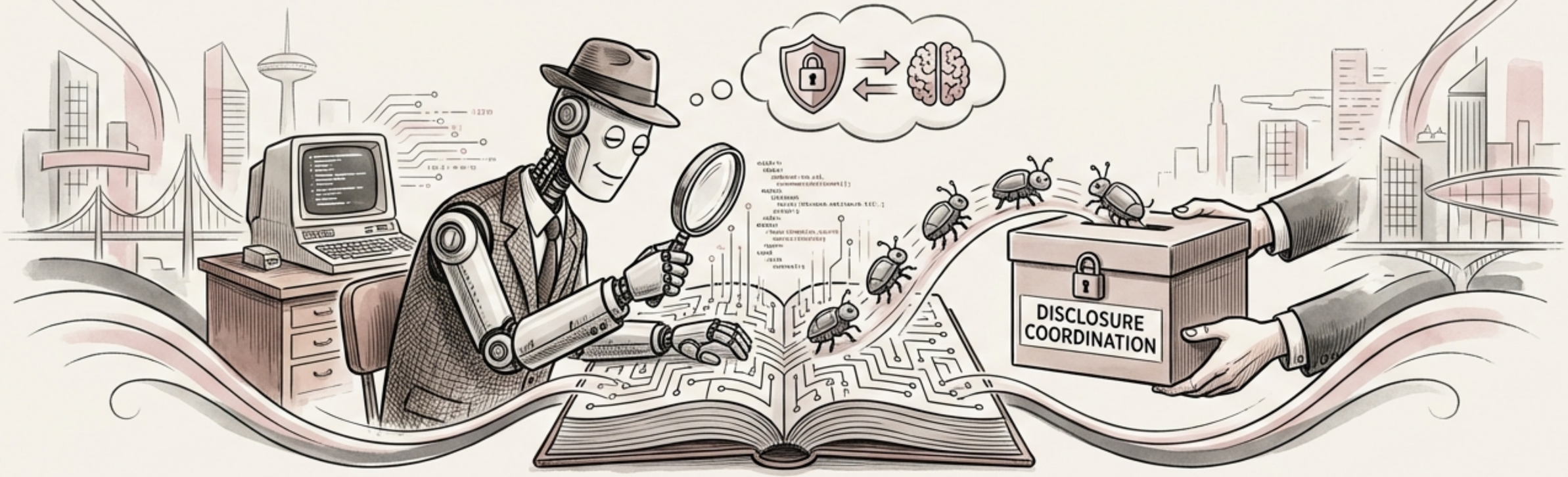
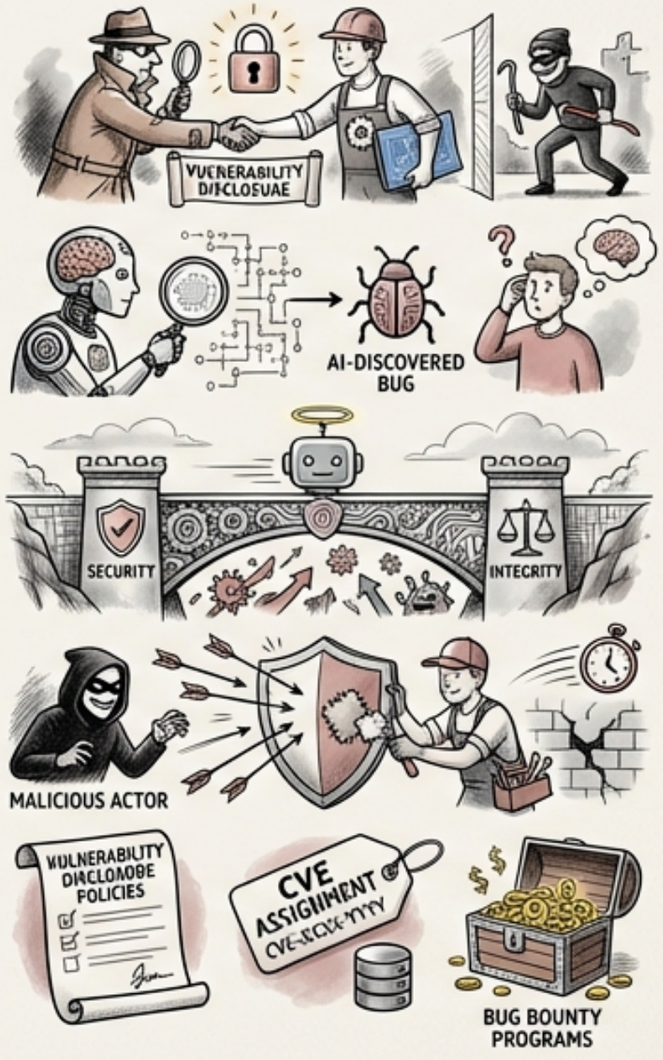
Securing AI-Augmented Development: Coordinated Vulnerability Disclosure



