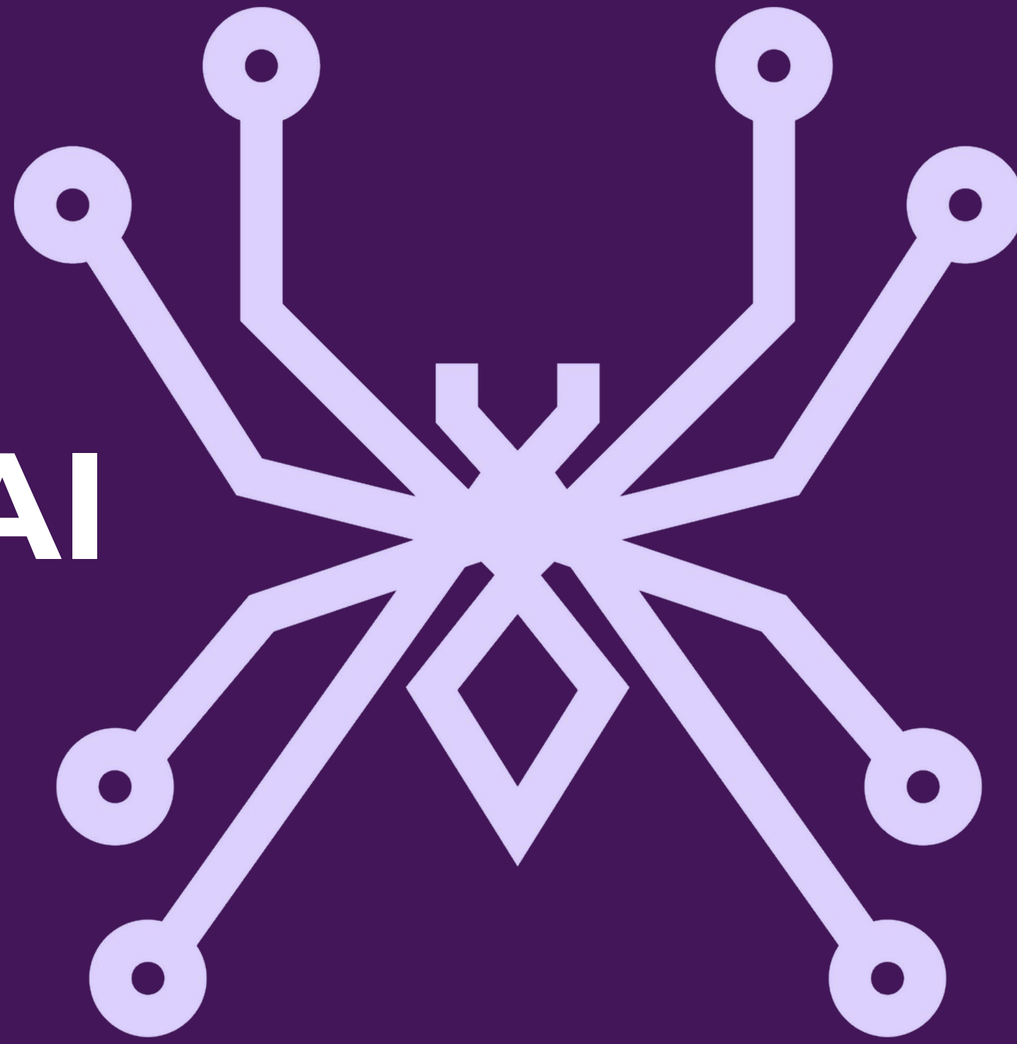
Susanna Cox
AI SECURITY RESEARCHER

Threat Modeling AI

Beyond the Hype & Theater to Proactive Security



About me

I help people apply **mission-critical, secure-by-design** engineering principles to AI.

OWASP AI Exchange

Policy & Standards | ISO ++

EU AI Act Requirements Team

Gap Analysis Research Lead

Red Team → Purple Team

AIX Core Red Team Member

Cloud Security Alliance *Agentic Red Team Guide* Contributor

SOCMED

Find me: **@disesdi** on most platforms

I’m looking forward to connecting with you!

