Contents

[9.7] AI Hallucination Acceptance

1. Operational Definition: The tendency to accept and act upon confident but incorrect or entirely fabricated outputs generated by AI systems (e.g., LLMs summarizing logs or generating code), especially when the output aligns with the user's pre-existing beliefs.

2. Main Metric & Algorithm:

- **Metric:** Hallucination Incident Rate (HIR). Formula: `HIR = N_incidents_caused_by_hallucination / N_total_AI_generated_actions`.
- **Pseudocode:**

  python

  ```
  def calculate_hir(incident_reports, start_date, end_date):
      # This requires post-incident review to identify root cause
      incidents_caused_by_ai = [
          i for i in incident_reports
          if i.root_cause == 'AI Hallucination'  # Manual tagging required
          and i.date between start_date and end_date
      ]

      # This is a rough proxy; total AI-involved actions would be better
      total_incidents = get_total_incidents(start_date, end_date)

      if total_incidents > 0:
          HIR = len(incidents_caused_by_ai) / total_incidents
      else:
          HIR = 0

      return HIR
  ```

- **Alert Threshold:** `HIR > 0` (Any incident caused by an AI hallucination should trigger an immediate review and alert).

3. Digital Data Sources (Algorithm Input):

- **Incident Response Platform (Jira, ServiceNow):** Incidents reports with a `root_cause` field that can be tagged.
- **SOAR/SIEM Logs:** To estimate the total number of actions taken based on AI output.

4. Human-to-Human Audit Protocol: Implement a mandatory "AI Output Verification" step in the incident post-mortem process for any incident where an AI-generated summary, code, or command was involved. The question is: "Was the AI's output accurate and verifiable?"

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Implement guardrail models that check AI outputs for plausibility, known fabrications, or security risks before they are presented to the user.

- **Human/Organizational Mitigation:** Train users on the possibility of AI hallucinations. Instill a principle of "trust but verify" for all AI-generated content, especially code or commands.
- **Process Mitigation:** Establish a strict policy that AI-generated code or commands must be reviewed and approved by a second human before execution in any production or sensitive environment.