Securing AI-Augmented Development: Coordinated Vulnerability Disclosure

- Coordinated Vulnerability Disclosure (CVD) is a collaborative process between security researchers and software vendors to address vulnerabilities before public disclosure.
- AI-augmented development introduces new CVD challenges due to AI-discovered vulnerabilities and their unique characteristics.
- A robust CVD process is crucial for maintaining the security and integrity of AI-powered applications.
- Effective CVD helps reduce the risk of exploitation by malicious actors by allowing vendors to patch vulnerabilities before they are widely known.
- This presentation will cover the key components of a successful CVD program, including vulnerability disclosure policies, CVE assignment, and bug bounty programs.



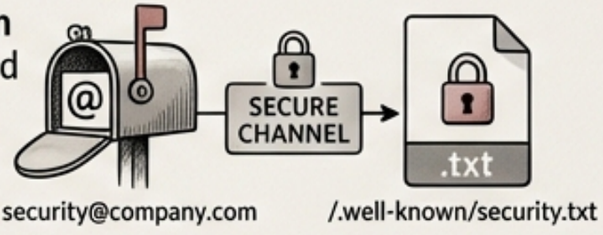
Vulnerability Disclosure Policy (VDP): Your Foundation for Responsible Disclosure

- Every organization that ships software needs a publicly accessible **Vulnerability Disclosure Policy (VDP).**

- **Safe harbor language** protects good-faith security researchers from legal repercussions for vulnerability discovery and reporting.

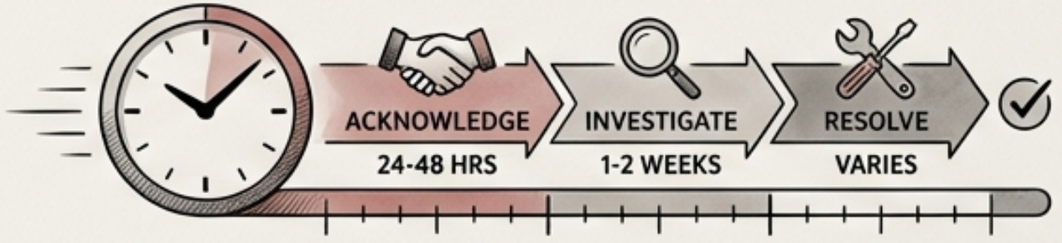
- The VDP should clearly define the **scope** of covered products and services.

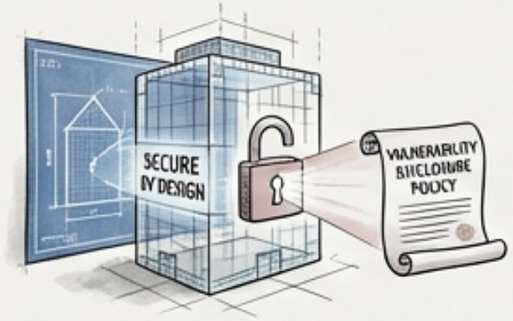
- Specify **clear communication channels**, such as a dedicated security email address (e.g., security@company.com) and a **security.txt** file.



- **Safe harbor language** protects good-faith security researchers from legal repercussions for vulnerability discovery and reporting.

