Contents

[9.5] Uncanny Valley Effects

1. Operational Definition: The sense of unease, distrust, or revulsion triggered by AI systems that mimic human behavior (e.g., conversational chatbots) in a way that is almost, but not perfectly, realistic, impairing effective collaboration.

2. Main Metric & Algorithm:

- **Metric:** Negative Sentiment Ratio (NSR) in AI-Human Interactions. Formula: `NSR = N_negative_interactions / N_total_interactions`.
- **Pseudocode:**

  python

  ```
  def calculate_nsr(chat_logs, ai_agent_id, start_date, end_date):
      # Get all interactions with the AI agent
      interactions = get_chat_interactions(ai_agent_id, start_date, end_date)

      # Use sentiment analysis model on human messages within these interactions
      negative_interactions = set()

      for interaction in interactions:
          for message in interaction.human_messages:
              sentiment = sentiment_analysis_model.predict(message.text)
              if sentiment == 'negative':
                  negative_interactions.add(interaction.id)
                  break  # One negative message marks the whole interaction

      N_total = len(interactions)
      N_negative = len(negative_interactions)

      if N_total > 0:
          NSR = N_negative / N_total
      else:
          NSR = 0

      return NSR
  ```

- **Alert Threshold:** `NSR > 0.3` (Over 30% of conversational interactions with the AI agent contain detectable negative sentiment).

3. Digital Data Sources (Algorithm Input):

- **Chat Platform API (Slack/Teams):** Logs of conversations with the AI chatbot/agent (`message`, `user`, `timestamp`, `thread_id`).
- **A pre-trained Sentiment Analysis Model:** To process the text of human messages in these conversations (e.g., using a library like VADER or a dedicated API).

4. Human-to-Human Audit Protocol: Conduct focus groups or interviews: "How do you feel about interacting with the security chatbot? Does anything about it feel 'off' or frustrating?" Analyze feedback for themes related to unnatural responses, frustration, or creepiness.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Re-design the AI's conversational style. Often, moving away from attempting to mimic humans and towards a clearly artificial but helpful and transparent persona (e.g., "I am an AI designed to…") can reduce uncanny valley effects.
- **Human/Organizational Mitigation:** Be transparent about the AI's capabilities and limitations. Explain it is a tool, not a human.
- **Process Mitigation:** Provide an easy and obvious escape hatch to a human operator if the AI interaction becomes frustrating or ineffective.