The AIRCTL Project

What it is and why we need AI Preparedness

Emily Soward
FOSSY 2025 Lightning Talks

Maintainers

Emily Soward

- **Job:** Scientist and Tech Lead in AI Security & Privacy
- **Notable Contributions:** HITRUST Cybersecurity Certification for Deployed AI Systems, AWS Well-Architected Framework for Machine Learning, AWS Cloud Adoption Framework for AI/ML/GenAI, AWS Responsible AI Framework, AWS SkillBuilder - Security, Compliance, and Governance for AI

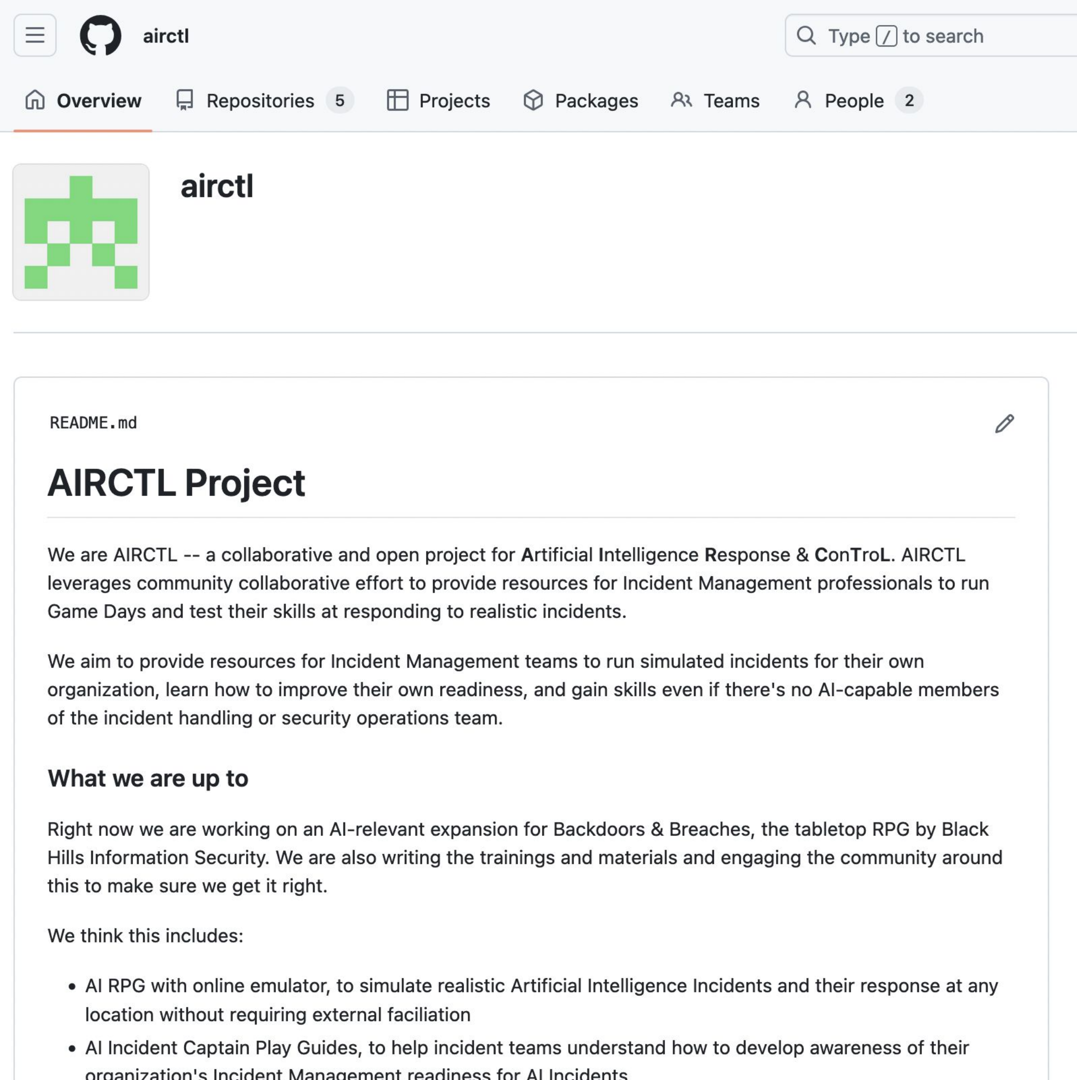
Jonathan Reiter

- **Job:** Principal Engineer
- **Notable Contributions:** Network Protocols Research, detections, and CTI in ICS-OT, Ethernet for Plant Automation (EPA) reversing IEC PAS 62409 and GB/T 20171-2006), DEFCON32 ICS Village “People’s Republic of FieldBus” speaker

What is AIRCTL?

a collaborative and open project for **A**rtificial **I**ntelligence **R**esponse & **C**on**T**ro**L**

https://www.github.com/airctl



AIRCTL Project

We are AIRCTL -- a collaborative and open project for **A**rtificial **I**ntelligence **R**esponse & **C**on**T**ro**L**. AIRCTL leverages community collaborative effort to provide resources for Incident Management professionals to run Game Days and test their skills at responding to realistic incidents.

