United States
Department of Energy
National Nuclear Security Administration

International Nuclear Security

Responsible Artificial Intelligence for Insider Threat Mitigation

Jessica Baweja, Jon Barr,
Chantell Murphy

23 October 2025

PNNL-SA-216591

INS International Nuclear Security
Reducing Risk of Nuclear Terrorism

Core Principles for Responsible AI in ITM

VALID AND RELIABLE

The system consistently produces accurate results that can be trusted for its intended purpose.

SAFE

The system operates without creating risks to people, property, or the environment.

SECURE AND RESILIENT

The system can protect itself from attacks and continue working even when problems occur.

ACCOUNTABLE

It's clear who is responsible for the system's actions and decisions.

EXPLAINABLE AND INTERPRETABLE

Users can understand how and why the system makes its decisions.

PRIVACY-ENHANCED

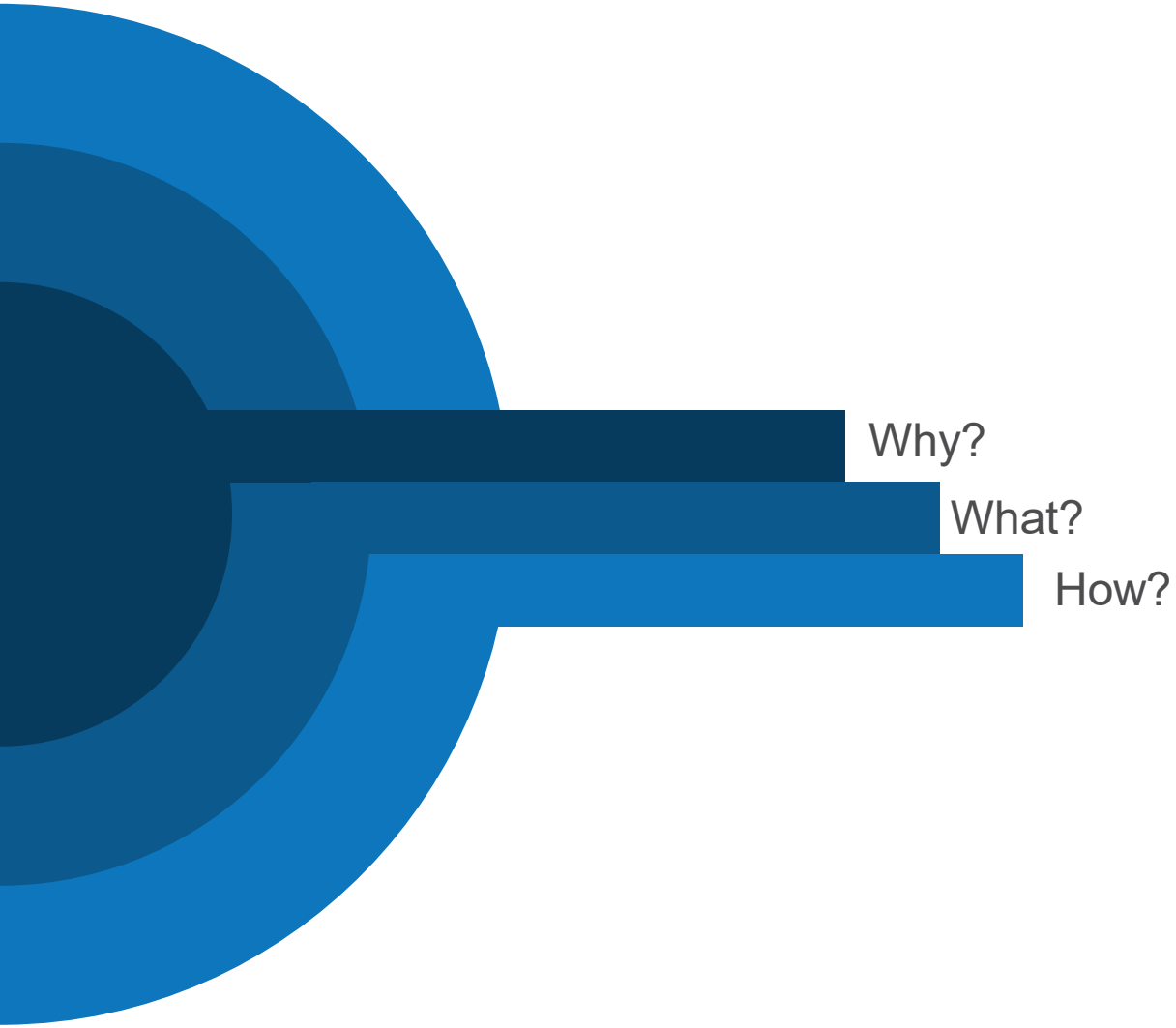
The system protects personal information and respects people's privacy rights.