Follow me on LinkedIn

in/disesdi

Subscribe to my Channel

youtube.com/@disesdi

Angles of Attack

The AI Security Intelligence Brief

Go beyond the hype to get real intel on the bleeding edge of AI Security & Policy

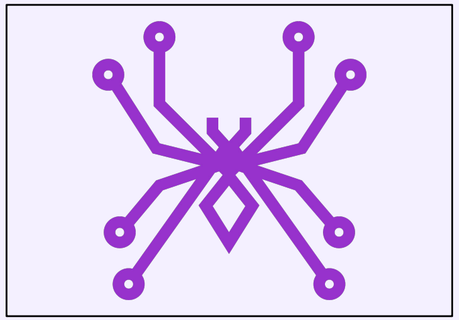
disesdi.substack.com

Subscribe to my newsletter

What Is Threat Modeling?

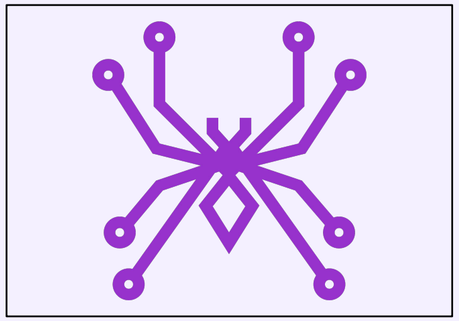
“Threat modeling is analyzing representations of a system to highlight concerns about security and privacy characteristics.”

Threat Modeling Manifesto



A structured, systematized approach

Clearly articulated, well documented, & consistently applied



A tool to contextualize risks

What vectors? How likely is each attack? And what might the effects be?

What Is Threat Modeling?

A means of preparing & documenting mitigations

Once we know what are threats are, we can begin to prepare our response

The original Purple Team technique

How we get defenders to think like attackers: Threat modeling

The 4 Questions of Threat Modeling

1. What are we working on?

Understanding the systems at hand, so we can understand risk

2. What could go wrong?

Quantifying, understanding, & communicating threats

3. What are we going to do about it?

Making mitigations & responsibilities explicit

4. Did we do a good job?

Defining outcomes: How will we know we've succeeded?

Shostack, Adam. Threat Modeling: Designing For Security. John Wiley & Sons, 2014.

When To Threat Model?

AI systems **require** a purple team approach throughout the development lifecycle

The Secure AI Development Lifecycle

Desiloing Data + Security teams requires threat modeling at every stage of development

Cybersecurity Meets AI Security

Uniting Red + Blue teams means creating shared understanding of the new threat landscape

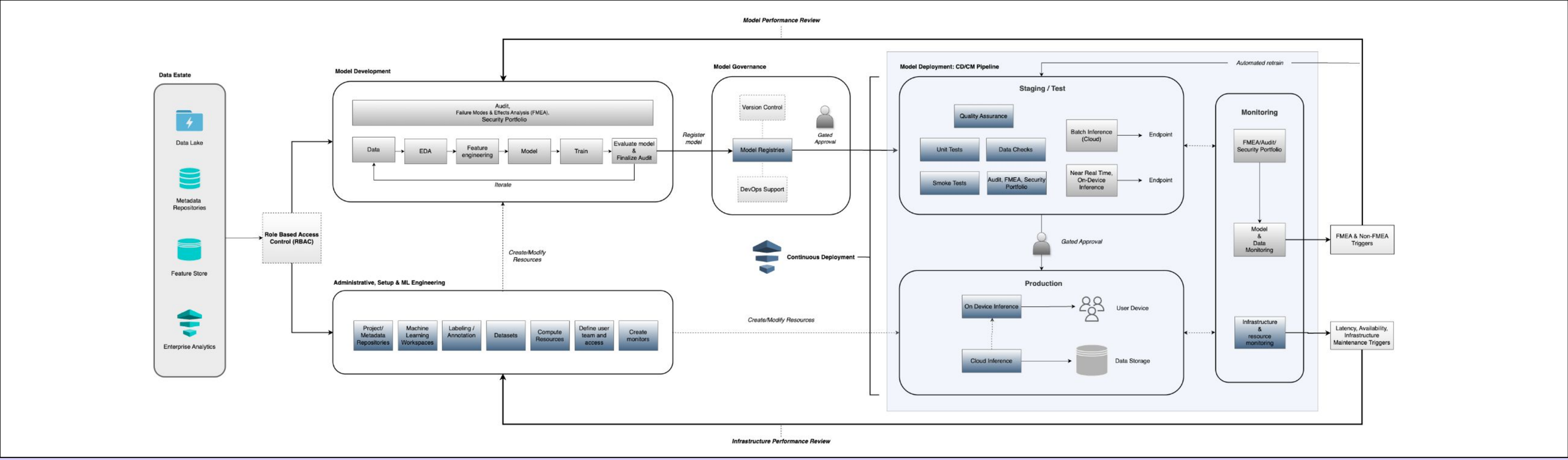
What are we working on?

