

INCIDENT RESPONSE FOR ARTIFICIAL INTELLIGENCE

PLAYTEST NEW TABLETOP EXERCISES WITH THE AIRCTL PROJECT FOR FUN AND STICKERS

Emily Soward & Jonathan Reiter

MEET YOUR INSTRUCTORS

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



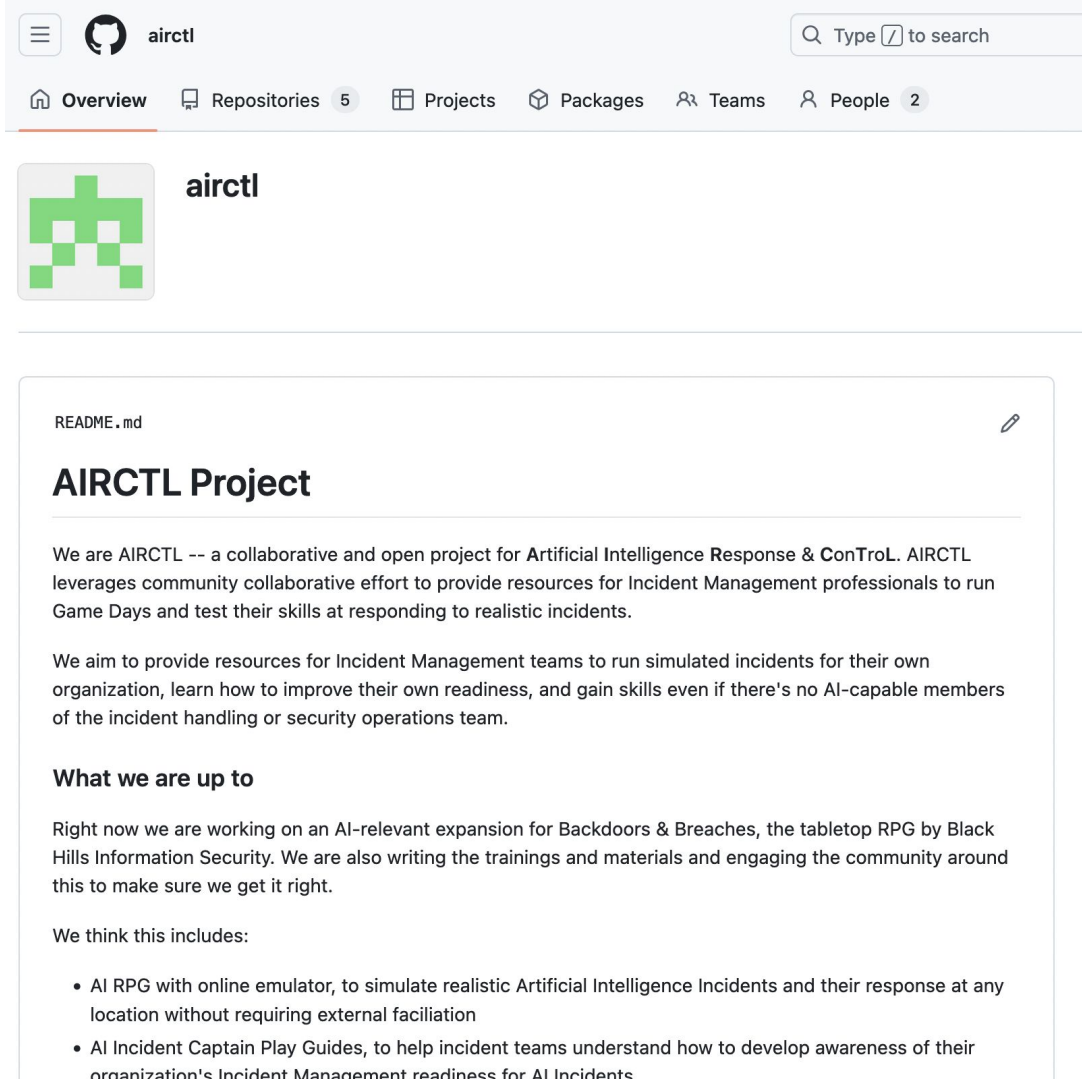
AI IS HAVING A
MOMENT RIGHT NOW...



