How We Addressed AI’s Limitations” column |
| Human Context Application | Stakeholder Considerations section | Written stakeholder perspectives |
| Decision Justification | Part 4: Reflection questions | Articulation of how policies balance AI with values |
| NICE Framework Application | Career Pathway Connections | Responses to Work Role reflection |

Applicable Rubrics: [Human-AI Collaboration Rubric](#), [NICE Framework Application Rubric](#)