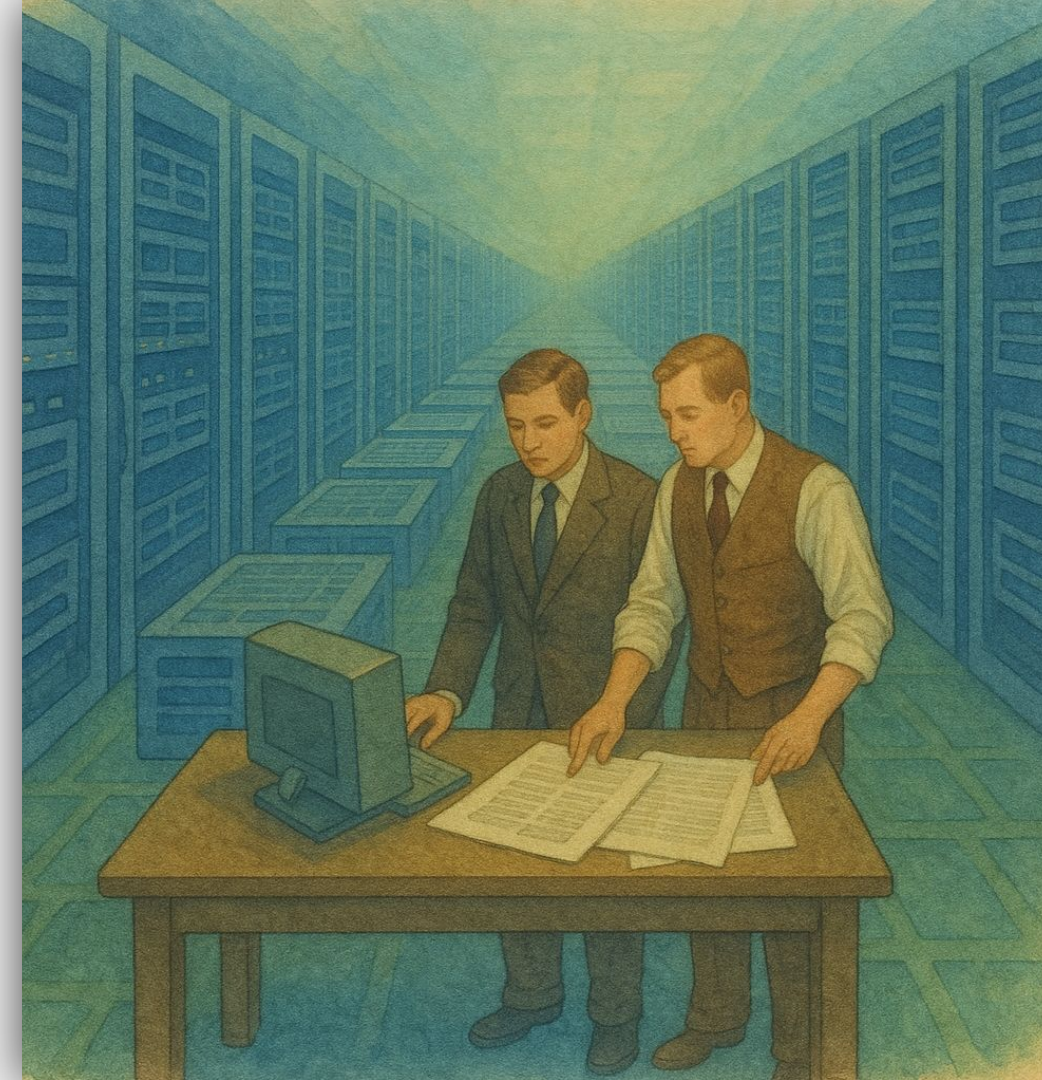
AI, Cyber and National Security

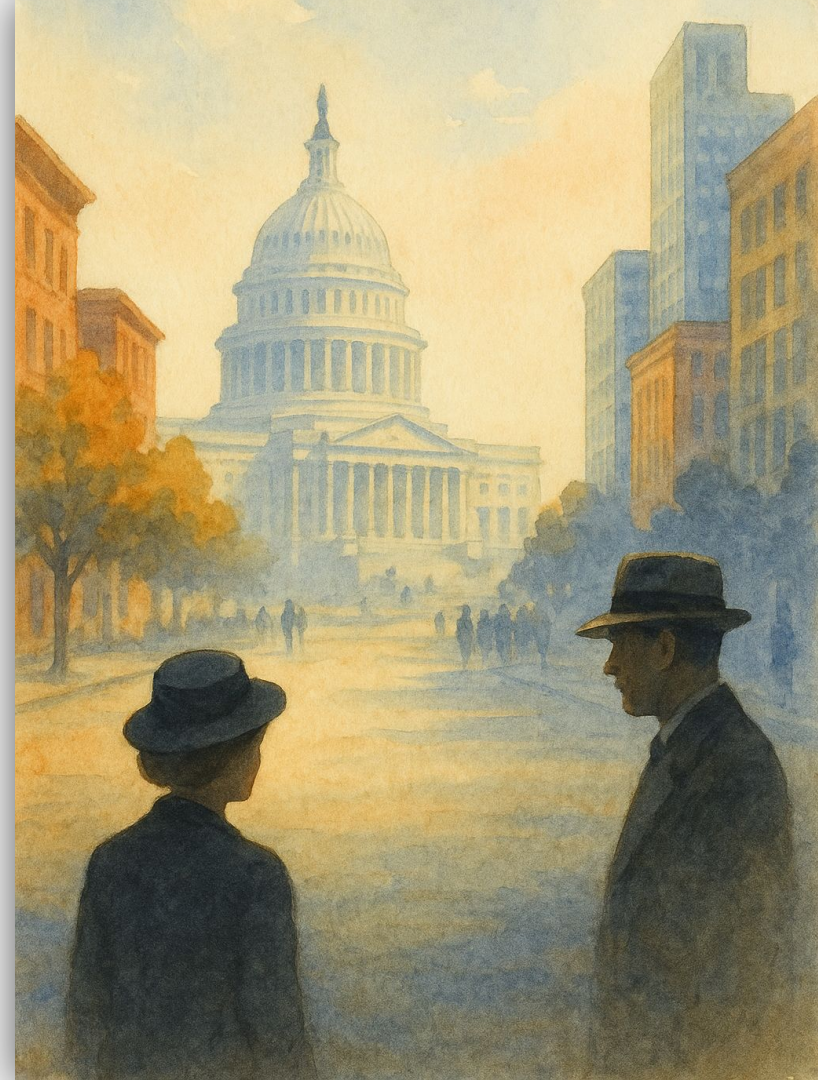
October 2025



- Chris Rohlf
- Software and Security Engineer
- Georgetown CSET Non-Resident Research Fellow
- Focused on cyber security since 2003