We aim to provide resources for Incident Management teams to run simulated incidents for their own organization, learn how to improve their own readiness, and gain skills even if there's no AI-capable members of the incident handling or security operations team.

What we are up to

Right now we are working on an AI-relevant expansion for Backdoors & Breaches, the tabletop RPG by Black Hills Information Security. We are also writing the trainings and materials and engaging the community around this to make sure we get it right.

We think this includes:

- AI RPG with online emulator, to simulate realistic Artificial Intelligence Incidents and their response at any location without requiring external faciliation
- AI Incident Captain Play Guides, to help incident teams understand how to develop awareness of their organization's Incident Management readiness for AI Incidents

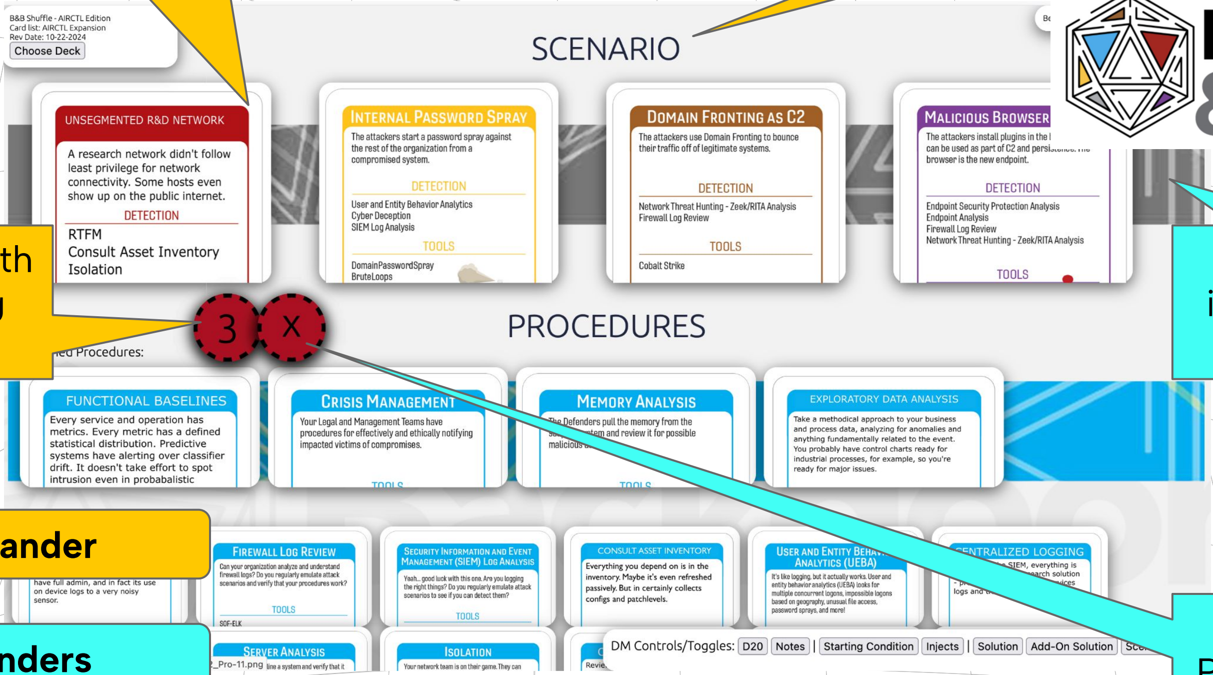
Explain details about the scenario if asked

Model an attacker's path

Simulate chaos with dice roll tracking and Injects

Work together to identify what went wrong

Find the right Procedures without losing ground


B&B Shuffle - AIRCTL Edition
Card list: AIRCTL Expansion
Rev Date: 10-22-2024

SCENARIO

UNSEGMENTED R&D NETWORK

A research network didn't follow least privilege for network connectivity. Some hosts even show up on the public internet.

