What are you doing, Dave?

Lessons for AI Security Preparedness

Shawn Richardson | CAMLIS 2023
Looking for Failure Modes
And addressing them
Lesson 1: People Fear the Unknown
AI is no different
“Human sacrifice! Dogs and cats living together! Mass hysteria!”

AI is no different

The price we'll pay for our AI future: more loneliness

Pausing AI Developments Isn't Enough. We Need to Shut it All Down

Artificial intelligence could pose serious threats to national and global security. Case in point: AI could one day enable a lone actor to set off a deadly pandemic.

Here are some ideas for safeguards to put in place now—so such scenarios never happen.

Frame nothing as a conflict between national interests, have it clear that anyone talking of arms races is a fool. That we all live or die as one, in this, is not a policy but a fact of nature. Make it explicit in international diplomacy that preventing AI extinction scenarios is considered a priority above preventing a full nuclear exchange, and that allied nuclear countries are willing to run some risk of nuclear exchange if that’s what it takes to reduce the risk of large AI training runs.

European Commission
@EU_Commission

Mitigating the risk of extinction from AI should be a global priority.

And Europe should lead the way, building a new global AI framework built on three pillars: guardrails, governance and guiding innovation.
Lesson 2: Don’t Believe the Hype
But prepare for it

Self-Inflicted Harms

Theoretical Attacks
No Proof of Concept...
For Now

Cool Stuff, but Niche Applications
Possible Concerns
The possibilities are endless

- New attack surfaces
- Lowering barriers to entry
- Proliferation
- Patching is not possible
- Loss of trust
Seeing Is Not Believing

“The post circulated widely enough on the platform to have a real-world impact. After Twitter user Deltaone, who has over 650,000 followers, shared the post at 10:06 am, the stock market fell 0.26 percent a mere four minutes later” - Yahoo! Money, May 22, 2023

Analysis:

Close inspection of the photo reveals the fence around the building appears to melt into the sidewalk and the windows are not aligned.
"JUST TWO DAYS before Slovakia’s elections, an audio recording was posted to Facebook. On it were two voices: allegedly, Michal Šimečka, who leads the liberal Progressive Slovakia party, and Monika Tó dová from the daily newspaper Denník N. They appeared to be discussing how to rig the election, partly by buying votes from the country’s marginalized Roma minority." - Wired, Oct 3, 2023

Analysis:
AI speech classifier made by an American company called Eleven Labs confirmed audio as generated.
“In testing ChatGPT, Ozarslan instructed the bot to write a phishing email, and it spat out a perfect mail within seconds. "Misspellings and poor grammar are often tell-tale signs of phishing, especially when attackers are targeting people from another region. Conversational AI eliminates these mistakes, making it quicker to scale and harder to spot them," he says." - Bank Info Security, Dec. 15, 2022

Analysis:

Use of current techniques to determine source of emails and links before interactions.
What Can You Trust?
What Do You Do?

The End of Privacy is a Taylor Swift Fan TikTok Account Armed with Facial Recognition Tech
Lesson 3: Start With What You’ve Got

Start small, dream big

Relationships  Processes  Communications
Lesson 3: Start With What You’ve Got

Start small, dream big

- Roles and Responsibilities
- Red Teaming and Penetration Testing
- Secure Development and Incident Response Lifecycles
- Crisis Communications
- Release Management
Secure Development Lifecycle

Secure infra, supply chain, design, implementation, and release

- Toolchain/Infra (SCM, Build CI/CD, Scanners, Data Management)
  - Setup Once | Verify Periodically

- Supply Chain (Dependency Managers, Central Repos - Approved OSS, 3rd Party)
  - Setup Once | Verify Periodically

- Plan and Design (Objectives, Requirements, Architecture, Threat Modeling)
  - Document | Iterate as Required

- Implement (Coding, Model-Training, Verification, Scanning, Pen-testing)
  - Execute at Each Release

- Release/Deploy (Signing, Model Cards, Reviews, Exception/Risk Tracking)
  - Execute at Each Release

- Monitor and Support (PSIRT, PSOC, Etc)
  - Continuous with Release
Incident Response Lifecycle

INTAKE and TRIAGE
Receives notification and routes to correct team for triage.

Analysis
Validates the vulnerability and assesses risk level.

Remediate
Determines the vulnerability mitigation.

Disclosure
Follows the Communication Plan to notify customers.

Feedback
Incorporates feedback and root cause actions.
Insecure Cloud, Container or Orchestration Configuration
Injection Flaws
Excessive Data Exposure
CI/CD Pipeline & Software Supply Chain Flaws
Improper Authentication & Authorization
Insecure Secrets Storage
Over-Permissive or Insecure Network Policies
Using Components with Known Vulnerabilities
Improper Assets Management
Inadequate ‘Compute’ Resource Quota Limits
Ineffective Logging & Monitoring

Prompt Injections
Insecure Output Handling
Training Data Poisoning
Supply Chain Vulnerabilities
Sensitive Information Disclosure
Insecure Plugin Design
Excessive Agency
Overreliance
Model Theft
Model Denial of Service
Lesson 4: Enhance and Iterate
Tools and Processes

- AI SDL
- Harm vs Vulnerability
- Playbooks
- Threat Intelligence
- Tabletop Exercises
Lesson 4: Enhance and Iterate

Human Factors

- Train Engineers
- Engagement Processes
- Build Relationships
Lesson 5: Industry Preparedness

Think Local Act Global

- Follow Coordinated Vulnerability Disclosure
- MITRE - CVEs for AI
- FIRST.org AI special interest group
- Bug Bounty Community of Interest
- OWASP Top 10 for LLM
Summary

• Prepare for the Unknown
• Don’t Believe the Hype Cycle…but Prepare for It
• Start with What You’ve Got
• Enhance and Iterate
• Industry Preparedness
Thank you