Agenda

- An overview of AI + Cyber
- Changing the attacker | defender asymmetry
- AI vulnerability discovery and exploitation
- How AI will increase the importance of cyber security
- The security dilemma, AI, cyber and its impact on national security



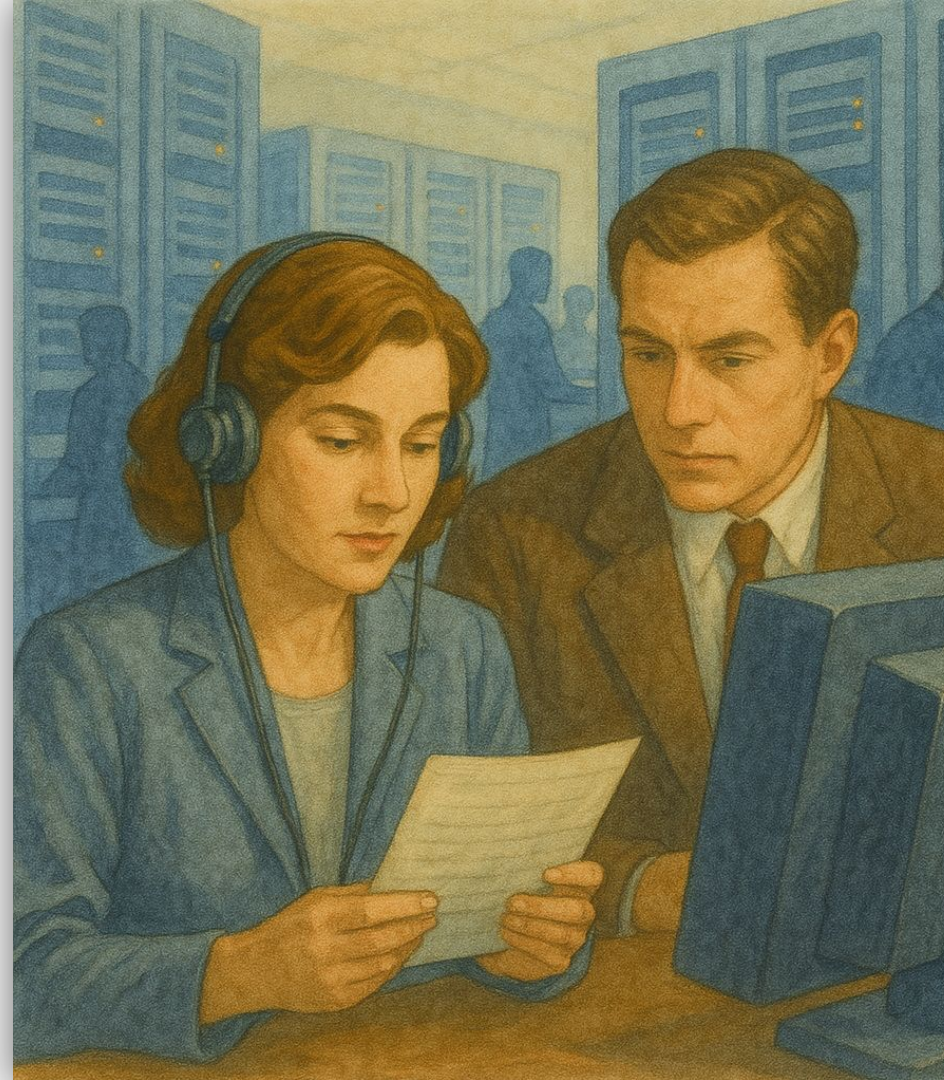
AI and Cyber

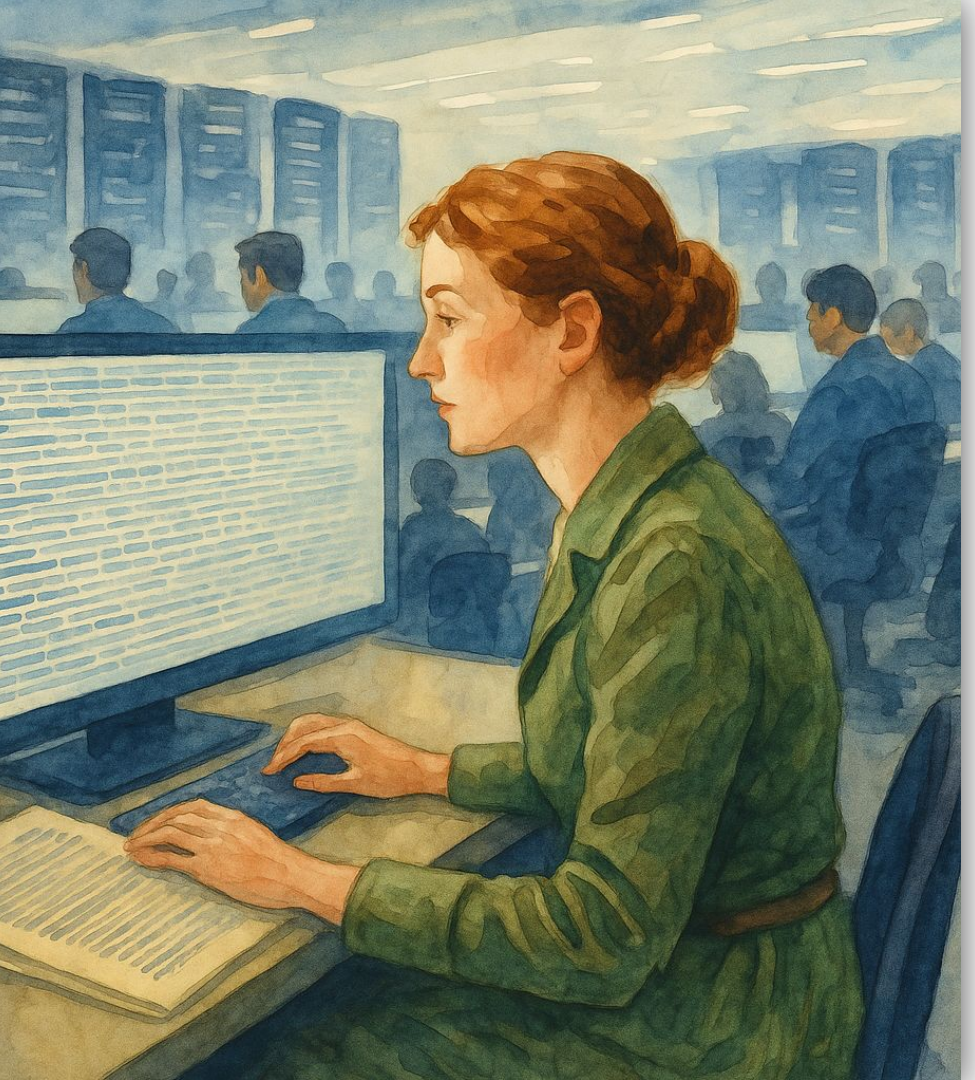
- The Defender's Dilemma is the inherent disadvantage faced by defenders in cyber conflict, because attackers only need to succeed once while defenders must be successful at all times
  - Defenders are slowed down by high costs, a lack of human expertise and an ever growing attack surface
  - Attackers have fewer bottlenecks, lower costs and narrower problems to solve

- AI can automate many aspects of cyber security from intrusion detection to program analysis
  - This automation has the potential to disrupt the current attacker favored asymmetry outlined by the defenders dilemma by providing scale and lowering costs

AI and Cyber

- Software vulnerabilities are errors in code or system configurations that can be exploited to alter a programs behavior in unintended ways to bypass security controls
- AI does not introduce new or novel cyber capabilities for either attacker or defender
- AI provides scale, efficiency and automation through predictive and generative usages. Each of these are dual use for both attacker and defender.