FAIR

The system treats all people equally and avoids harmful bias.

US National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0).

Validity & Reliability



WHY it matters:

- AI must work reliably in all conditions
- Poor performance creates security weaknesses
- Different uses need different levels of reliability

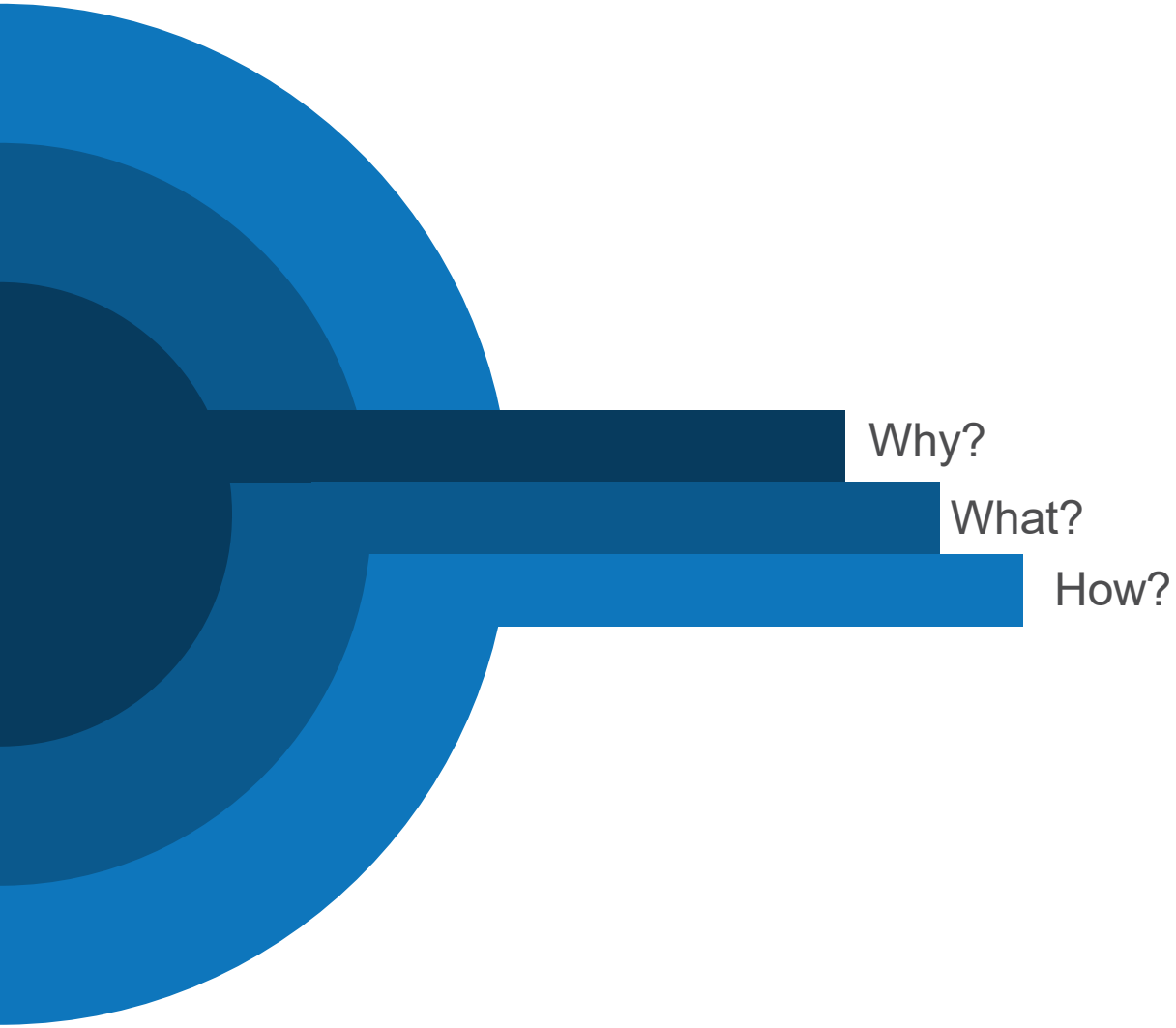
WHAT it involves:

