Contents

[9.10] Algorithmic Fairness Blindness

1. Operational Definition: The failure to recognize, audit, or mitigate biases within AI security systems that lead to disproportionately high false positive rates or targeted scrutiny against specific individuals, groups, or system behaviors based on non-relevant characteristics.

2. Main Metric & Algorithm:

- **Metric:** Group Disparity Index (GDI). For a protected group (e.g., a specific team, location), calculate: `GDI = (FP_Group_A / N_Actions_Group_A) / (FP_Group_B / N_Actions_Group_B)`.
- **Pseudocode:**

  python

  ```
  def calculate_gdi(alert_data, protected_group, control_group, start_date, end_date):
      # Get alerts for Group A (protected group)
      alerts_group_a = get_alerts_for_group(protected_group, start_date, end_date)
      fp_group_a = count_false_positives(alerts_group_a)
      actions_group_a = count_actions(alerts_group_a)

      # Get alerts for Group B (control group)
      alerts_group_b = get_alerts_for_group(control_group, start_date, end_date)
      fp_group_b = count_false_positives(alerts_group_b)
      actions_group_b = count_actions(alerts_group_b)

      # Calculate False Positive Rate for each group
      if actions_group_a > 0:
          fpr_a = fp_group_a / actions_group_a
      else:
          fpr_a = 0

      if actions_group_b > 0:
          fpr_b = fp_group_b / actions_group_b
      else:
          fpr_b = 0

      # Avoid division by zero
      if fpr_b > 0:
          GDI = fpr_a / fpr_b
      else:
          GDI = float('inf')  # Undefined, but indicates a severe issue

      return GDI
  ```

- **Alert Threshold:** `GDI > 4.0 or GDI < 0.25` (The false positive rate for one group is 4x

higher or 4x lower than for another, indicating potential bias).

3. Digital Data Sources (Algorithm Input):

- **SIEM/SOAR:** Alert data enriched with metadata about the "actor" or "target" that can define groups (e.g., `user_department`, `geo_location`, `job_role`).
- **Ticketing System:** Data to determine the final, ground-truth classification of an alert as True/False Positive.

4. Human-to-Human Audit Protocol: Conduct an audit meeting: "Let's analyze our alert data. Are there any teams, locations, or types of behavior that seem to get flagged much more often than others? And when we investigate, are those alerts more likely to be false positives?" This is a qualitative review of the quantitative GDI metric.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Regularly run bias audits on the AI's output using metrics like GDI. Implement fairness-aware machine learning techniques during model (re)training.
- **Human/Organizational Mitigation:** Form a diverse oversight committee including members from HR, legal, and diverse business units to review bias audit results.
- **Process Mitigation:** Establish a formal process for individuals or teams to appeal and request a review of AI-generated alerts they believe are biased or unfair.