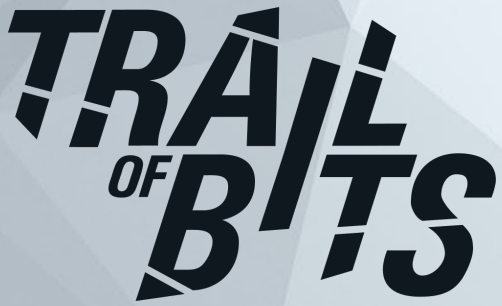

TRAIL
OF
BITS

About Us

Since 2012, Trail of Bits has helped secure some of the world's most targeted organizations and devices. We combine high-end security research with a real-world attacker mentality to reduce risk and fortify software.

Fast Facts

EXPERTISE Application Security | Cryptography
ML/AI Assurance | Research | Engineering

FOUNDED 2012 EMPLOYEES 120 LOCATIONS 14 PROJECTS 500+ LANGUAGES / CERTIFICATIONS 20+

ML Assurance Practice

Mission: Identify and taxonomize classes of failure modes which directly impact AI/ML model performance and novel hazards that threaten the AI/ML operations pipeline for mission-critical applications.

Works closely with Research practice to make novel advances in techniques and tools for AI/ML assurance.

Heidy Khlaaf - Engineering Director Michael Brown - Principal Researcher

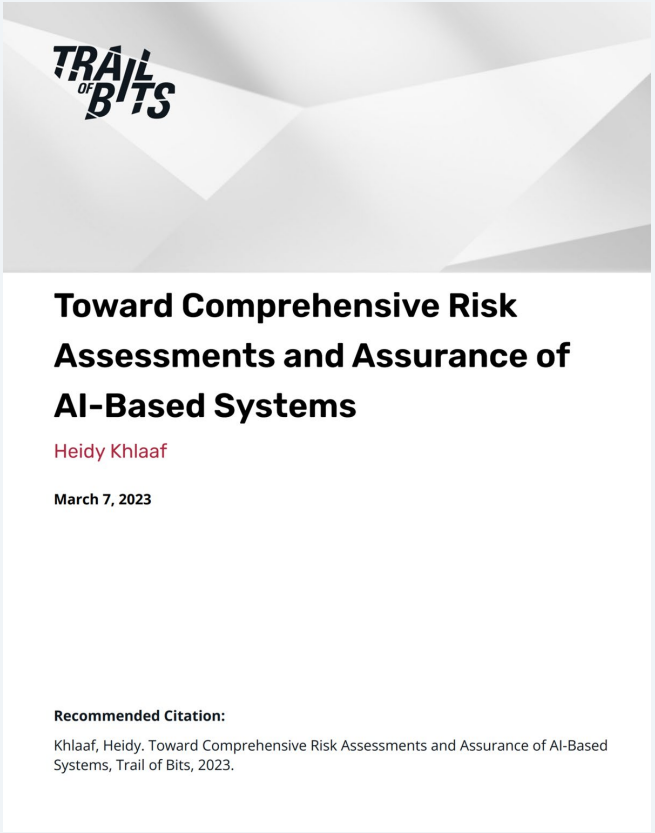
Kelly Kaoudis - Senior Research Engineer

AI/ML - A New Assurance Frontier

- Lack of *model robustness* that break safety and security properties (e.g., adversarial attacks, prompt "injections")
  - unpredictable and non-deterministic
  - difficult to measure and assure model performance
- Novel *structural vulnerabilities* and *supply chain intrusions* due to the use of AI in downstream dependencies
  - Poisoning web-scale training datasets
  - Sleeper agents: behaves like a normal model under most circumstances, but activates and generate commands when a specific code phrase is used
- New attack surfaces: *AI/ML ops/pipeline vulnerabilities* and exploits
  - degradation of model performance
  - exploitation of the collection and processing of data and parameters
- Proliferation and *misuse of AI/ML model capabilities* (e.g., offensive cyber)

