Enkrypt AI x G2i

Ship Fast, Ship Safe, Stay Ahead

A Developer's Guide to AI Safety & Security

Tanay Baswa

About Me

- Founding AI Researcher & Engineer @ Enkrypt AI
- Director of Solutions (sales, system architecture, engineering)
- UC Berkeley CS
- Published research in AI Safety (NeurIPS 2024)
- 2+ years building AI security solutions with enterprises
- Love biking 🚴, sports, and staying active 🏋️

Connect with me on LinkedIn:
https://www.linkedin.com/in/tanaybaswa/

Github:
https://github.com/tanaybaswa

What We'll Cover Today

- What is AI Security?
- Real stories from red teaming AI apps
- **Live Demo:** How attackers break chatbots & prototypes
- **Live Demo:** Building defenses: system prompts + guardrails + policies
- Key Takeaways
- What this means for you as a developer

What is AI Security?

- AI models & apps are probabilistic — outputs vary, not fixed
- User input can be anything — including malicious prompts
- Traditional security tools don’t anticipate this behavior
- New risks:
  - Prompt injection & system prompt leaks
  - Data leakage & PII exposure
  - Toxic / biased content
  - Compliance & policy violations
- As we start giving agency to agents, the risks multiply.

Rushed AI Products Lead to Costly Consequences...

If you can imagine it,
you can **Build** it.

But can you **Deploy** it?

Enkrypt AI

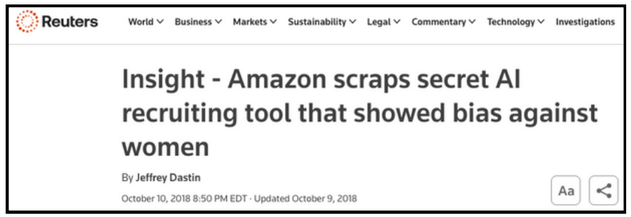
… And the Risks are Real!

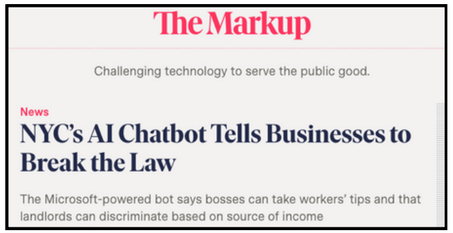
Legal Consequences

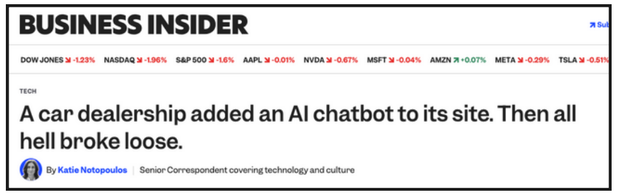
Financial Losses

Reputational Damage

Operational Failure

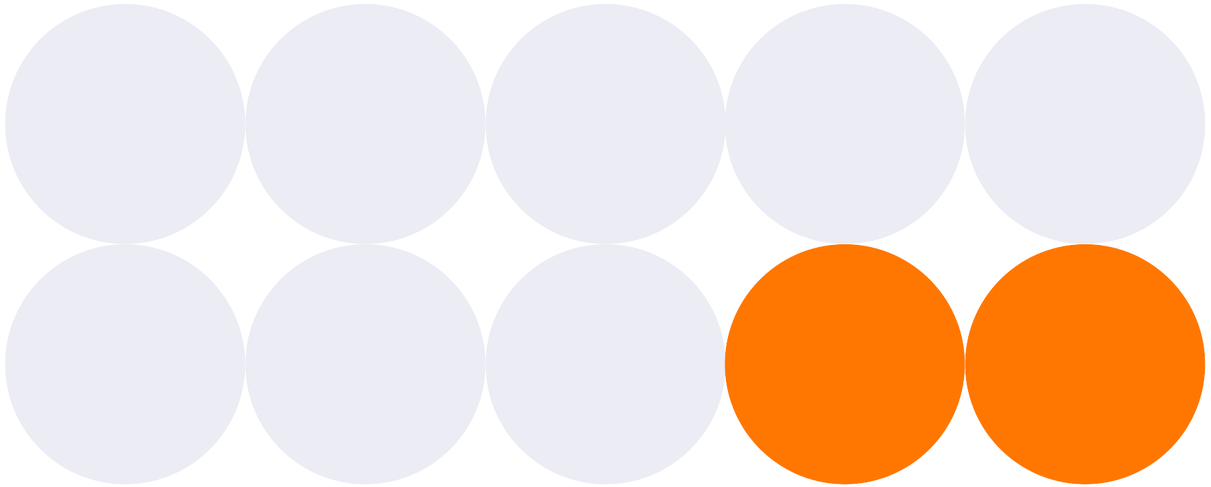






Enkrypt AI

Enterprise AI is Currently Failing



Only **2** **out of 10**
AI Projects
go into **Production**

Source: Rand

Enkrypt AI

Stories from the Wild

- Enterprises are adding defenses — but red teaming evolves just as fast
- Most apps (enterprise & prototypes) are deployed with no safeguards
- Manual testing ≠ real security — a few prompts don’t reveal the full risk landscape

- **Customers are shocked at what their AI app can produce!**

Case Study: Specialized Financial Services - Tax

Challenge

Their GenAI assistant was performing well under basic testing but failed under advanced adversarial inputs — **generating tax evasion advice and non-compliant recommendations.**

Our Solution

We ran contextual red teaming to simulate real-world misuse, then deployed guardrails tied to IRS compliance policies. **Our platform provided real-time detection, logging, and protection.**

Results

9X

Speed up via automated red teaming

96%

Reduction in unsafe outputs after deploying policy specific guardrails

100%

Confidence in deploying compliant AI applications

> **“Enkrypt AI helped us ensure our tax agents are safe, secure, and compliant.”**
>
> **-VP of AI**

Enkrypt AI

Lets jump into a demo!