DETECTION

RTFM
Consult Asset Inventory
Isolation

INTERNAL PASSWORD SPRAY

The attackers start a password spray against the rest of the organization from a compromised system.

DETECTION

User and Entity Behavior Analytics
Cyber Deception
SIEM Log Analysis

TOOLS

DomainPasswordSpray
BruteLoops

DOMAIN FRONTING AS C2

The attackers use Domain Fronting to bounce their traffic off of legitimate systems.

DETECTION

Network Threat Hunting - Zeek/RITA Analysis
Firewall Log Review

TOOLS

Cobalt Strike

MALICIOUS BROWSER

The attackers install plugins in the can be used as part of C2 and persi browser is the new endpoint.

DETECTION

Endpoint Security Protection Analysis
Endpoint Analysis
Firewall Log Review
Network Threat Hunting - Zeek/RITA Analysis

TOOLS

PROCEDURES

ed Procedures:

FUNCTIONAL BASELINES

Every service and operation has metrics. Every metric has a defined statistical distribution. Predictive systems have alerting over classifier drift. It doesn't take effort to spot intrusion even in probabalistic

CRISIS MANAGEMENT

Your Legal and Management Teams have procedures for effectively and ethically notifying impacted victims of compromises.

TOOLS

MEMORY ANALYSIS

The Defenders pull the memory from the tem and review it for possible malicious

TOOLS

EXPLORATORY DATA ANALYSIS

Take a methodical approach to your business and process data, analyzing for anomalies and anything fundamentally related to the event. You probably have control charts ready for industrial processes, for example, so you're ready for major issues.

have full admin, and in fact its use on device logs to a very noisy sensor.

FIREWALL LOG REVIEW

Can your organization analyze and understand firewall logs? Do you regularly emulate attack scenarios and verify that your procedures work?

TOOLS

SOF-ELK

SECURITY INFORMATION AND EVENT MANAGEMENT (SIEM) LOG ANALYSIS

Yeah... good luck with this one. Are you logging the right things? Do you regularly emulate attack scenarios to see if you can detect them?

TOOLS

CONSULT ASSET INVENTORY

Everything you depend on is in the inventory. Maybe it's even refreshed passively. But in certainly collects configs and patchlevels.

USER AND ENTITY BEHA ANALYTICS (UEBA)

It's like logging, but it actually works. User and entity behavior analytics (UEBA) looks for multiple concurrent logons, impossible logons based on geography, unusual file access, password sprays, and more!

CENTRALIZED LOGGING

SIEM, everything is arch solution vices logs and

SERVER ANALYSIS

_Pro-11.png line a system and verify that it

ISOLATION

Your network team is on their game. They can

Revie

DM Controls/Toggles:

Incident Commander

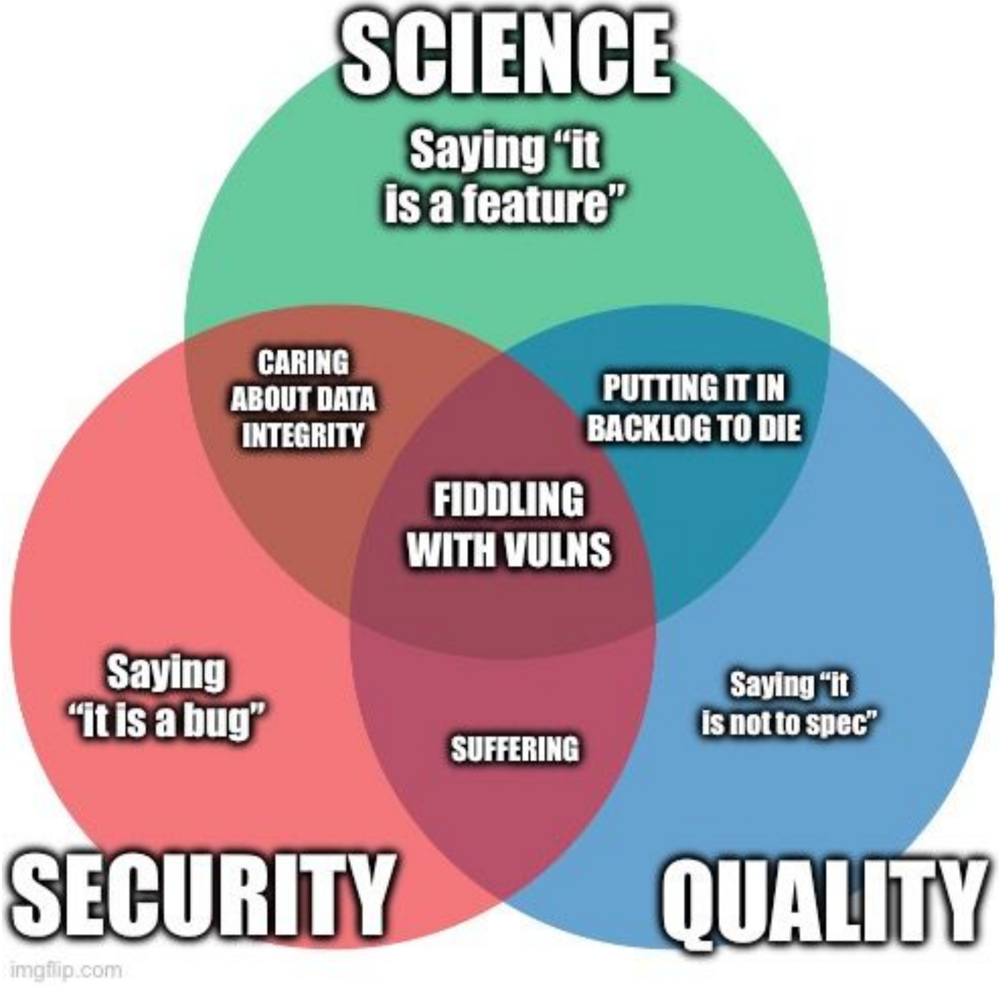
Incident Responders

Incident Handling: Managing active Incident Response

Why AIRCTL?

Despite the enormous effort in so many ways and channels to advance security of AI by characterizing the challenges of AI, we know that readiness is critical to the ability of anyone to address risks once realized.

We see the need to bring the community together to solution and prepare for bad days with AI systems and programs, may we never see them.



Emily SOward. Creative Commons. https://github.com/eksoward/culture-image-assets/blob/main/security-science-quality-venn.jpeg

What do we do?

We provide resources for Incident Management teams to get started even if there’s no AI-capable members of the incident handling or security operations team.

TRAIN: run simulated incidents for their own organization

IDENTIFY: learn how to improve and identify gaps through play

PREPARE: gain skills and improve readiness

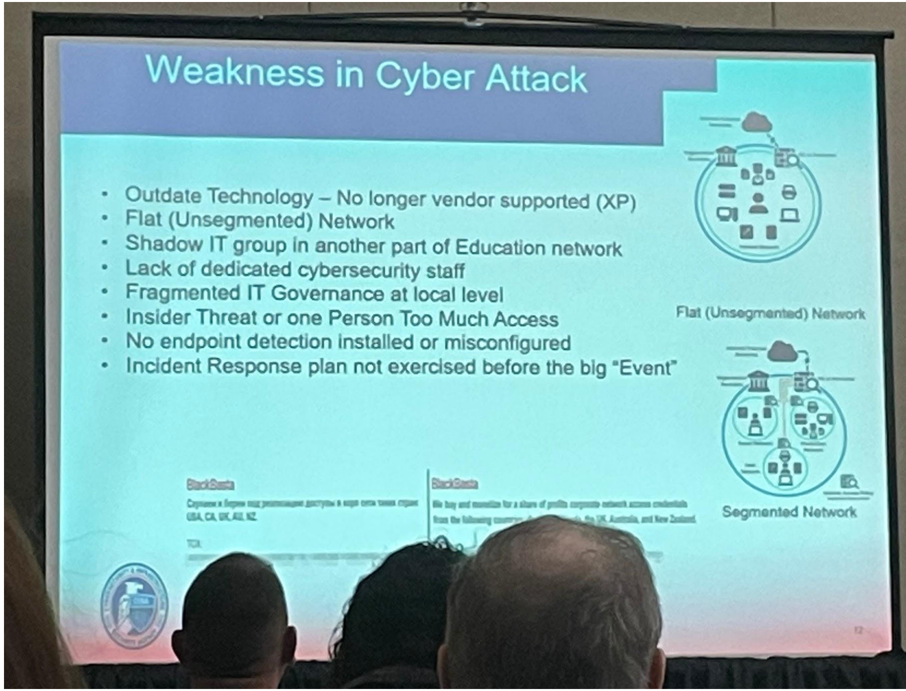
Emily SOward. Creative Commons. https://github.com/eksoward/culture-image-assets/blob/main/seo-ddos.jpeg

What could gO wrong in R&D?