WHAT IS AIRCTL?

a collaborative and open
project for **Artificial
Intelligence Response &
ConTroL**

<https://www.github.com/airctl>



The screenshot shows the GitHub repository page for 'airctl'. At the top, there's a navigation bar with the GitHub logo, the repository name 'airctl', and a search bar. Below this, there's a secondary navigation bar with tabs for 'Overview', 'Repositories' (5), 'Projects', 'Packages', 'Teams', and 'People' (2). The main content area features a green and white pixelated logo for 'airctl'. Below the logo, there's a section for the 'README.md' file, which includes the title 'AIRCTL Project' and the following text:

We are AIRCTL -- a collaborative and open project for **Artificial Intelligence Response & ConTroL**. 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 facilitation
- AI Incident Captain Play Guides, to help incident teams understand how to develop awareness of their organization's Incident Management readiness for AI Incidents

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.



imgflip.com

EMILY SOWARD. CREATIVE COMMONS. [HTTPS://GITHUB.COM/EKSOWARD/CULTURE-IMAGE-ASSETS/BLOB/MAIN/SECURITY-SCIENCE-QUALITY-VENN.JPEG](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](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"

Flat (Unsegmented) Network

Segmented Network

Stock Alerts

Скорее всего ваши инвестиции будут в убыток. Ваши акции упали.

USA, CA, UK, AU, NZ

TCR

Мы бы и рекомендовали для a share of profits corporate network access credentials from the following countries: USA, Canada, the UK, Australia, and New Zealand.

12

PHOTO TAKEN AT TALK BY DHS CISA REGION 10 & FBI CYBER ON CYBER IN CRITICAL INFRASTRUCTURE, INTERFACE 2024

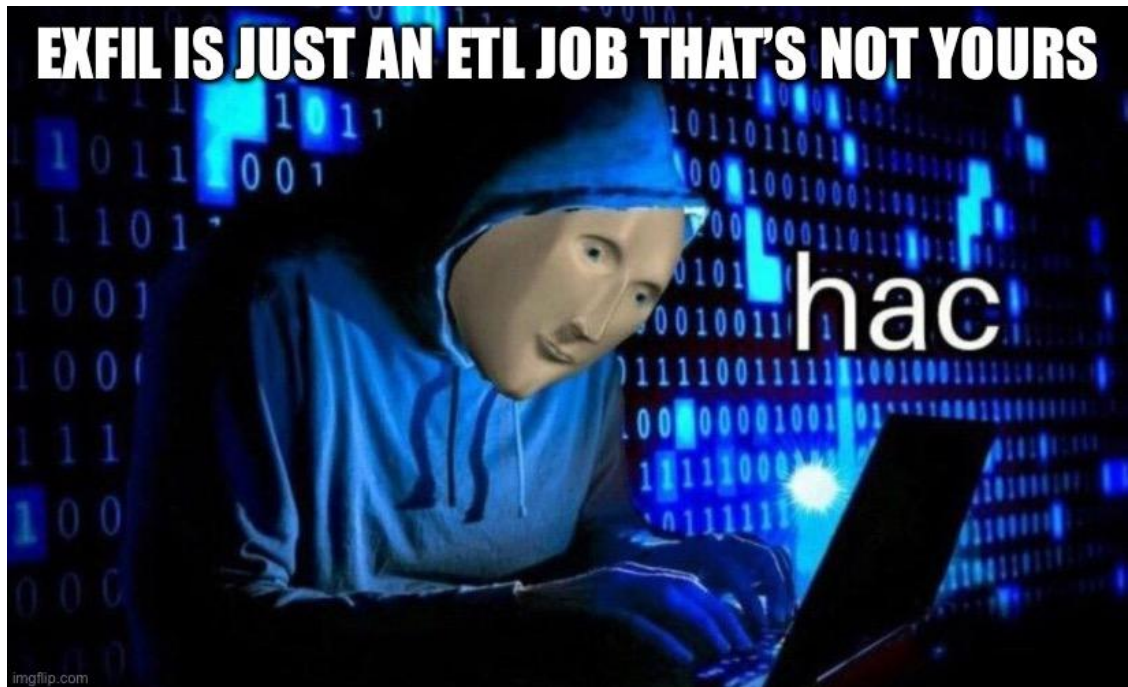
MATERIAL BREACH

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](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