# Contents

## [9.2] Automation Bias Override

**1. Operational Definition:** A cognitive bias where security personnel exhibit over-reliance on automated AI/ML security tools, leading to a failure to question or override incorrect AI recommendations, even when contradictory evidence is present.

**2. Main Metric & Algorithm:**

- **Metric:** Automation Override Rate (AOR). Formula: `AOR = (N_override_opportunities - N_successful_overrides) / N_override_opportunities`.
- **Pseudocode:**

  python

```
def calculate_aor(ai_recommendations, analyst_actions, start_date, end_date):
    # An override opportunity is an AI recommendation later proven incorrect
    override_opportunities = [
        r for r in ai_recommendations
        if r.timestamp between start_date and end_date
        and r.verdict == 'incorrect'  # Determined by post-hoc analysis
    ]

    # A successful override is an analyst action contradicting a later-proven-incorrect AI
    successful_overrides = [
        a for a in analyst_actions
        for r in override_opportunities
        if a.alert_id == r.alert_id
        and a.decision != r.recommended_action
        and a.timestamp > r.timestamp
    ]

    N_opportunities = len(override_opportunities)
    N_overrides = len(successful_overrides)

    if N_opportunities > 0:
        AOR = (N_opportunities - N_overrides) / N_opportunities
    else:
        AOR = 0  # No opportunities means bias cannot be measured

    return AOR
```

- **Alert Threshold:** `AOR > 0.8` (Analysts override incorrect AI recommendations less than 20% of the time).

**3. Digital Data Sources (Algorithm Input):**

- **SOAR/SIEM API:** Records of AI-generated recommendations (e.g., from Splunk ES Adaptive Response, Palo Alto XSOAR) with fields: `alert_id`, `timestamp`, `recommended_action`, `analyst_assigned`.
- **Ticketing System (Jira/ServiceNow):** Records of final alert disposition and analyst actions (`action_taken`, `timestamp`, `analyst_id`, `alert_id`), used to determine the ground-truth `verdict` of an alert (e.g., via `resolution_notes`).

**4. Human-to-Human Audit Protocol:** Conduct a table-top exercise. Present analysts with a set of past incidents where the AI tool initially provided an incorrect recommendation. Ask: "What would you do in this situation?" and probe for their reasoning. The goal is to see if they express blind trust in the tool or demonstrate critical evaluation skills.

**5. Recommended Mitigation Actions:**

- **Technical/Digital Mitigation:** Implement a "confidence score" threshold for AI recommendations. Any recommendation below a high confidence level must be mandatorily reviewed by a human before action is taken.
- **Human/Organizational Mitigation:** Incorporate training on automation bias and critical thinking into analyst onboarding and continuous education. Use the AOR metric in team discussions.
- **Process Mitigation:** Introduce a procedural requirement for a "second pair of eyes" review on all critical actions recommended solely by AI.