- Standard Issues Apply
- R&D Networks are typically treated as disposable in AI development
- AI Application Layer and Infrastructure for hosting/training AI are outside of scan sweeps
- There’s a lot of data
  - (and information)

Example Scenarios

- Suddenly your competitor is launching the same product - (R&D Compromise) - Espionage
- Cryptomining on your R&D Cluster (R&D Compromise) - Resource Hijacking



Weakness in Cyber Attack

- Outdate Technology – No longer vendor supported (XP)
- Flat (Unsegmented) Network
- Shadow IT group in another part of Education network
- Lack of dedicated cybersecurity staff
- Fragmented IT Governance at local level
- Insider Threat or one Person Too Much Access
- No endpoint detection installed or misconfigured
- Incident Response plan not exercised before the big “Event”

Photo taken at talk by DHS CISA Region 10 & FBI Cyber on Cyber in Critical Infrastructure. INterface 2024

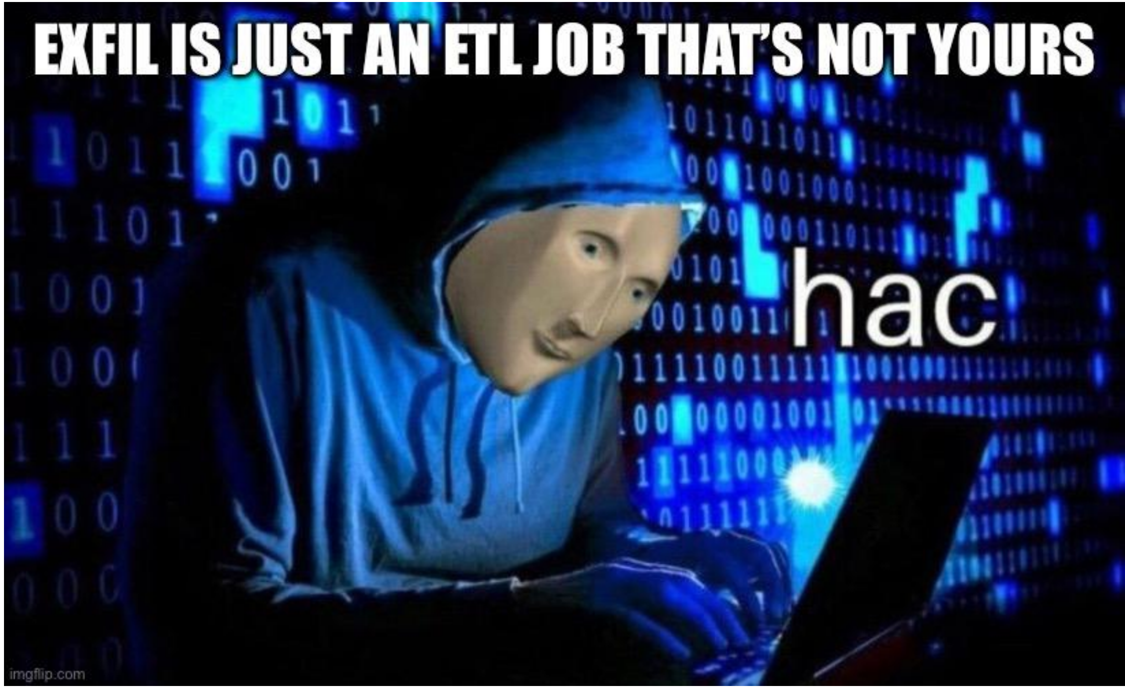
Material Breach

AI is used inline of a material breach, spill, or aggregation

- An AI system experiences an intrusion or exposure
- Data and information is used in an unauthorized system
- Data is aggregated and creates a different class of data

Example Scenarios

- Application Logs with Full-Take I/O Stolen (Material Breach) - Exfiltration
- Your Field Team is writing really great email campaigns with the client database - (Material Breach) - Insider Risk



Emily SOward. Creative Commons. https://github.com/eksoward/culture-image-assets/blob/main/seo-ddos.jpeg

AI Specific Issues

Defacement and Manipulation

Damaging AI utility and integrity

- Poisoned to Death by uselessness - (AI Defacement & Manipulation) - Defacement
- Your search engine has been trained for free to sell soap - (AI Defacement & Manipulation) - Manipulation

Exploitation

Directly using AI as a vector

- Let me send you a copy / talk to me via the logs - (AI Exploitation) - Complex exfiltration / C2 & comms
- Using your AI App to persist chat conversations (AI Exploitation) - Resource Hijacking

THANKS!

https://www.github.com/airctl

Emily Soward | *emily@airctl.org* | AIRCTL

Jonathan Reiter | *jonathan@airctl.org* | AIRCTL