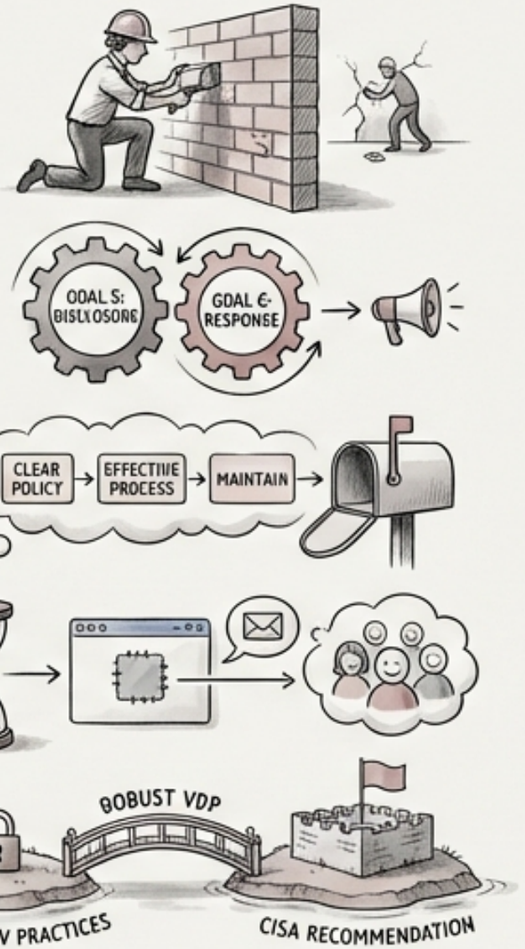
- Define expected **response timelines** for acknowledging, investigating, and resolving reported vulnerabilities.





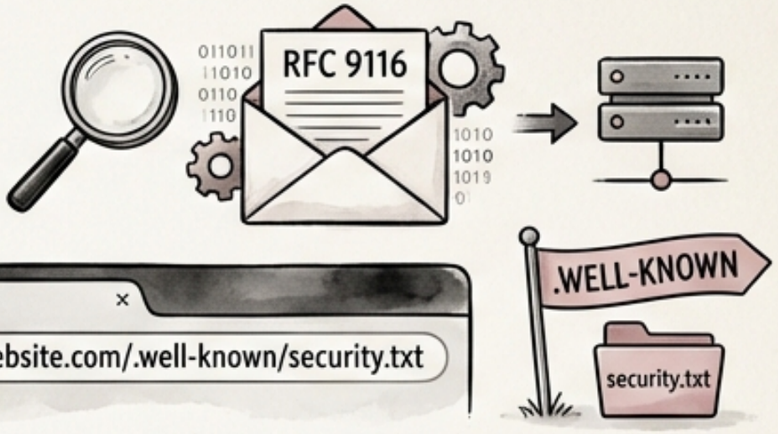
CISA's 'Secure by Design' Goals Emphasize Vulnerability Disclosure

- CISA's Secure by Design initiative promotes proactive security measures during software development.
- Goals 5 and 6 of CISA's Secure by Design directly address vulnerability disclosure and incident response.
- Goal 5 likely focuses on establishing and maintaining a clear and effective vulnerability disclosure policy.
- Goal 6 likely addresses timely patching and communication of vulnerabilities to users.
- Implementing a robust VDP aligns with CISA's recommendation to prioritize secure development practices.



STREAMLINING SECURITY COMMUNICATION WITH SECURITY.TXT

- RFC 9116 standardizes the `security.txt` file for machine-readable security contact information.
- `security.txt` should be placed at `/.well-known/security.txt` on your web server.
- The file can contain: contact email/URL, encryption key, preferred language, disclosure policy URL, and hiring URL.
- Implementation is trivial and signals security maturity to researchers.
- Without it, researchers may disclose vulnerabilities publicly or not report them at all.



CVE: The Foundation for Tracking and Addressing Vulnerabilities

- **CVE** (Common Vulnerabilities and Exposures) provides a unique identifier for each publicly known vulnerability.
- **CNA** (CVE Numbering Authority) are authorized organizations that assign CVE IDs.

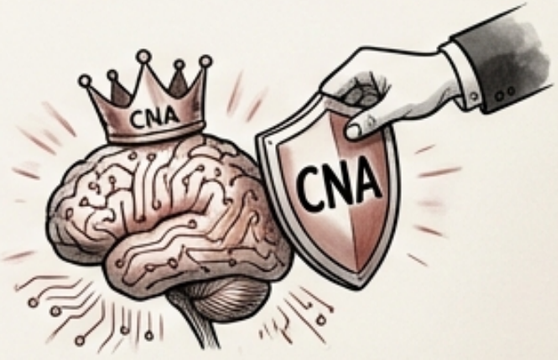
- The process involves: vulnerability discovery → request CVE from CNA → coordinate fix → publish advisory with CVE → update NVD.

Vulnerability Discovery → REQUEST → Request CVE from CNA → Coordinate Fix → Publish Advisory with CVE

- Request a CVE early in the disclosure process; don't wait for the fix to be completed.

Update NVD

- Organizations handling AI tool vulnerabilities should consider becoming CNAs for their products.