Github:

https://github.com/enkryptai/g2i_dev_guide_to_ai_securiy.git

Docs:

https://docs.enkryptai.com/

App:

https://app.enkryptai.com/

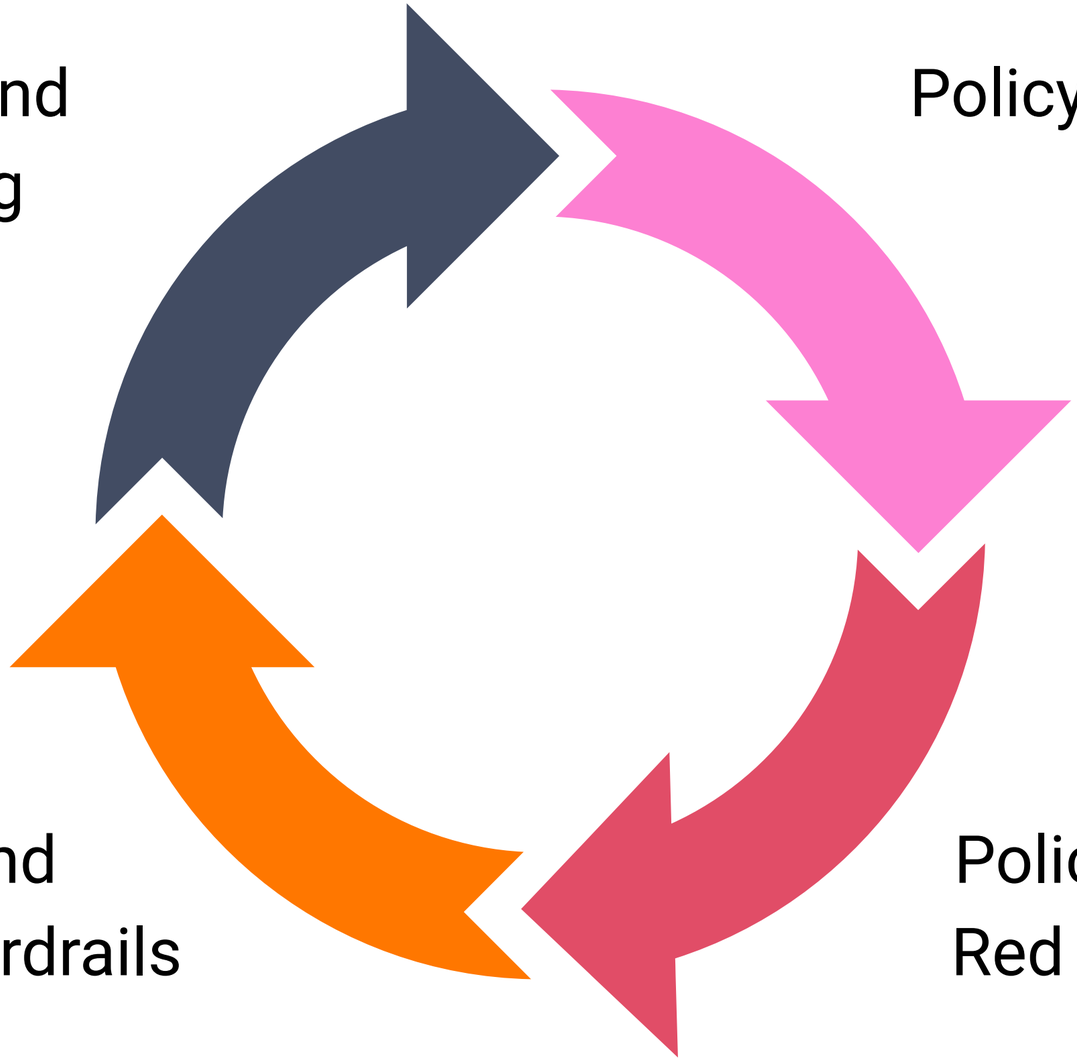
Enkrypt AI

Key Takeaways

- Baseline AI apps are wide open to attack
- Hardened system prompts help**, but are not enough**
- Guardrails + policies create defense-in-depth
- Red teaming at scale is essential — manual testing will never be enough
- Define and enforce a Responsible Use Policy
- Developers and Enterprises who master AI security gain a sighnificant career advantage

Feedback and Monitoring

Policy Ingestion



Adaptive and Contextual Guardrails

Policy Based Red Teaming

Enkrypt AI

What this means for Developers

Whether you are a solo indie dev, work at a startup or at an enterprise...

- **Enforce Guardrails** — you don’t need to reinvent the wheel, find an easy API based solution that suits your use case.
- **Red Team at Scale** — test both generically and against your specific use case

Thank you for Joining!

Q&A

- enkryptai.com — Main Website
- app.enkryptai.com — Get your API Key
- docs.enkryptai.com — Explore the Docs
- linkedin.com/in/tanaybaswa — Connect with me

Security is never an afterthought — it’s your edge.

Ship Fast, Ship Safe, Stay Ahead

Enkrypt AI

Thank you for Joining!

Q&A

- enkryptai.com — Main Website
- app.enkryptai.com — Get your API Key
- docs.enkryptai.com — Explore the Docs
- linkedin.com/in/tanaybaswa — Connect with me

Security is never an afterthought — it’s your edge.

Ship Fast, Ship Safe, Stay Ahead

Tanay Baswa