Contents

[9.3] Algorithm Aversion Paradox

1. Operational Definition: The counter-intuitive tendency to distrust or disregard accurate and reliable AI/automated systems, often after experiencing a small number of previous errors, leading to under-utilization of a valuable tool.

2. Main Metric & Algorithm:

- **Metric:** System Disuse Index (SDI). Formula: `SDI = (N_actions_taken_without_consulting_AI) / N_total_actions` for tasks where AI support is available and recommended.
- **Pseudocode:**

  python

  ```
  def calculate_sdi(analyst_actions, ai_capable_alerts, start_date, end_date):
      # Get all actions taken by analysts on alerts that AI could have assisted with
      relevant_actions = [
          a for a in analyst_actions
          if a.alert_id in ai_capable_alerts
          and a.timestamp between start_date and end_date
      ]

      # Check if the analyst consulted the AI before taking the action
      # This requires logs from the AI system showing query/access events
      actions_without_ai = [
          a for a in relevant_actions
          if not exists_ai_access_log(a.alert_id, a.analyst_id, a.timestamp)
      ]

      N_total = len(relevant_actions)
      N_disused = len(actions_without_ai)

      if N_total > 0:
          SDI = N_disused / N_total
      else:
          SDI = 0

      return SDI
  ```

- **Alert Threshold:** `SDI > 0.75` (Analysts are not using available AI support for over 75% of relevant actions).

3. Digital Data Sources (Algorithm Input):

- **SOAR/SIEM API:** Logs of all analyst actions on alerts (`analyst_id`, `alert_id`, `action`, `timestamp`).

- **AI System Audit Logs:** Records of each time an analyst queries the AI for a recommendation on a specific alert (`analyst_id`, `alert_id`, `query_timestamp`).
- **CMDB/Asset DB:** A list of `alert_types` or `asset_classes` for which AI support is available and recommended (`ai_capable_alerts`).

4. Human-to-Human Audit Protocol: Interview team members: "Can you describe a time you disagreed with the AI tool's suggestion? What did you do? How often do you check its recommendations?" Look for narratives of distrust based on isolated past events rather than current performance metrics.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Improve AI system transparency by providing concise, understandable explanations for its recommendations (XAI - Explainable AI).
- **Human/Organizational Mitigation:** Showcase the AI's overall accuracy and value in team meetings, addressing the "why" behind its recommendations to rebuild trust.
- **Process Mitigation:** Initially make AI consultation a mandatory step in the playbook for specific alert types, forcing re-engagement to demonstrate improved reliability.