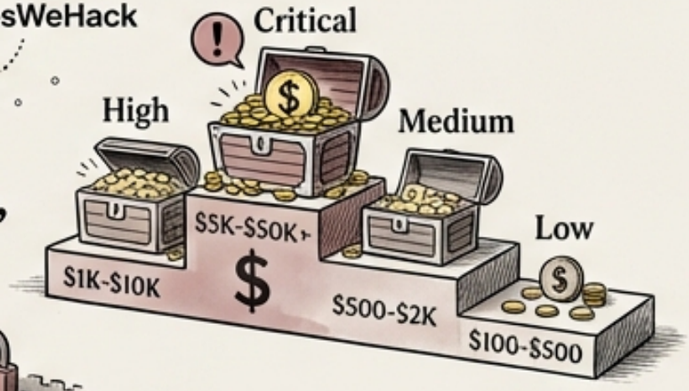


Bug Bounty Programs: Incentivizing **Vulnerability** Discovery

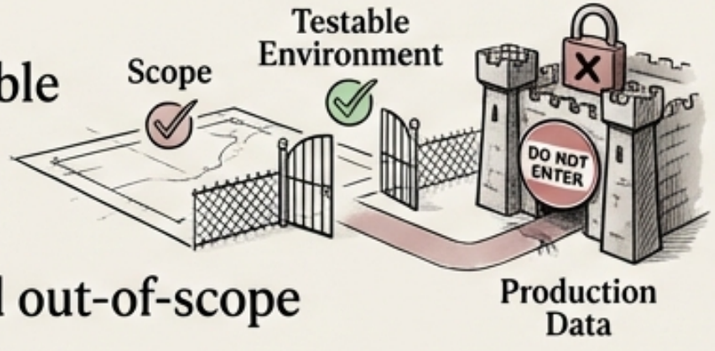
- Bug bounty programs incentivize security researchers to find and report vulnerabilities in exchange for rewards

- Popular platforms include: HackerOne, Bugcrowd, Intigriti, YesWeHack

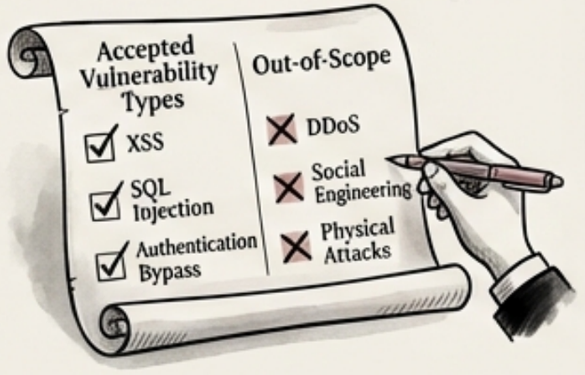
- Reward tiers are based on severity and impact (critical: $5K-$50K+, high: $1K-$10K, medium: $500-$2K, low: $100-$500)



- Clearly define the scope of what is testable and exclude access to production data