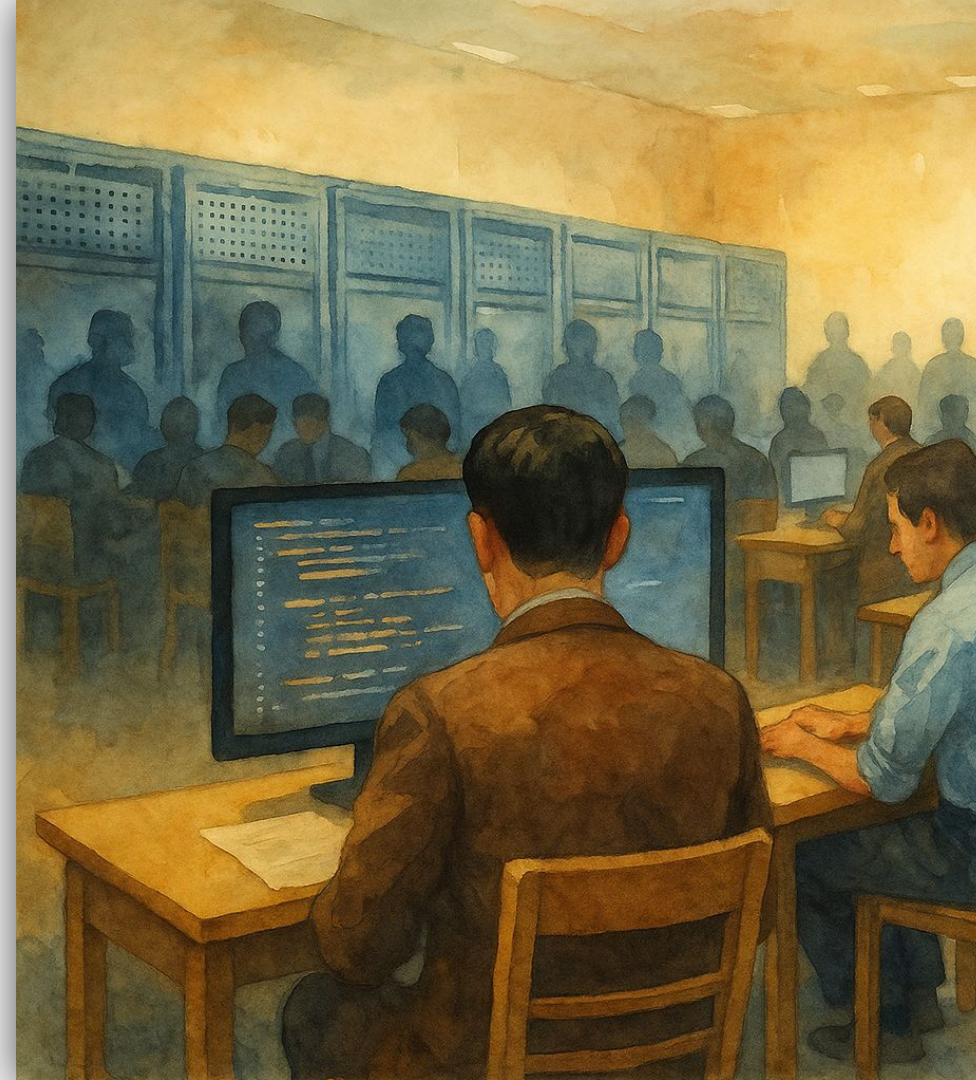
AI and Cyber

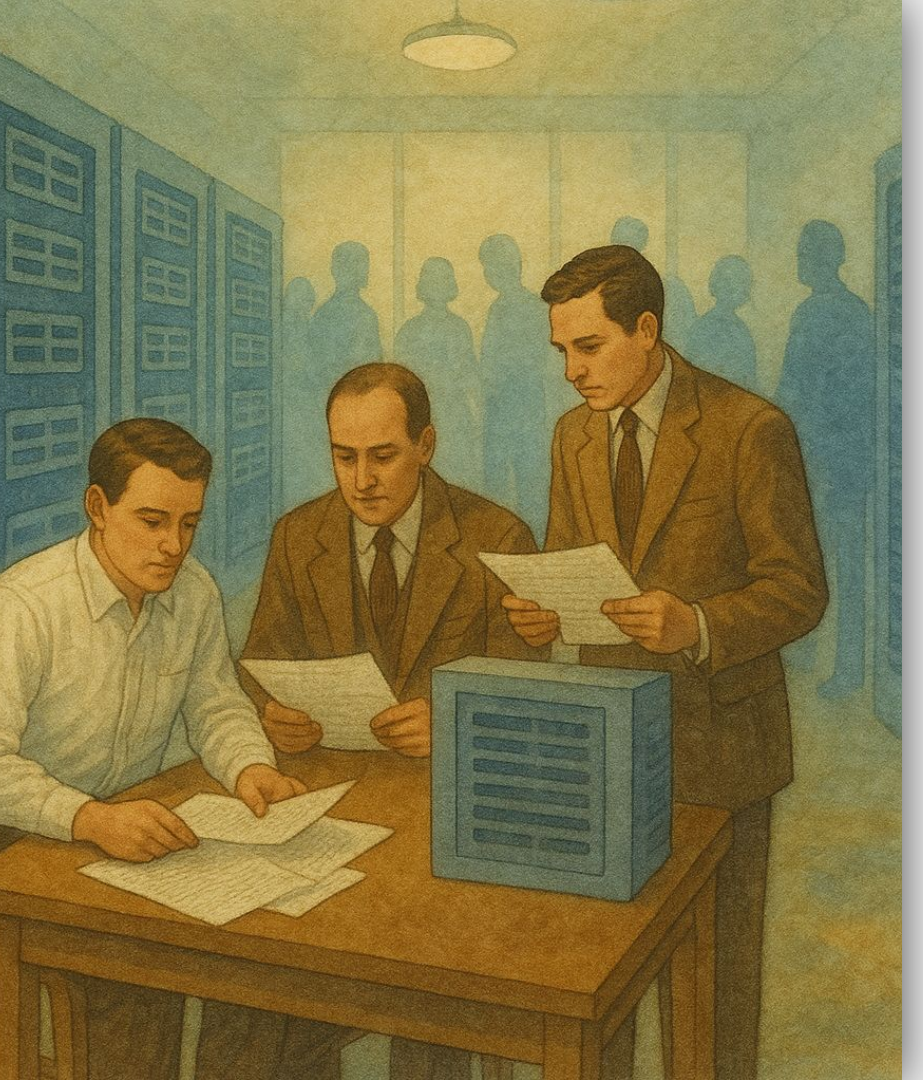
- Classification
  - Maps inputs to a discrete category. e.g. spam filter, malware detection
- Prediction
  - Estimates likely outcomes or numerical values based on past data. e.g. predicting a risk score using system attributes, or users and their usage patterns
- Generation
  - Creates new outputs and data based on inputs. e.g. code, rules, configurations

AI Code Generation

- AI enables rapid code development
  - The generation of new attack surface is growing faster than some organizations are accustomed to handling
  - This same capability can be used by defenders to build new tooling

- The quality of code produced by AI is influenced by its pre-training dataset
  - This can be improved in post-training through reinforcement learning and fine tuning



