The New Frontline: Your AI Is Under Attack

Algorithms vs. Algorithms
The End of Easy Trust
Welcome to the Era of AI Security (AISec)

K7 ACADEMY

Your 25-Minute AISec Upgrade

Why Now: The AI shift is here

The Model: A simple way to think about AI

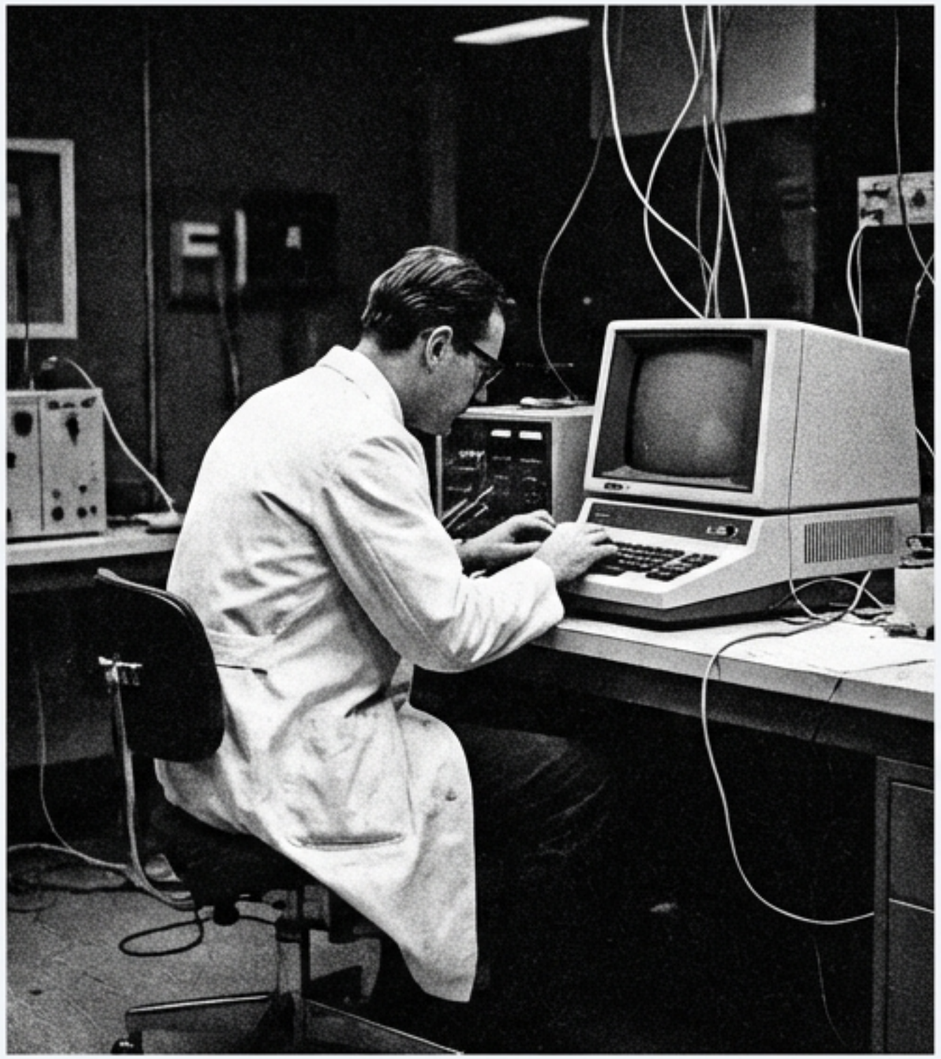
The Risks: 5 threats that actually matter

The Controls: 3 ‘gates’ to secure any AI app

Your Roadmap: Go from learner to builder

K7 ACADEMY

We’ve Crossed The AI Rubicon



AI is no longer just a lab experiment.

It’s now a core production system.

This creates a “Red Queen Race.”

Attackers and defenders are both using AI.

You have to run just to stand still.