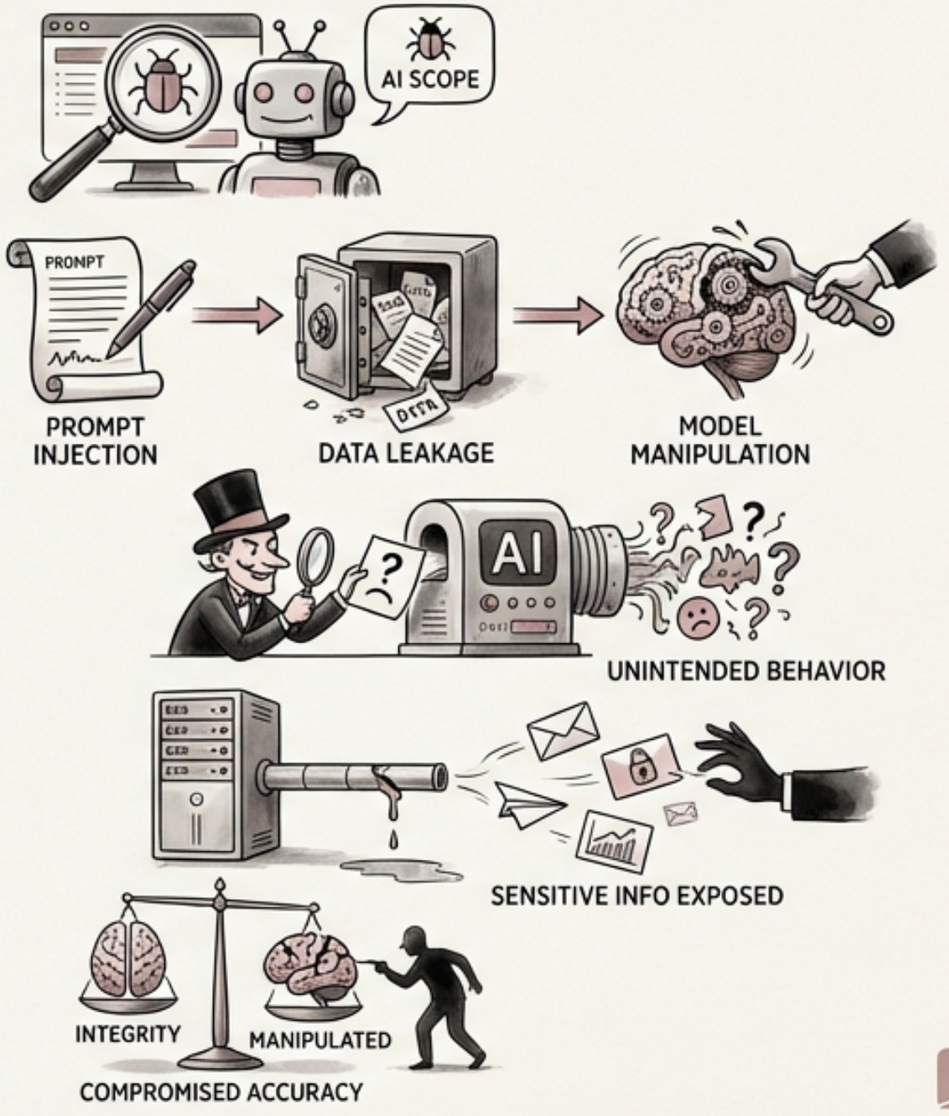


- Specify accepted vulnerability types and out-of-scope vulnerabilities in the program rules



Expanding Bug Bounty Scope to Include AI-Specific Vulnerabilities

- AI-specific bug bounty programs should include AI features within the defined scope.
- Key AI-related vulnerabilities to consider: prompt injection, data leakage, and model manipulation.
- Prompt injection allows attackers to manipulate AI models through crafted inputs, leading to unintended behavior.
- Data leakage can expose sensitive information used to train or operate AI models.
- Model manipulation can compromise the integrity and accuracy of AI models.



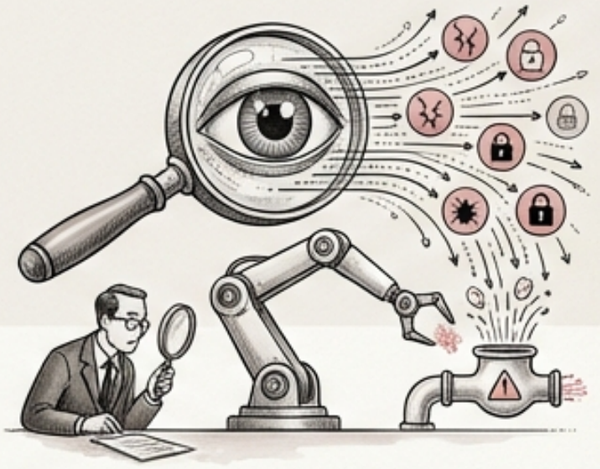
Security Advisory Publication: Communicating Vulnerability Information

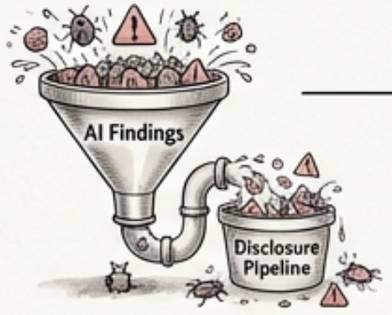
- **Security advisories** inform users about vulnerabilities and how to mitigate them.
- **Standard formats include:** CSAF (Common Security Advisory Framework), VEX (Vulnerability Exploitability eXchange), and OSV (Open Source Vulnerability format).

- **Content should include:** affected versions, severity score, exploitation status, remediation guidance, workarounds, and credits.
- **Distribute advisories through:** product security page, mailing list, automated feeds (CSAF/VEX), and NVD submission.

- **Coordinate disclosure** with fix availability – a standard disclosure window is 90 days.