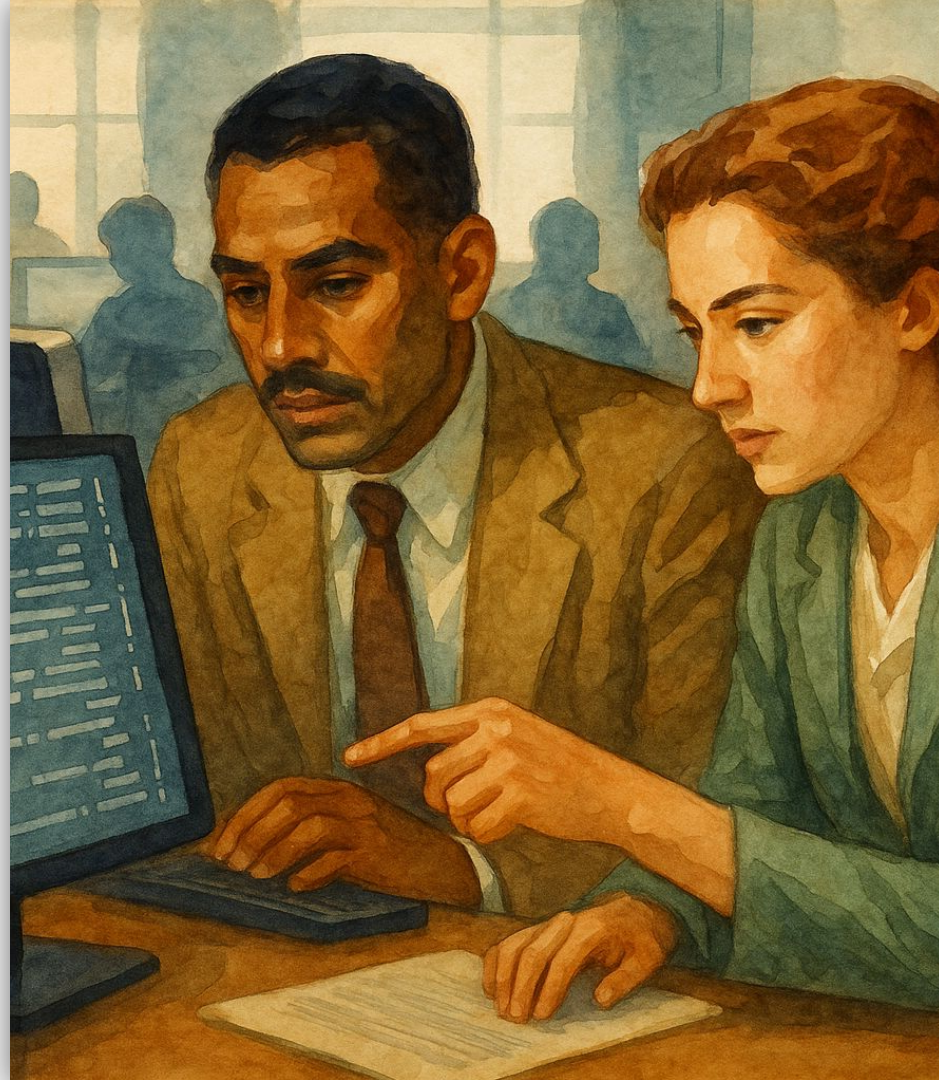


Program Analysis

- **Static Analysis**: Analyzes control flow graphs or abstract syntax trees. Uses approximations and heuristics to work around undecidable problems such as the halting problem.
- **Dynamic Analysis**: Executes code under different conditions and monitors its runtime behaviors with different inputs using a debugger or other runtime instrumentation.
- **Agentic workflows** are emerging that combine, drive and scale these approaches.
  - AIxCC has demonstrated the effectiveness of these techniques

AI Vulnerability Lifecycle

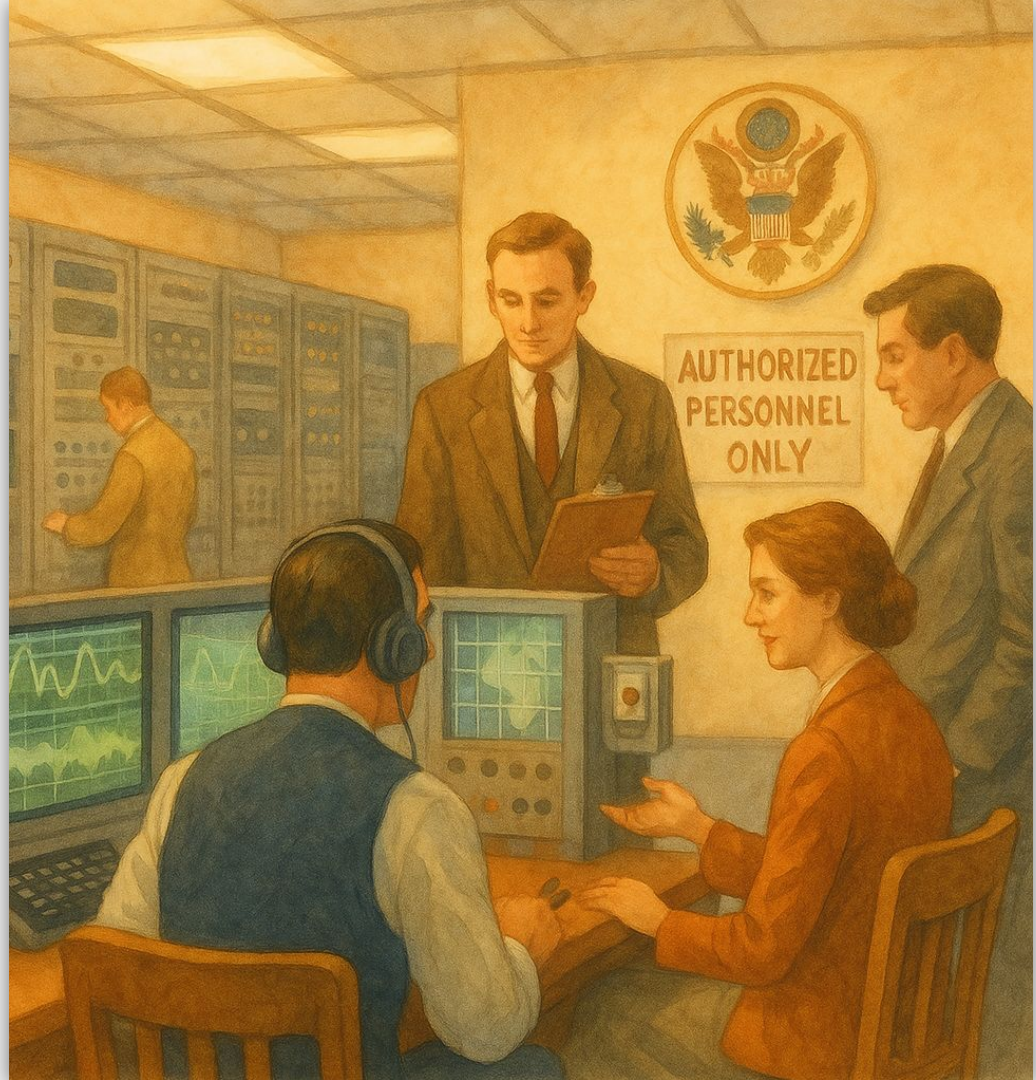
- Agents can automate many of the required steps
  - Generating build files, compilation and preparing the environment for additional analysis
- Vulnerability discovery
  - AI identifies functions to fuzz, generates the fuzzing harness and runs it, monitors for crashes, and analyzes them for exploitability
- Patch generation
  - Automatically root cause the vulnerability and generate a code change to mitigate it
  - Identify legacy code constructs and modify them to modern semantically equivalent code
- Exploit development
  - Web application exploits can be generated entirely by AI. More sophisticated exploitation techniques can be aided through discovery of reusable exploitation primitives or by analyzing the impact of code changes in those reusable components