- Ensuring systems work well in all environments
- Balancing between too many and too few alerts
- Regularly checking that systems meet requirement

HOW to implement?

- Test systems in many different conditions
- Regularly check performance with known test cases
- Adjust sensitivity settings as needed
- Track and review performance regularly
- Set minimum performance standards for each application

Safety



WHY it matters:

- Depending too much on AI can create security gaps
- Both false alarms and missed threats cause problems
- Automated systems may create safety risks

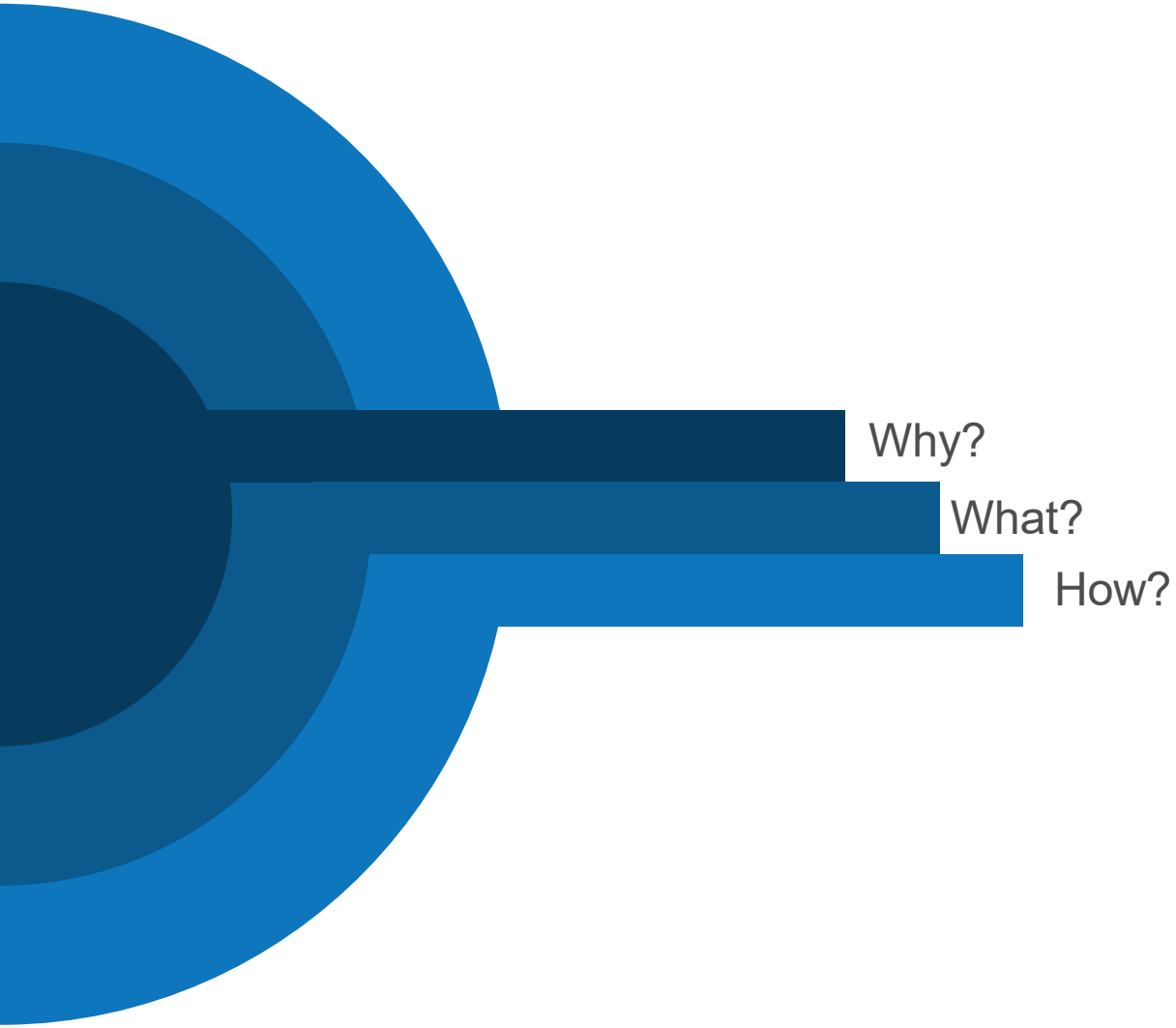
WHAT it involves:

- Making sure AI improves rather than weakens security
- Keeping human skills sharp alongside AI tools
- Preventing errors that could affect safety

HOW to implement?

- Keep traditional security methods working alongside AI
- Create balanced plans for responding to different risks
- Use multiple, overlapping security measures
- Practice security tasks without AI regularly
- Create simple override procedures for automated systems

Security



WHY it matters:

- These systems are targets for attackers
- The systems could be misused to monitor people unfairly
- Access to sensitive data creates risks of internal misuse

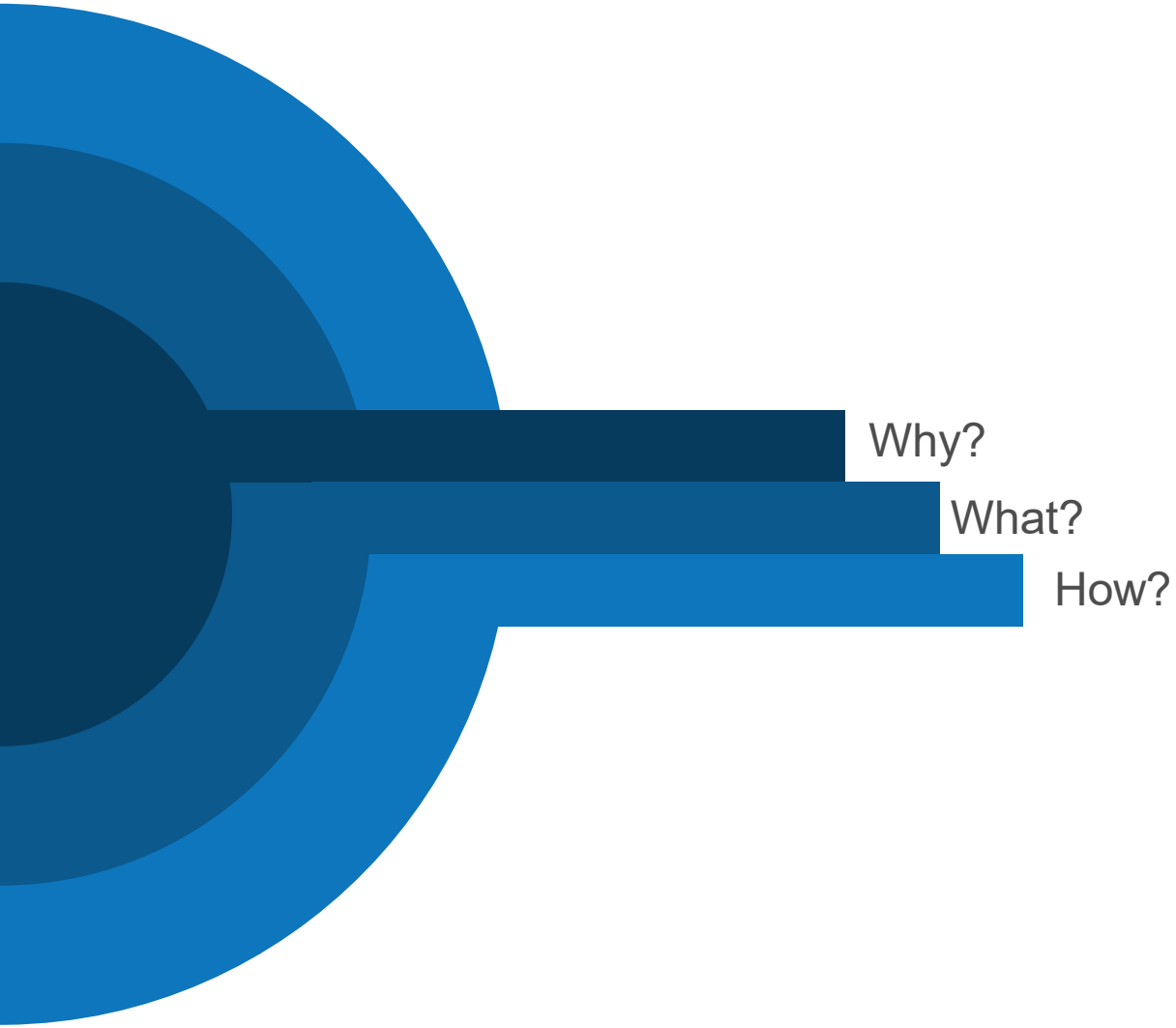
WHAT it involves:

- Protecting AI systems from being hacked or manipulated
- Preventing misuse by people with authorized access
- Maintaining strong security for the AI system itself

HOW to implement?

- Create multiple layers of access controls with detailed logs
- Build safeguards that prevent targeting specific individuals
- Regularly test security controls for weaknesses
- Have independent oversight of system usage patterns
- Create ways to detect and prevent system misuse

Accountability



WHY it matters:

- AI systems help make important security decisions
- Without clear responsibility, no one "owns" the decisions
- Security gaps happen when responsibility is unclear

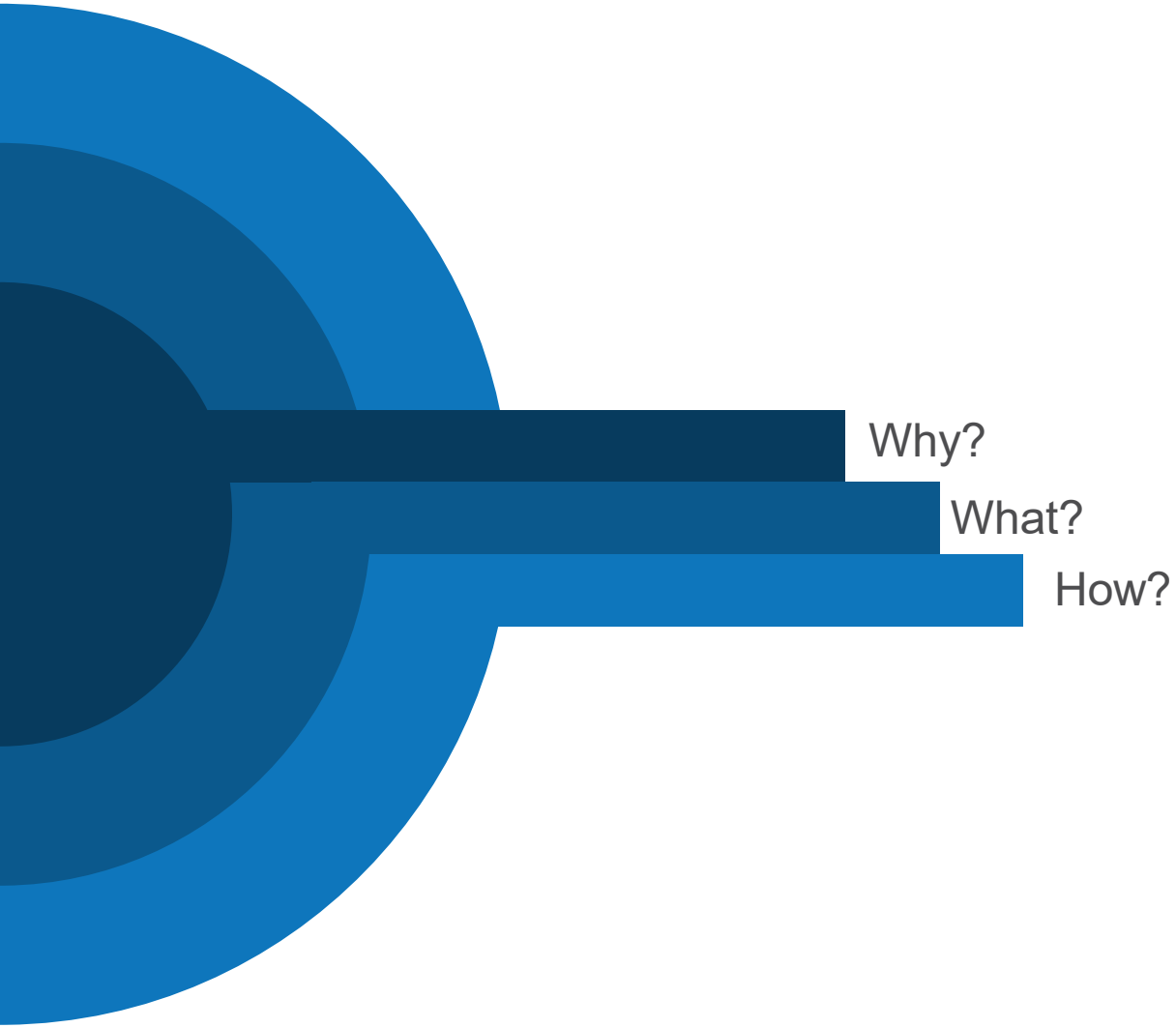
WHAT it involves:

- Clear ownership of AI system decisions
- Defined roles for human oversight
- Transparent decision-making process

HOW to implement?

- Create clear policies showing who reviews AI decisions
- Set up simple frameworks for handling system alerts
- Keep records of all decisions and who made them
- Create clear steps for escalating different types of alerts
- Ensure humans always supervise high-risk AI applications

Explainability



WHY it matters:

- AI systems help make important security decisions
- Without clear responsibility, no one "owns" the decisions
- Security gaps happen when responsibility is unclear

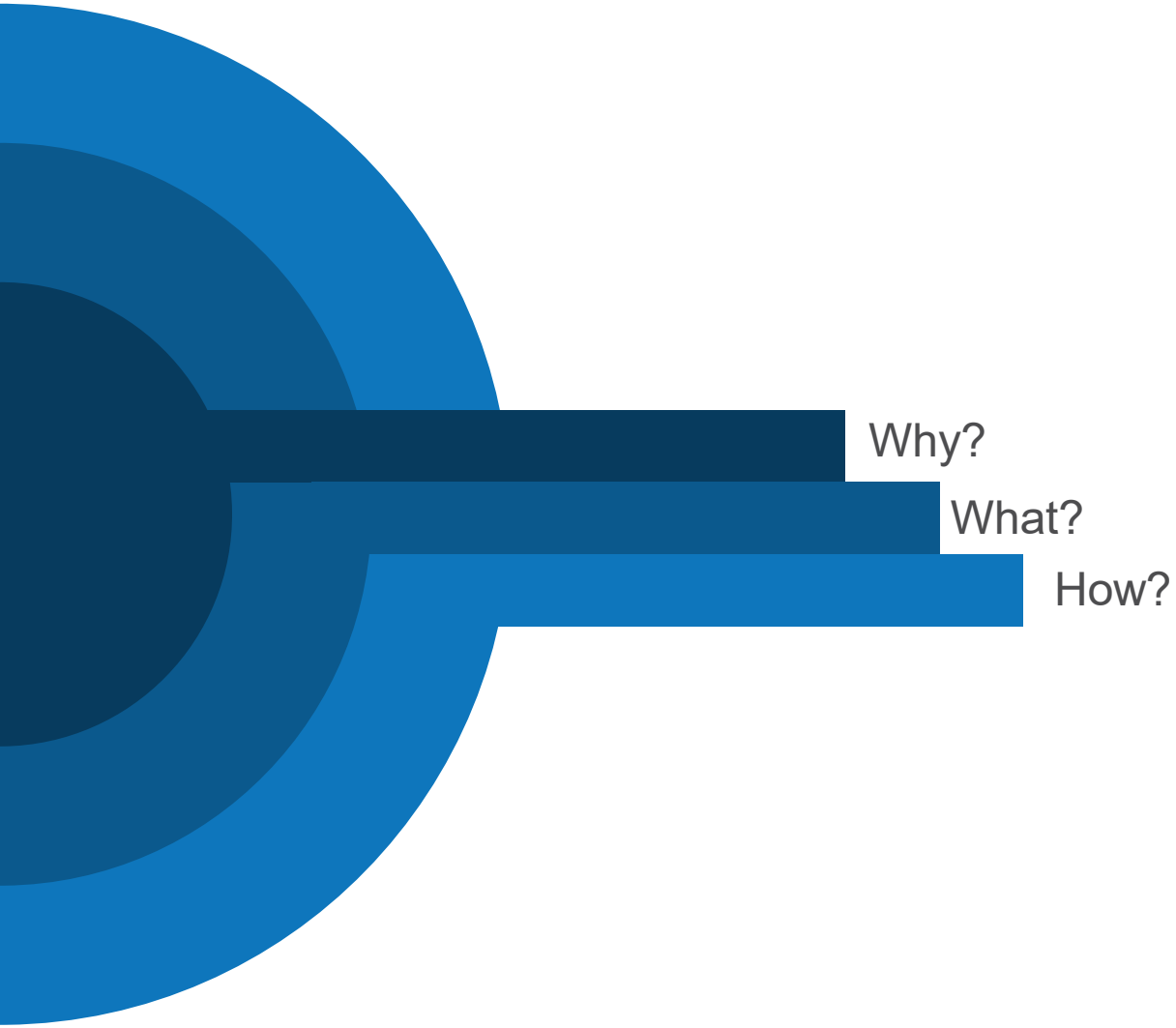
WHAT it involves:

- Clear ownership of AI system decisions
- Defined roles for human oversight
- Transparent decision-making process

HOW to implement?

- Create clear policies showing who reviews AI decisions
- Set up simple frameworks for handling system alerts
- Keep records of all decisions and who made them
- Create clear steps for escalating different types of alerts
- Ensure humans always supervise high-risk AI applications

Privacy



WHY it matters:

- These systems use very personal information
- Privacy problems hurt trust and may break laws
- Systems tend to collect more data over time than needed

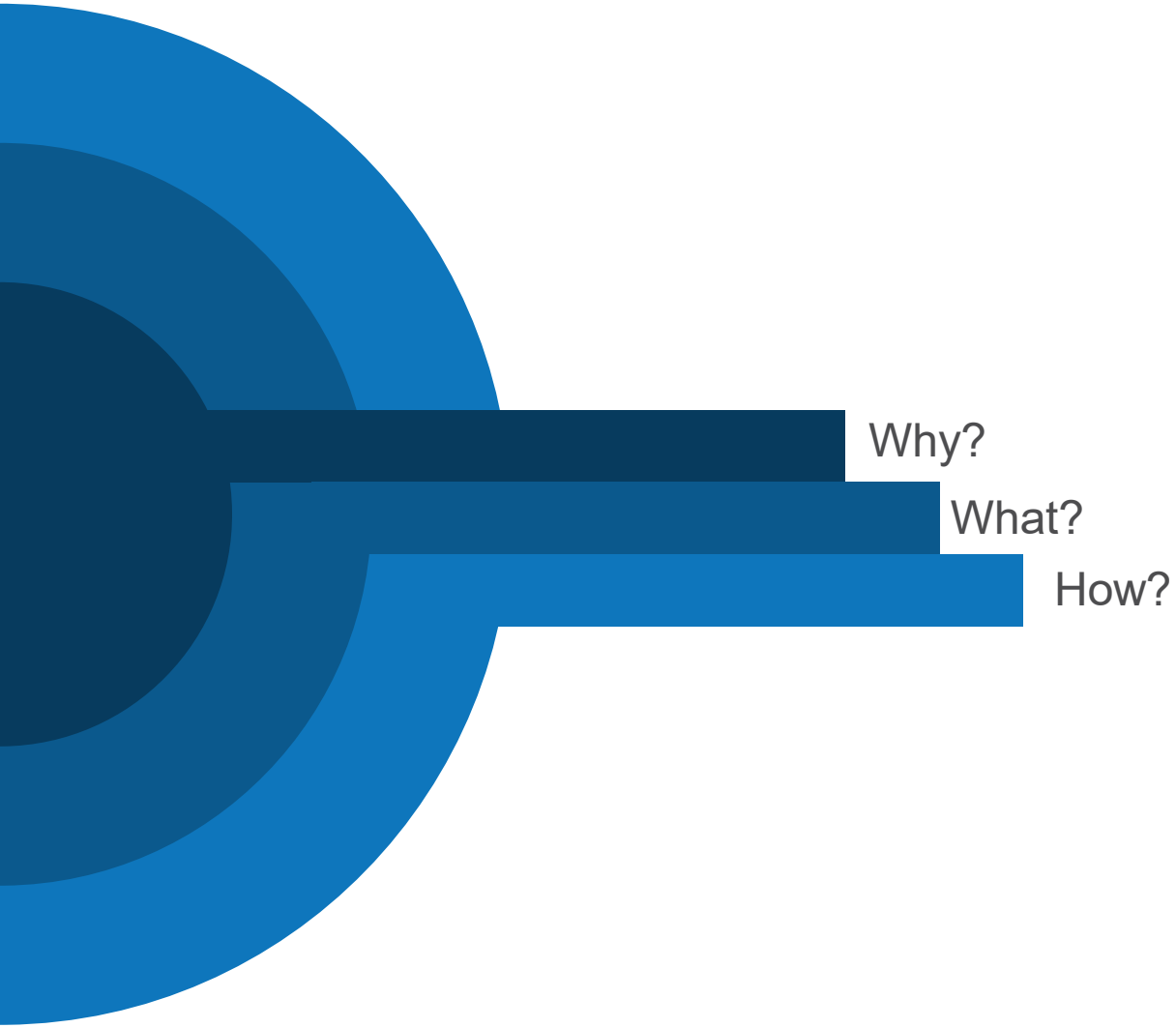
WHAT it involves:

- Protecting personal data from unauthorized access
- Balancing security needs with privacy rights
- Limiting data collection to what's truly necessary

HOW to implement?

- Collect only the information you really need
- Create strict controls for who can access data
- Set clear limits on how data can be used
- Keep security monitoring separate from job evaluations
- Create and follow clear data deletion schedules

Fairness



WHY it matters:

- AI systems may work differently for different groups
- Data from the past often contains bias
- Unfair systems damage trust and create security problems

WHAT it involves:

- Making sure the system works well for all employees
- Preventing old biases from affecting new decisions
- Creating fair processes for everyone

HOW to implement?

- Test system performance across different groups before using it
- Include diverse team members when reviewing alerts
- Create clear standards that consider possible bias
- Use multiple methods to verify information
- Regularly check if the system works equally well for everyone

Risk Management Ecosystem

- Shared Responsibility for AI Risk Management:
  - Security Leadership: Sets risk thresholds and approves policies
  - System Operators: Evaluate daily performance and effectiveness
  - IT/Cyber Teams: Secure systems and ensure data integrity
  - Compliance Officers: Verify regulatory alignment and privacy controls
  - Human Resources: Address workforce concerns and ensure fair application
- Everyone has a role in responsible AI for insider threat mitigation.

Responsible AI Implementation

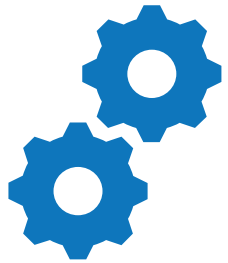
Prerequisites

Clear security policies and procedures

Integration with existing security systems

Defined data governance

Training for security personnel



Process

Define security objectives of the proposed AI system

Assess data availability and quality

Select appropriate AI applications

Implement with proper controls

Monitor performance and adjust (continuous improvement!)

Summary & Key Takeaways

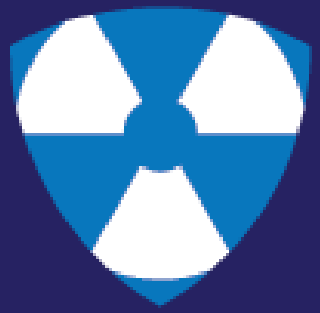
AI offers powerful capabilities for enhancing insider threat mitigation

Different applications carry varying levels of risk and complexity

Responsible implementation requires balancing security benefits with potential risks

Human oversight remains critical, with AI serving as a tool to enhance human capabilities

A principled approach ensures AI strengthens security posture while respecting rights and values



INS International Nuclear Security

Reducing Risk of Nuclear Terrorism