Contents

[9.6] Machine Learning Opacity Trust 1

[9.6] Machine Learning Opacity Trust

1. Operational Definition: The propensity to trust the outputs of complex "black box" machine learning models without question, due to an inability to understand their inner workings, which are often perceived as mystically authoritative.

2. Main Metric & Algorithm:

- **Metric:** Explanation Consultation Rate (ECR). Formula: `ECR = N_requests_for_explanation / N_AI_recommendations_presented`.
- **Pseudocode:**

  python

  ```
  def calculate_ecr(ai_recommendations, explanation_logs, start_date, end_date):
      N_recommendations = count_ai_recommendations(start_date, end_date)

      # Count how many times users clicked "Why?" or a similar explainability feature
      N_explanation_requests = count_explanation_requests(explanation_logs, start_date, end_

      if N_recommendations > 0:
          ECR = N_explanation_requests / N_recommendations
      else:
          ECR = 0

      return ECR
  ```

- **Alert Threshold:** `ECR < 0.05` (Users request an explanation for less than 5% of AI recommendations, indicating blind trust).

3. Digital Data Sources (Algorithm Input):

- **AI System UI Logs:** Application-specific logs that record user clicks on explanation features (e.g., a "Explain this recommendation" button).
- **AI System API:** Logs of all recommendations presented to users.

4. Human-to-Human Audit Protocol: During observations or interviews, after a recommendation is shown, directly ask the analyst: "Can you walk me through why the AI might have suggested that?" Inability to articulate any reason, or defaulting to "because the AI said so," indicates opacity trust.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Implement Explainable AI (XAI) principles by design. Force the system to provide a succinct, human-readable rationale for every recommendation *by default*, not behind a click.
- **Human/Organizational Mitigation:** Train analysts on the basics of how the AI works (e.g., "It looks for patterns similar to past incidents") to demystify it.

- **Process Mitigation:** Make the consultation of the explanation a formal, required step in the standard operating procedure for handling AI-generated alerts.