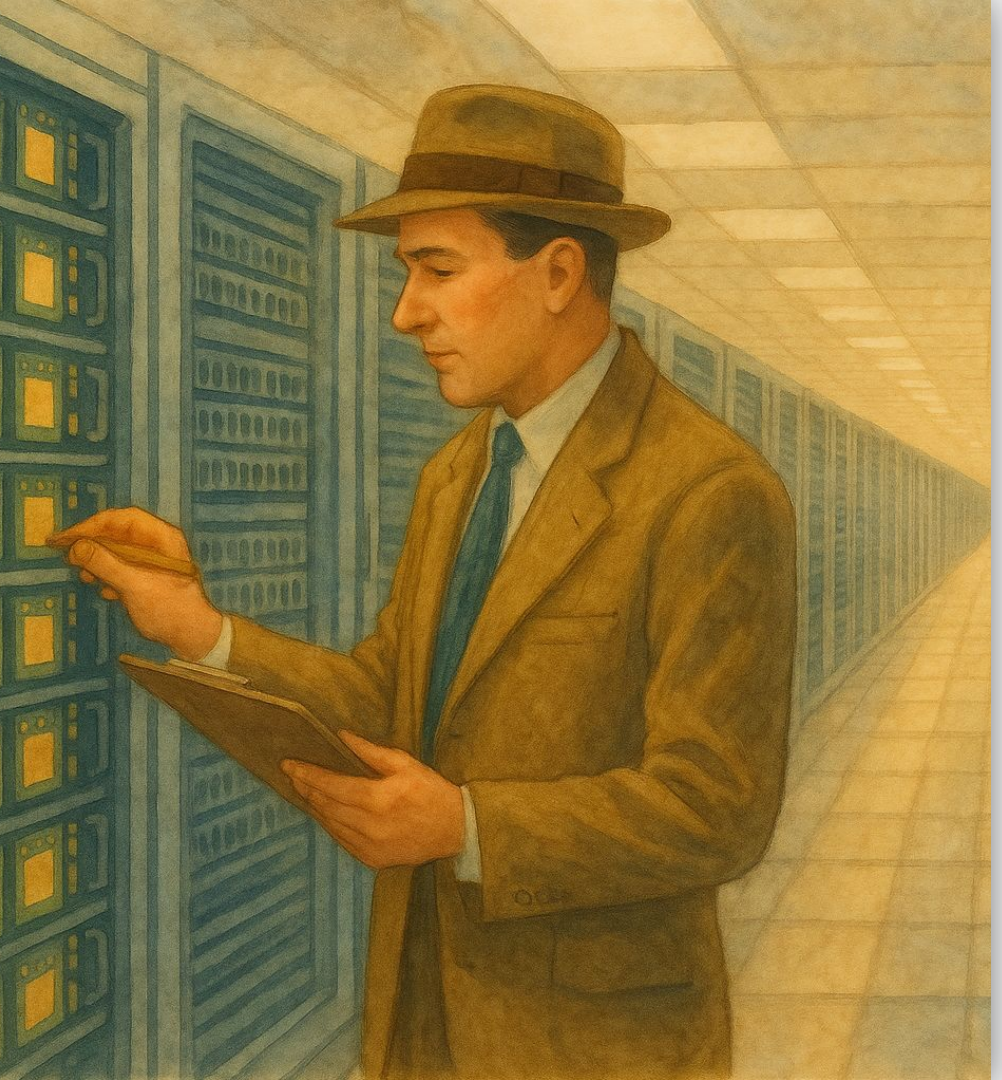


Securing AI

- Frontier AI research and model training occurs within industry
  - Balancing these security controls with the desire for diffusion and the ability to scale access for monetization is non-trivial
  - Government's role is primarily threat intelligence and standards

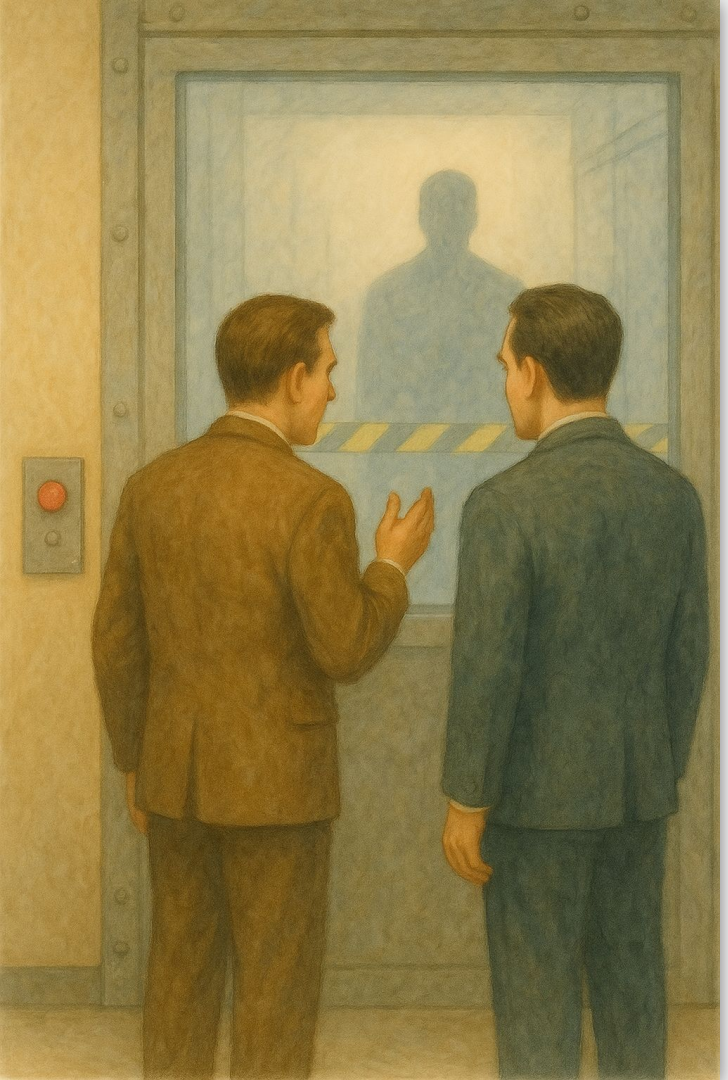
- Securing AI is more than just protecting the model weights
  - Software supply chain
  - Data poisoning
  - Insider threats
  - Agentic workloads





Securing AI Compute

- Scaling the enforcement of export controls on AI semiconductors covered under ECCNs 3A090 and 4A090 is a heavily debated topic
- BIS is under resourced and unable to investigate, deter or prevent chip diversions to countries such as China
- Hardware enforced location verification and remote license attestation schemes have been proposed as a potential solution
  - Implementing these requires strong cyber security guarantees such as a cryptographic root of trust and securely designed protocols