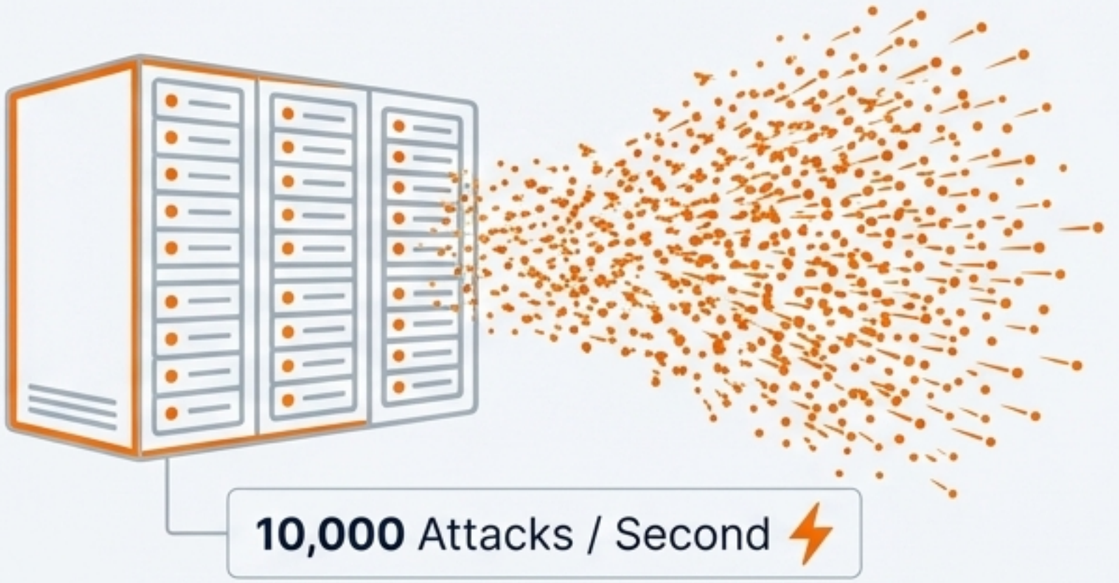
K7 ACADEMY

The Threat: Speed and Scale Beyond Humans

Old Threat: Human-speed attacks (hours/days)

- ~~Hyper-personalized~~
- ~~Infinitely scalable~~
- ~~Impossible for humans to review manually~~

New Threat: Machine-speed attacks (milliseconds)



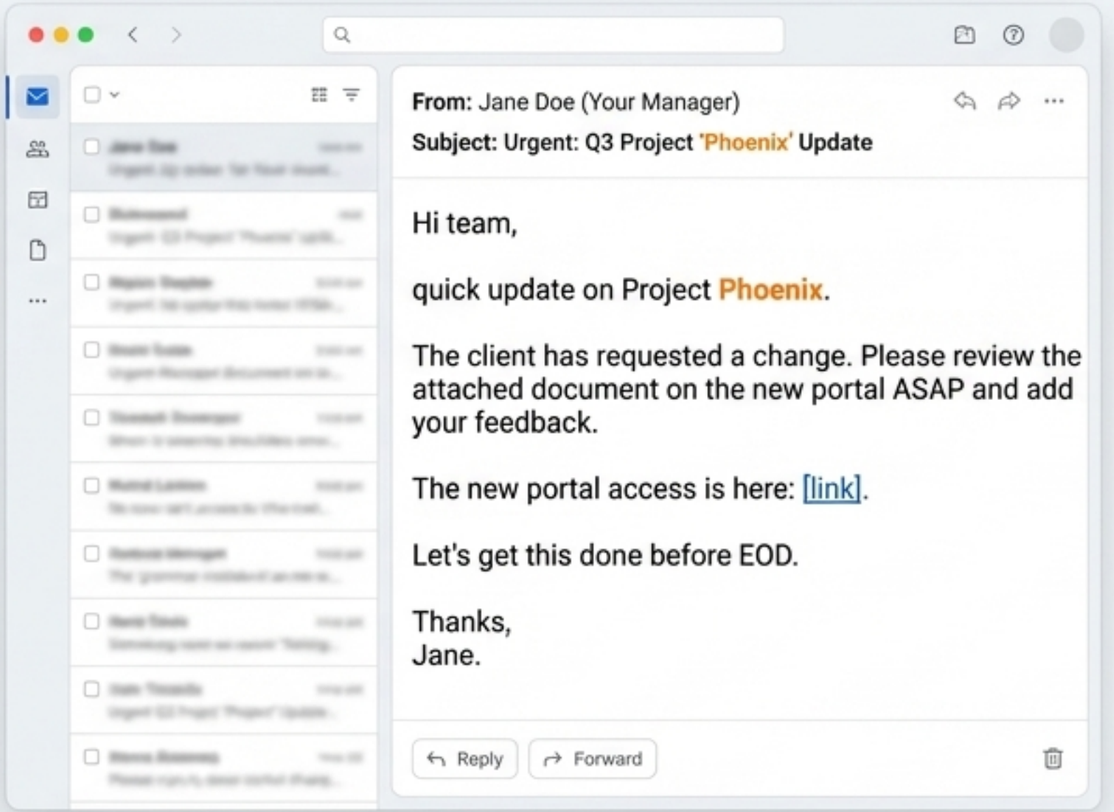
- ✓ Hyper-personalized
- ✓ Infinitely scalable
- ✓ Impossible for humans to review manually

Attack In The Wild: The Deepfake CFO

- Target: A multinational firm in Hong Kong
- Method: AI-cloned voices and video
- The ‘Tell’: A convincing group video call
- Result: A $25 million fraudulent transfer
- The employee was the only real human on the call.