Assessing Model Robustness

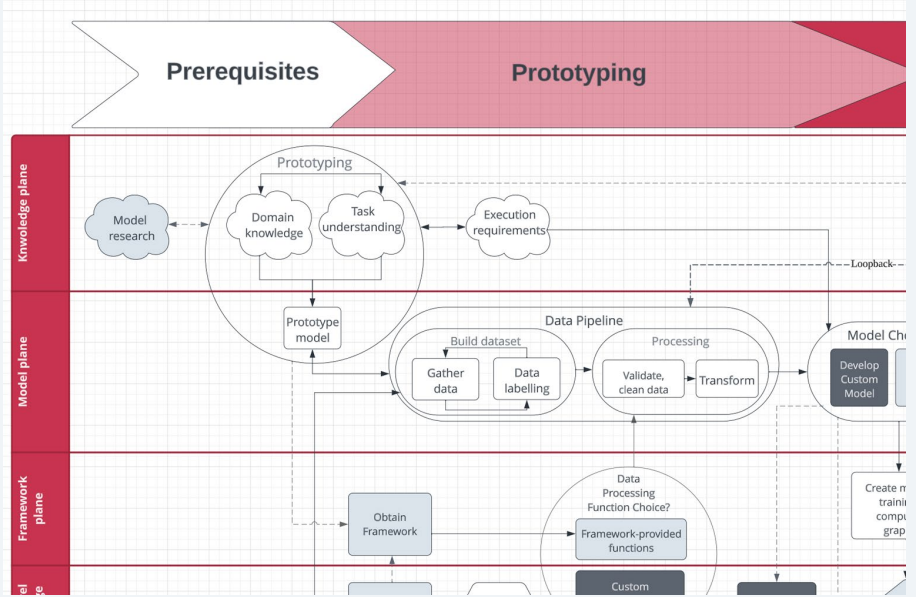
- Unique deployment risks and failure modes
  - Must assume output can be manipulated by attackers
- Designed an AI risk framework using Operational Design Domains (ODD) to assess AI-based systems
- ODDs describe specific operating conditions for which an AI-system is designed to properly behave
  - System hazards and mitigations determined against this safety and security envelope



Supply Chain Research and Assessment

- Currently supporting the UK Government's AI Taskforce

- Assessing and taxonomizing new, undetectable threats from downstream systems created by or using LLMs that may lead to the subversion of existing supply chain integrity

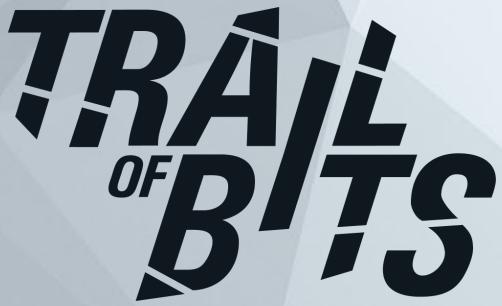


Novel ML Attack Surfaces

- AI/ML frameworks and tooling have unknown and uncharted attack surfaces
  - Model and assets can be compromised or degraded
- AI/ML systems development cycles forego established security practices in favor of rapid innovation
  - New file and serialization formats for model weights have resulted in new vulnerabilities
- Built Fickling: A pickle file analysis tool for identifying malicious files
- Conducting safety and security audits for commercial AI/ML systems

Proliferation of Cyber Capabilities

- What are the implications of LLMs being used (or misused) by adversaries?
  - Can LLMs make adversaries more capable? Give them access to speed and scale?
  - Already evident for social engineering - malicious actors can quickly and easily make high fidelity phishing emails and fake images at volume.
  - Potential to use summarization and contextualization features to lower barrier of entry for low-level attackers
- Supporting the UK Government's AI Taskforce with an national security risk assessment on the proliferation of offensive AI cyber capabilities via LLMs
  - Created a framework to rigorously evaluate emergent offensive cyber capabilities in LLMs
  - Conducting a preliminary evaluation of foundation models, their risks, and findings


TRAIL
OF
BITS