AI Architectures Are Different

Understanding AI Architectural Patterns

Predictive AI, Generative AI, and Agentic AI deployments all have canonical architectural patterns

Understanding these patterns–and their operationalization–is key to securing AI systems



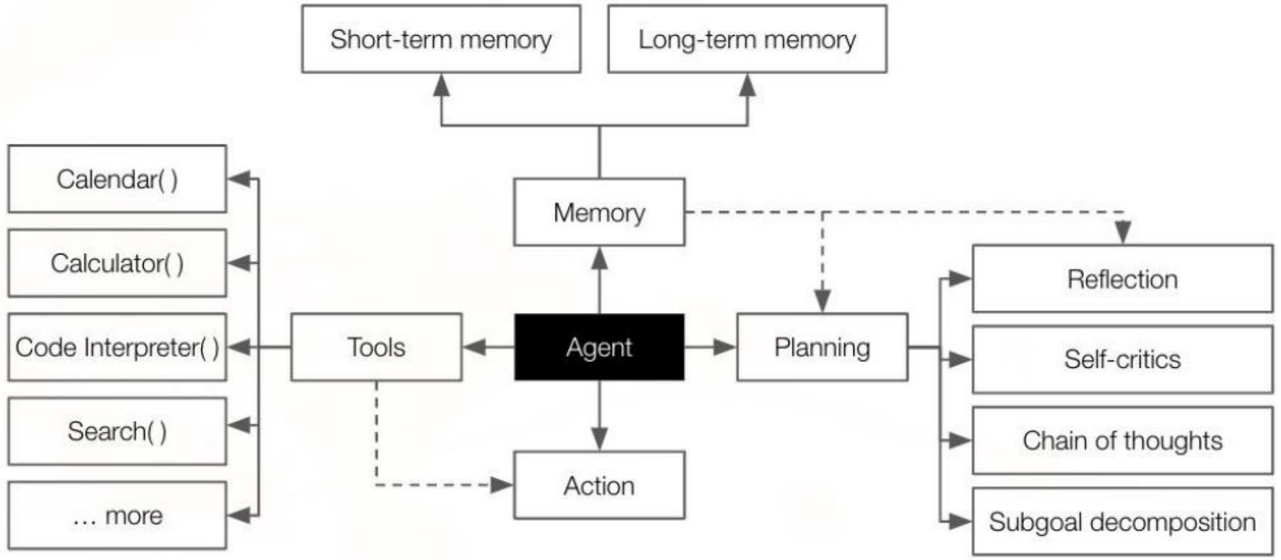
Predictive AI

Architectural Patterns

MLOps & MLSecOps

Structures like continuous monitoring, model and data registries, and gated model approval