Attack In The Wild: Phishing On Autopilot

- **Tool:** ‘WormGPT,’ an ‘evil’ LLM 😈
- **Goal:** Mass-produce believable phishing emails
- **Method:** Scrapes context, mimics writing styles
- **Scale:** From 100 emails a day to 100,000+
- The "grammar mistakes" are gone.



How To Think About AI: Brain, Memory, Hands

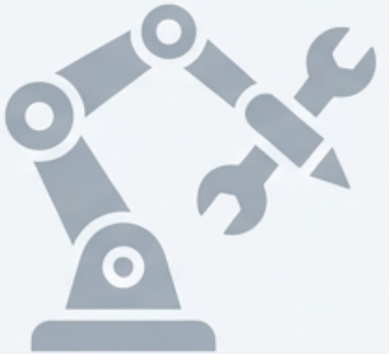
The Memory

Private data via RAG.
(The private library)

The Brain

The Large Language Model
(LLM) itself.
(The core reasoner)

The Hands



Agents and tools for action.
(The connection to the world)

The New Kill Chain: 5 Prominent Risks

1. Prompt Injection: Tricking the AI with hidden instructions.
Consequence: Bypassing safety filters.

2. Insecure Output Handling': Trusting AI output blindly.
Consequence: Code execution on your system.

3. Training Data Poisoning: Corrupting the AI's 'Memory.'
Consequence: Biased or backdoored responses.

The New Kill Chain: 5 Prominent Risks (Cont.)

4. Model Theft: Stealing the ‘Brain’ itself.
Consequence: Loss of your most valuable IP.

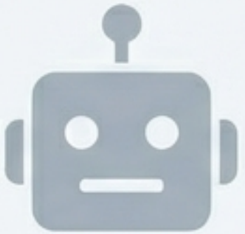
5. Over-reliance: Humans trusting the AI too much.
Consequence: Critical errors from hallucinations go unnoticed.

GATE 1

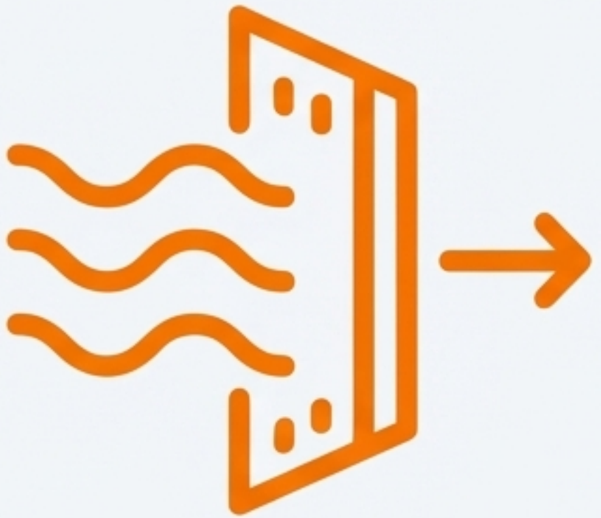
The 3 Gates: How We Secure The System

IDENTITY

- Who (or what) is making the request?
- Authenticate every AI, agent, and user.
- Apply the principle of least privilege.
- Control *what* data can even be retrieved.



GATE 2



The 3 Gates: Guarding The Flow

DATA

- What information is flowing in and out?
- Use allowlisted sources for RAG.
- Redact sensitive PII and secrets automatically.
- Create immutable audit trails for every query.



GATE 3

The 3 Gates: Vetting The Action

ACTIONS

- What is the AI trying to do?
- Strictly limit which tools the AI can use.
- Require a human-in-the-loop for risky actions.
- Validate outputs before they are executed.

Your Roadmap: From Zero to AISec Hero

Phase 1: LEARN (This Month)

- Read OWASP Top 10 for LLMs.
- Follow 2-3 AISec experts on social media.

Phase 2: BUILD (Next 3 Months)

- Build one of the portfolio projects on the next slide.
- Focus on implementing the “3 Gates.”

Phase 3: PROVE (Next 6 Months)

- Contribute to an open-source AISec tool.
- Write a blog post explaining your project.

Build Your Future. Secure Theirs.

Your Mission: Become the security-minded builder everyone wants to hire.

LLM Firewall

A proxy that filters prompts and responses.

Gate 2 Gate 3

Permission-Aware RAG

A chatbot that respects user permissions.

Gate 1

Secure Agent

An AI that requires approval for risky actions.

Gate 1 Gate 3