Contents

[9.4] AI Authority Transfer

1. Operational Definition: The unconscious ceding of human decision-making authority and responsibility to the AI system, diminishing the analyst's sense of agency and critical engagement with the security process.

2. Main Metric & Algorithm:

- **Metric:** Agency Loss Score (ALS). Formula: `ALS = (N_AI_recommendations_accepted_verbatim) / N_total_AI_recommendations_accepted`.
- **Pseudocode:**

  python

```
def calculate_als(ai_recommendations, analyst_actions, start_date, end_date):
    # Get all AI recommendations that were accepted by an analyst
    accepted_recommendations = [
        r for r in ai_recommendations
        if exists_analyst_action(r.alert_id, r.recommended_action)
        and r.timestamp between start_date and end_date
    ]

    # Filter for those accepted without any modification or additional note
    verbatim_acceptances = [
        r for r in accepted_recommendations
        if not exists_analyst_notes(r.alert_id)  # No additional reasoning provided
        and not was_modified(r.alert_id, r.recommended_action) # Action taken matches AI r
    ]

    N_total_acceptances = len(accepted_recommendations)
    N_verbatim = len(verbatim_acceptances)

    if N_total_acceptances > 0:
        ALS = N_verbatim / N_total_acceptances
    else:
        ALS = 0

    return ALS
```

- **Alert Threshold:** `ALS > 0.9` (Over 90% of accepted AI recommendations are followed exactly without human modification or documented reasoning).

3. Digital Data Sources (Algorithm Input):

- **AI System & Ticketing System APIs:** As in 9.2 and 9.3, requiring logs of AI recommendations (`recommended_action`), analyst actions (`action_taken`), and analyst notes (`resolution_notes`, `comments`).

4. Human-to-Human Audit Protocol: Observe an analyst using the AI tool. Note if they scroll past the AI recommendation to look at raw data themselves. In interviews, ask: "Who is ultimately responsible if an action based on an AI recommendation leads to a breach?" Listen for answers that deflect responsibility to the tool.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Design the UI to require a mandatory "Analyst Rationale" text field to be filled before proceeding with an AI-recommended action on high-severity alerts.
- **Human/Organizational Mitigation:** Leadership must explicitly communicate that the analyst, not the tool, is responsible for decisions. Reinforce this in training.
- **Process Mitigation:** Implement a policy where a certain percentage of AI-driven decisions are randomly selected for mandatory peer review.