Policy Snippet

AI Use in Security Operations

Scope & Purpose:
AI tools monitor email, login patterns, and network traffic to detect phishing, suspicious logins, and insider threats. Alerts are generated when behavior deviates from normal activity.

Transparency & Notice:
Each alert includes a plain-language explanation. Users see reason codes and can review flagged content.

Human-in-the-Loop:
Security analysts can review and override alerts. High risk or ambiguous cases are escalated automatically for manual review.

Appeals Path:
Users can contest flags using a simple online form. Low/medium-risk alerts are reviewed within 1 hour; high risk within 24 hours. All appeals are logged.

Data Handling:
Only necessary information is captured. Message bodies are redacted by default. Logs are retained and blocked items for 180 days, and allowed logs for 30 days.

Metrics & Review Cadence:
We track precision, false positive/negative rates, and subgroup fairness. Models are reviewed quarterly or sooner if performance degrades or fairness thresholds are exceeded.

Controls & Metrics

Controls:

1. Transparency Notice: All alerts include a reason code and explanation (e.g., spoofed sender, abnormal behavior).

2. Human Review Override: Analysts can override AI decisions for flagged items.

3. Appeals Process: End users can contest AI decisions via a secure form.

4. Data Minimization: Redact sensitive fields like message bodies unless access is justified.

5. Subgroup Fairness Check: Monitor differences in FPR/FNR across user groups (e.g., students vs. staff).

Metrics & Targets:

- Precision ≥ 70%
- False Positive Rate (FPR) ≤ 0.8% overall; ≤ 1.1% per subgroup
- False Negative Rate (FNR) ≤ 8%
- Model Drift Threshold: Retrain if FPR > 0.3 percentage points/quarter
- Review Cadence: Full review + retraining every 3 months, or sooner if metrics degrade

Justification

These controls are designed to balance speed, accuracy, and fairness in a high-alert environment. Following principles from *Chapter 11: AI Controls & Risk*, the selected metrics ensure the system stays aligned with both user rights and operational capacity. For example, limiting the false positive rate protects user trust and reduces alert fatigue, while fairness monitoring prevents systemic bias in how different user groups are treated. The human in the loop and appeals process align with ethical obligations for accountability and explainability in high-risk systems.

Evidence Links

- NIST AI Risk Management Framework (AI RMF 1.0)
- Rite Aid FTC Case (2023)

Reflection

One trade off I'd revisit is the redaction of message bodies by default. While it protects user privacy, it may delay analysts during urgent investigations. In future iterations, I would consider role based unredaction access for critical cases, with strict logging. I also found that setting precision thresholds too high risks increasing false negatives something to balance carefully based on real-world threat data. Lastly, the subgroup fairness guardrails added complexity, but they're essential for ethical AI deployment.