1. What are we working on?



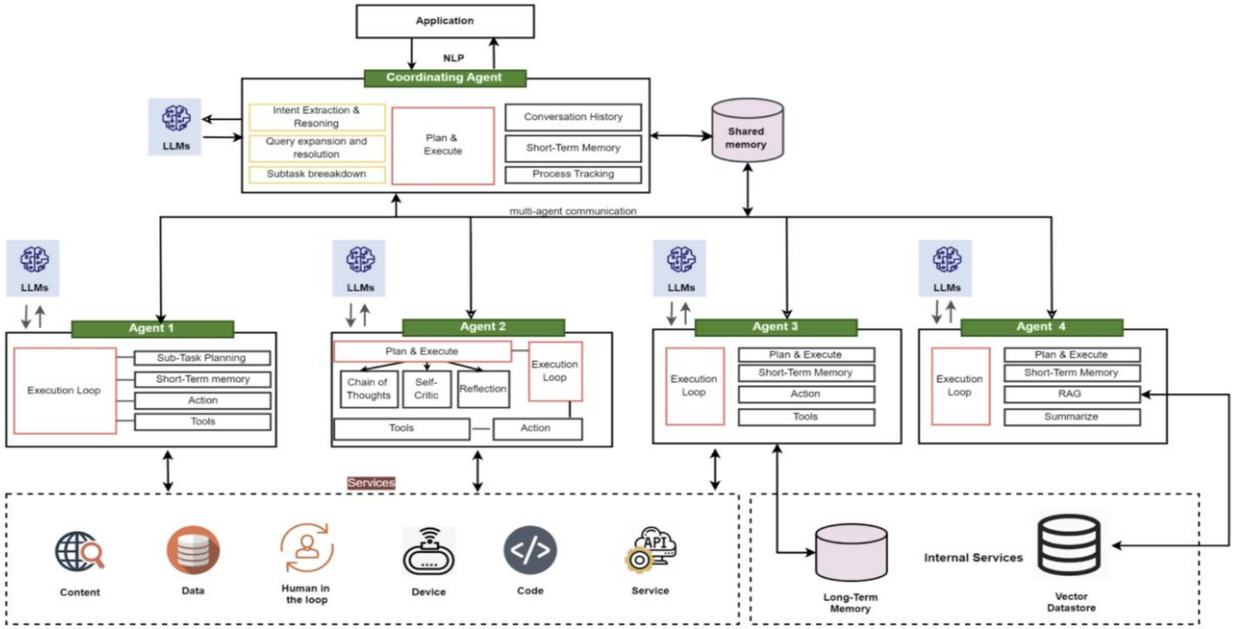
Agentic AI

Architectural Patterns I

Single Agent Deployment

Structures like planning, reflection, memory & subgoal decomposition

1. What are we working on?



Agentic AI

Architectural Patterns II

Multi-Agent Deployment

Structures like coordinating agent(s), planning, & validation

1. What are we working on?

What could go wrong?

AI threats are new.

Understanding the new threats in AI systems requires models that recognize the importance of data & scale.

There are many ways to organize and understand security threats.
AI security is no exception. In AI security, traditional models meet a new reality.

What could go wrong?

The New AI Threat Landscape

Data is now the vector.

2012: Data is the “**new oil**”

2024: Data is the **new attack vector**

What could go wrong?

The CIA Model In The New AI Era

Traditional CIA: Confidentiality, Integrity, Availability

- What does this mean for AI?
- NIST AI 100-2e2025 Taxonomy refers to CIA model ++

What could go wrong?

The CIA Model In The New AI Era

Availability Breakdown

- **Data poisoning**: when the attacker controls a fraction of the training set
- **Model poisoning**: when the attacker controls the model parameters
- **Energy-latency attacks** via query access

What could go wrong?

The CIA Model In The New AI Era

Privacy Compromise at Deployment Time

Attacker objectives: compromising the privacy of training data, such as

- **Data Reconstruction**
- **Membership-Inference Attacks**
- **Data Extraction (GenAI)**
- **Property Inference (data distribution)**
- **Model Extraction**

What could go wrong?

An AI-Tailored Approach: Understanding Threats in Their Lifecycle Phases

Many AI-specific vulnerabilities occur during key phases in the AI development lifecycle

- Training Phase
- Deployment Phase

What could go wrong?

AI Threats in Their Lifecycle Phases

AI-specific lifecycle threats: Development time threats, Threats through use, & Runtime security threats

- Training time: Poisoning
- Deployment time: Evasion & Privacy, Model theft

What are we going to do about it?