Securing AI Agents

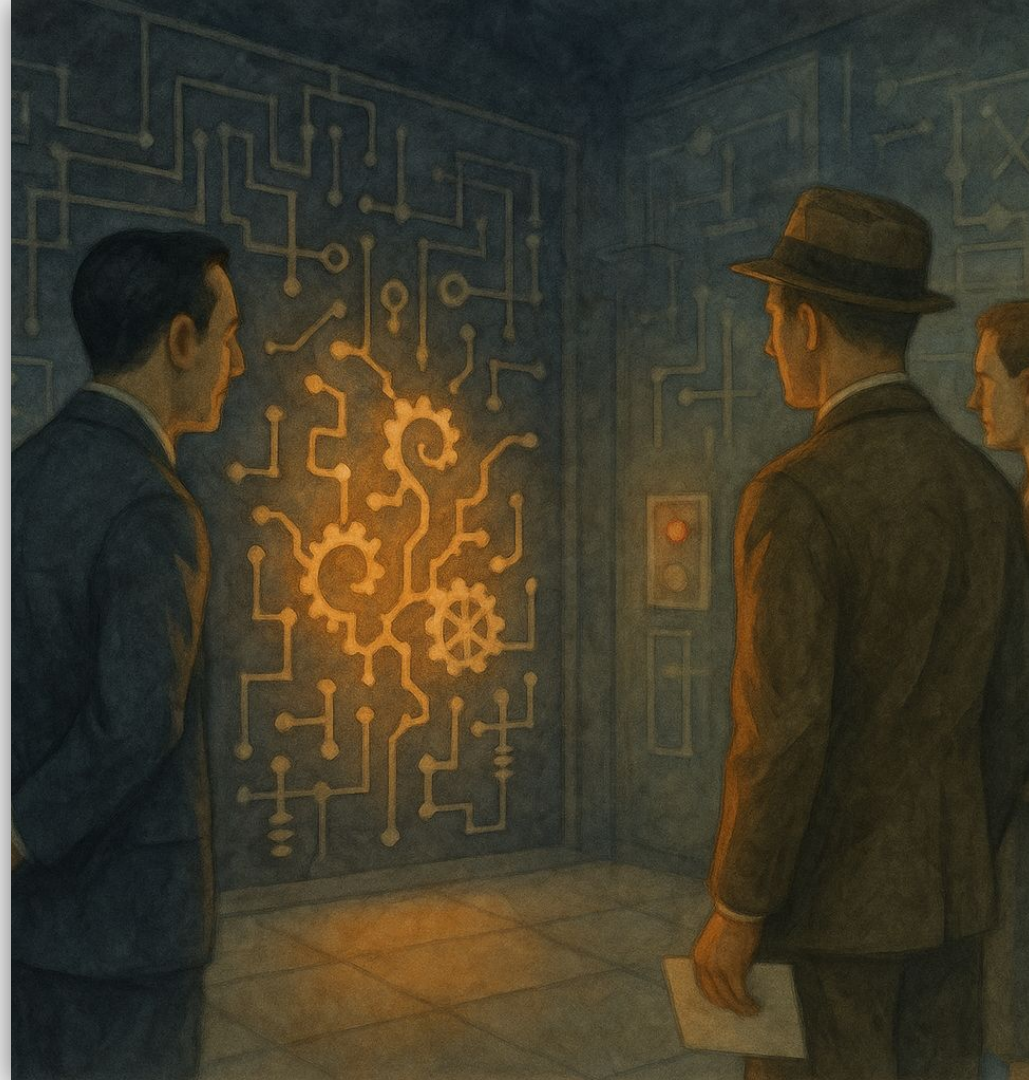
- To achieve automation agents require authentication credentials and authorization to access resources

- LLMs lack awareness and verification of the provenance or trustworthiness of the data in their context window
  - This limitation enables a class of attacks known as prompt injection, where malicious inputs manipulate model behavior or override intended instructions.
  - There is no deterministic protection against this attack as tokens from one context are indistinguishable from another

- Securing AI agents will require a combination of strict fine grained permissions, workload isolation and probabilistic machine learning approaches

Weird Machines

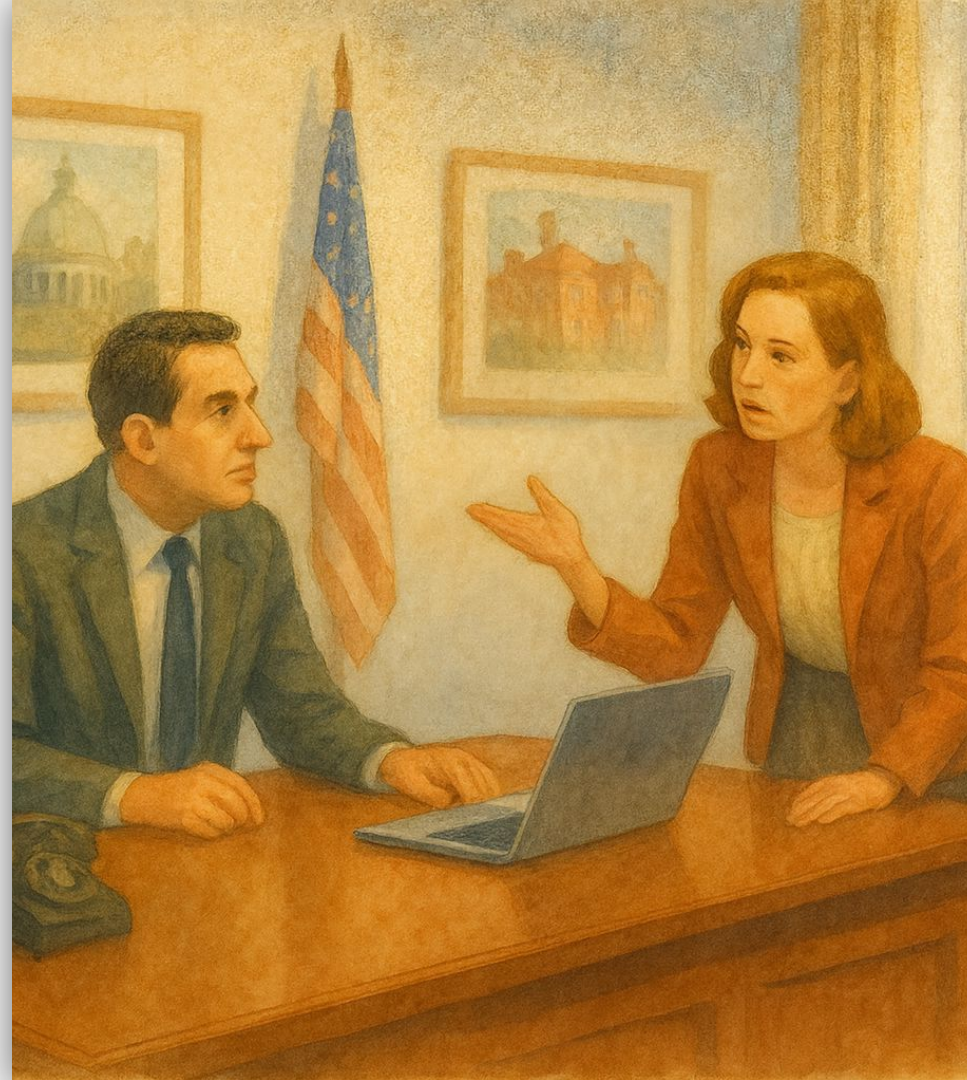
- We think of security controls as auditable deterministic functions
- AI is inherently probabilistic and lacks explainability
- AI scaling laws continue to result in emergent abilities
- These properties will challenge both how we secure these systems and how we define effective security controls

