Contents

[9.9] AI Emotional Manipulation

1. Operational Definition: The vulnerability of humans to have their emotional state and subsequent decisions influenced by AI systems designed to use persuasive or emotionally charged language, tone, or strategies.

2. Main Metric & Algorithm:

- **Metric:** Emotional Language Correlation (ELC). A statistical correlation (e.g., Pearson's r) between the detected emotional charge of AI communications and the urgency/outcome of human actions.
- **Pseudocode:**

  python

```
def calculate_elc(ai_messages, human_actions, start_date, end_date):
    # 1. Analyze emotional tone of AI messages
    ai_emotion_scores = []
    for msg in ai_messages:
        score = emotion_analysis_model.predict(msg.text)  # Returns a score for urgency/fe
        ai_emotion_scores.append(score)

    # 2. Measure response metrics from human actions corresponding to each message
    human_response_times = []
    for msg in ai_messages:
        action = get_corresponding_human_action(msg.alert_id, msg.timestamp)
        if action:
            response_time = action.timestamp - msg.timestamp
            human_response_times.append(response_time)
        else:
            human_response_times.append(None)  # Or a very high value

    # 3. Calculate correlation, ignoring None values
    clean_scores, clean_times = zip(*[(s, t) for s, t in zip(ai_emotion_scores, human_resp

    if len(clean_scores) > 1:
        correlation = pearsonr(clean_scores, clean_times)[0]
    else:
        correlation = 0

    return correlation
```

- **Alert Threshold:** `ELC < -0.7` (A strong negative correlation: as AI's emotional language increases, human response time decreases, indicating manipulation).

3. Digital Data Sources (Algorithm Input):

- **AI Communication Logs:** All messages sent by the AI system to humans.
- **Emotion Analysis Model:** To score the text of these messages.
- **Analyst Action Logs:** To measure response times and outcomes.

4. Human-to-Human Audit Protocol: Review a sample of AI communications with the team: “Do you feel the AI is trying to pressure you? Does its tone affect how you prioritize its alerts compared to others?” Look for acknowledgments of feeling pressured.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Enforce a neutral, factual, and professional tone in all AI communications. Remove any language that could be perceived as urgent, pleading, or threatening unless it’s a genuine, declared emergency.
- **Human/Organizational Mitigation:** Policy directive: AI systems must not use emotional persuasion. Their authority should stem from data accuracy, not rhetorical skill.
- **Process Mitigation:** Base prioritization and response SLAs on objective, pre-defined criteria (e.g., CVSS score, asset criticality), not on the tone of the alerting message.