Applying AI Controls in the Lifecycle

Mapping & Securing the Attack Surface: Operationalization & Data Intelligence

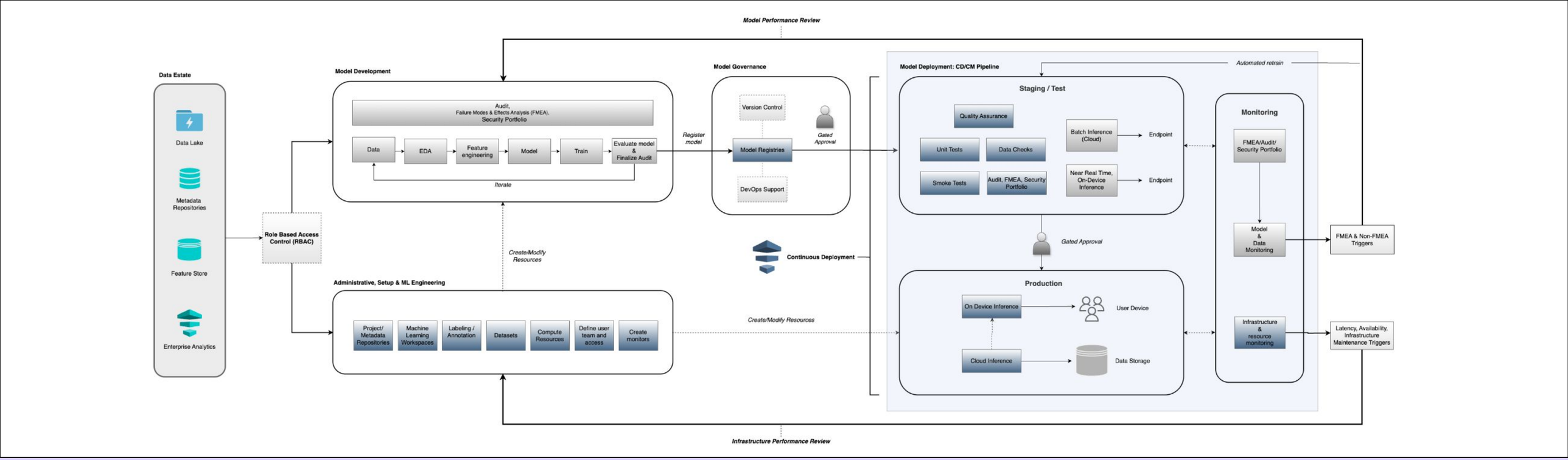
- Use the OWASP AI Exchange
- Apply Controls & Mitigations to Your Data & Process Flows

What are we going to do about it?

Applying AI Controls in the Lifecycle: 3 Steps

Three steps to understanding your AIML system attack surfaces

1. Know your data **flows**
2. Know your data **provenance**
3. Know your data **governance**



AI Architectures Revisited

The Importance of Lifecycle Ops

Security controls must be **operationalized** in order to be effective.

Attacks happen **at scale**: monitoring, data ops, ++ become critical

3. What are we going to do about it?

Did we do a good job?

Quantifying Success In AI Threat Modeling

Three Questions To Ask:

1. Is it **secure**?
2. Can we **operationalize**?
3. Does it **scale**?

Key Takeaways

Threat model throughout the development lifecycle

Proactive security always beats reactive response. An ounce of prevention beats a pound of cure.

Use AI-specific tools & frameworks like the OWASP AIX

The OWASP AI Exchange is your go-to for SOTA resources, made for AI. Don’t make the perfect the enemy of the good–start now.

Remember the importance of Data, Ops & Scale

New tech, new paradigm: AI security requires robust operationalization, and recognition that data is now a threat vector.

Susanna Cox
AI SECURITY RESEARCHER

Thank You!

Find me: **@disesdi** on most platforms

I’m looking forward to connecting with you!

Follow me on LinkedIn

in/disesdi

Subscribe to my Channel

youtube.com/@disesdi

Thanks so much! Let’s connect.

Susanna Cox
AI SECURITY RESEARCHER