Phase II - Identification

❏ 1. Assess nature and materiality of incident through:

    ❏ Disparate impact analysis
    ❏ Input distribution drift
    ❏ Residual analysis
    ❏ Reject on negative impact (RONI) analysis
    ❏ Prediction distribution drift
    ❏ Sensitivity analysis

    ❏ Scanning new and past AI system traffic for:
        ❏ Training data
        ❏ Duplicate data
    ❏ Score insiders with affected AI system (in case of insider attack)

❏ 2. Categorize incident as AI failure or AI attack:

AI failures:

❏ API mismatches
❏ Data drift
❏ Data entanglement
❏ Discrimination
❏ Error propagation
❏ Feedback loops
❏ Inability to scale
❏ Instability
❏ Lack of accountability:
    ❏ Inability to explain predictions
    ❏ Inconsistent or inaccurate explanations
    ❏ No consumer-appeal capability
❏ Silencing of monitoring alerts
❏ System failures:
    ❏ Application software
    ❏ Hardware
    ❏ Network
❏ Unintended or "off-label" use
❏ Unauthorized data usage

AI attacks:

❏ Adversarial examples
❏ Deep fake
❏ Denial of service (DoS/DDoS)
❏ Data poisoning
❏ Evasion
❏ Impersonation
❏ Man-in-the-middle
❏ Membership inference
❏ Model backdoor
❏ Model extraction
❏ Model inversion
❏ Third-party Trojan
❏ Training data breach
❏ Transfer learning Trojan

B N H
. A I

Phase II - Identification (Cont.)

- ❏ 3. Notify management and response staff, depending on materiality and in accordance with existing incident response plans:
    - ❏ Encrypted channels
    - ❏ Known impact
    - ❏ Need to know basis
    - ❏ Agree on update/communications cadence