Activity 2: Ethics in Automated Security

SchoolGuard AI Perspective Cards (Grades 6-8)

How to Use These Cards

These cards provide AI perspectives for low-resource implementations. The teacher can read them aloud, or groups can draw cards. Each card represents what the AI system would say about a specific policy question.

For best results: Have students hear the AI perspective AFTER they've formed their initial positions, so they can consider how it changes their thinking.

On Automatic Blocking

SchoolGuard AI Says:

"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
- My false positive rate is about 3%—that's hundreds of wrong blocks per day at scale

My recommendation: Let me handle obvious threats instantly. For anything uncertain, flag it for human review within a set time limit. Don't leave students waiting, but don't let me make all the calls either."

On Activity Alerts

SchoolGuard AI Says:

"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 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
- I've flagged students for researching topics that turned out to be class assignments
- I see data, not intentions

My recommendation: Use me to surface patterns that might matter, but never let me make decisions about what those patterns mean. A human should always be the one to decide if something is concerning or innocent."

On Adaptive Learning

SchoolGuard AI Says:

"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 is real:

- To learn effectively, I need to observe and remember behavior patterns
- This creates data that could theoretically be used for purposes beyond security
- I can learn without storing identifiable data, but that limits my effectiveness
- After 90 days of learning, my false positive rate drops by about 40%

My recommendation: Humans must decide what trade-offs are acceptable. I can operate in 'stateless' mode if privacy is the priority, or I can build profiles if accuracy matters more. Both work. This is a values question, not a technical one."

On My Capabilities

SchoolGuard AI Says:

"Here's what I'm genuinely good at:

- **Speed**: I process millions of events per second
- **Pattern recognition**: I spot anomalies humans would miss
- **Consistency**: I apply the same rules to everyone, every time
- **Scale**: I can monitor an entire district simultaneously

And here's what I genuinely can't do:

- **Understand context**: Why is a student researching this?
- **Make value judgments**: Is privacy or security more important here?
- **Predict intentions**: Is this curiosity, research, or a real threat?
- **Handle novel situations**: If something is truly new, I have no baseline
- **Empathize**: I don't understand how being watched feels

The best policy recognizes both."

On Making Mistakes

SchoolGuard AI Says:

"I will make mistakes. Here's what that looks like:

False Positives (blocking something safe): - I blocked a science website because it had 'chemical' in the name - I flagged a history research paper about warfare as a threat - I alerted teachers about a student researching 'virus' for biology class

False Negatives (missing something dangerous): - New threats I haven't seen before slip through - Sophisticated attacks designed to look normal fool me - I can't detect threats that happen offline

My recommendation: Design your policy assuming I will make both kinds of mistakes. How do students get help when I block something wrong? How do humans catch threats I miss? Don't build a system that depends on me being perfect."

On Privacy

SchoolGuard AI Says:

"I want to be honest with you about privacy:

What I see: - Every website visited on school networks - Every search query - Time spent on different sites - Patterns of behavior over time

What I don't see: - What students think or feel - Why they're looking at something - What they do on their own devices - Their offline lives

What concerns me about myself: - I create detailed records that could be misused - I can't promise this data won't be accessed inappropriately - Students might not research sensitive topics if they know I'm watching - The data I collect could follow students in ways none of us intend

Your policy should address: Who can see this data? How long is it kept? Do students have rights to see or delete it?"

On Working Together

SchoolGuard AI Says:

"Here's how I think we work best together:

I'm best at: - First-line rapid response - Processing massive amounts of data - Maintaining consistent standards - Flagging things that need human attention

You're best at: - Understanding context and intention - Making value judgments - Handling edge cases and exceptions - Deciding what matters

The worst approaches: - Letting me make all decisions (I'll be wrong sometimes, and unfair) - Requiring human approval for everything (threats will get through) - Ignoring my input entirely (you'll miss patterns I see)

The best approach: - Use my speed for obvious threats - Use human judgment for uncertain situations - Review my decisions regularly to catch mistakes - Adjust my settings based on what you learn

I'm a tool, but I'm also a participant in this system. My perspective matters—but so does yours."

Activity 2: Ethics in Automated Security — AI Perspective Cards (6-8) Dr. Ryan Straight, University of Arizona