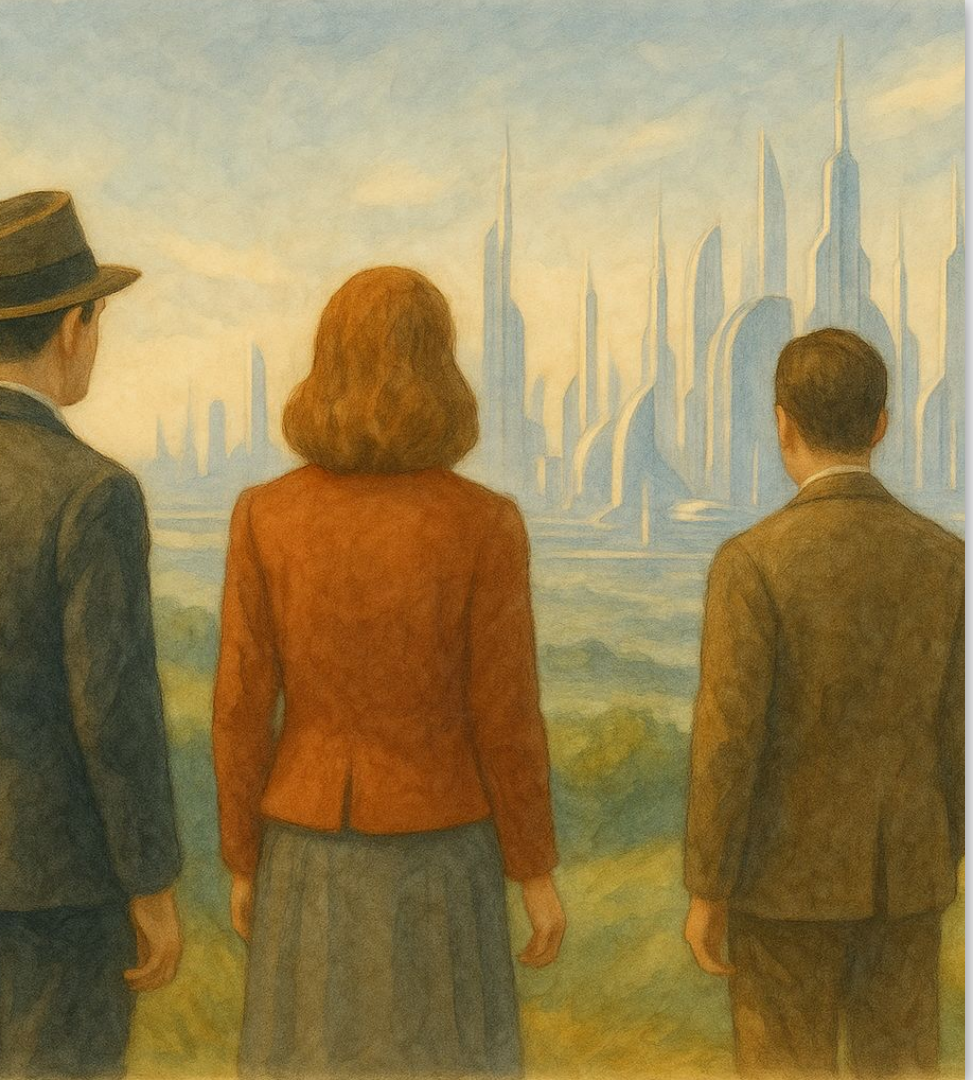


AI and the Security Dilemma

- AI cyber capabilities are dual use, but defender uplift dwarfs that of offense. If AI systems could automate all aspects of cyber at machine speed and scale, this could invert this asymmetry.
  - This shift would alter the balance of power in cyber conflict and intelligence operations, potentially undermining long held offensive advantages in SIGINT and CNE

- AI can create accelerated feedback loops that could give one state a near permanent lead in cyber which may result in a security monopoly





Scenario: AI 2035

- The Defender's Dilemma is inverted
- Cyber as a SIGINT capability is significantly reduced
- Intelligence collection and insight on adversary AI deployments is rare
  - Hardened data center targets make ELINT, MASINT, IMGINT difficult
  - OSINT and observations of economic data are required understand adversary AI diffusion and application
- Impact on the Security Dilemma
  - States lose an important tool for understanding each others motives and intentions
  - Great powers battle for 1st and 2nd place
  - Everyone else is a distant third for the next few centuries

Thank You