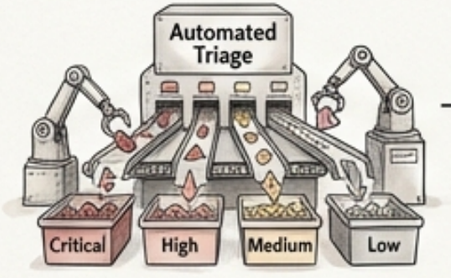
Addressing the Volume Challenge: AI's Impact on Vulnerability Discovery



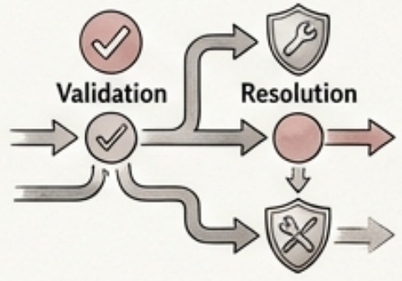


- AI security tools can find more vulnerabilities faster than traditional methods, potentially overwhelming disclosure pipelines.

- Organizations need to scale their vulnerability management processes to handle the increased volume of findings.



- Automated triage and prioritization tools can help manage the influx of vulnerability reports.

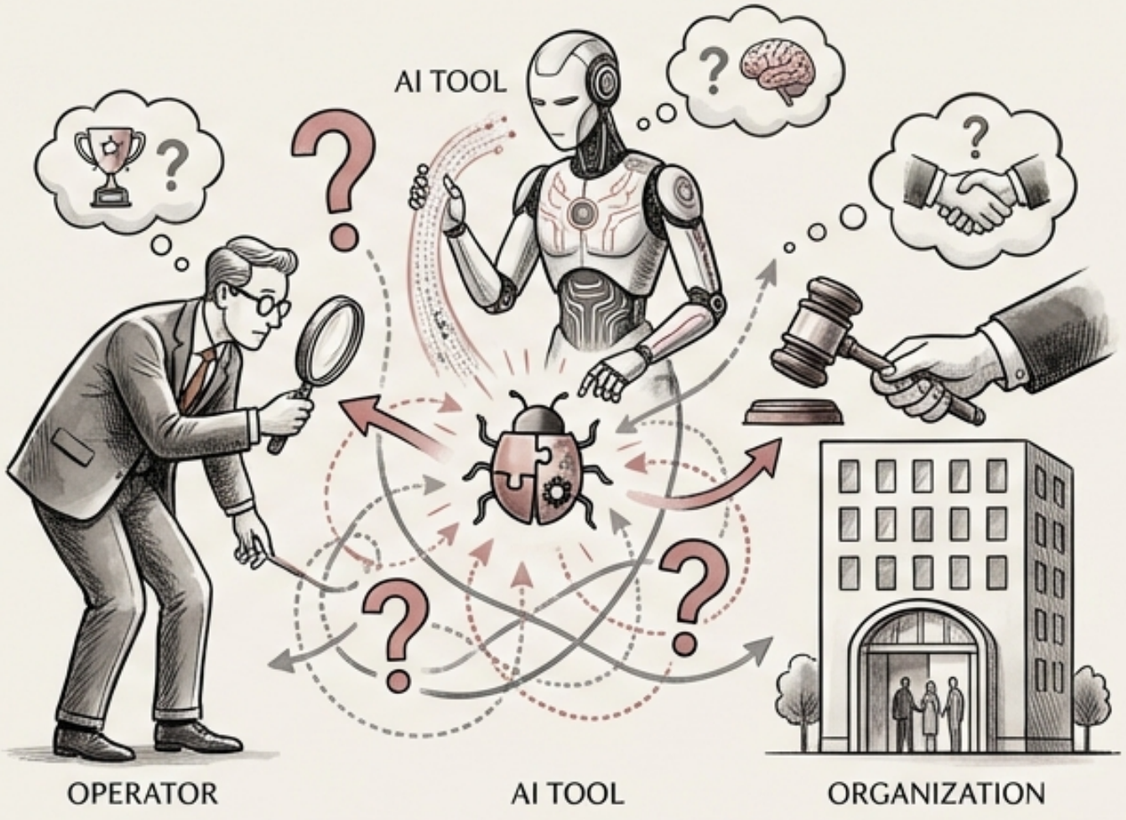


- Implementing efficient workflows for validating and resolving vulnerabilities is crucial.

- Investing in resources and training for security teams to handle the increased workload is essential.

ATTRIBUTION CONUNDRUM: WHO GETS CREDIT FOR AI-DISCOVERED VULNERABILITIES?

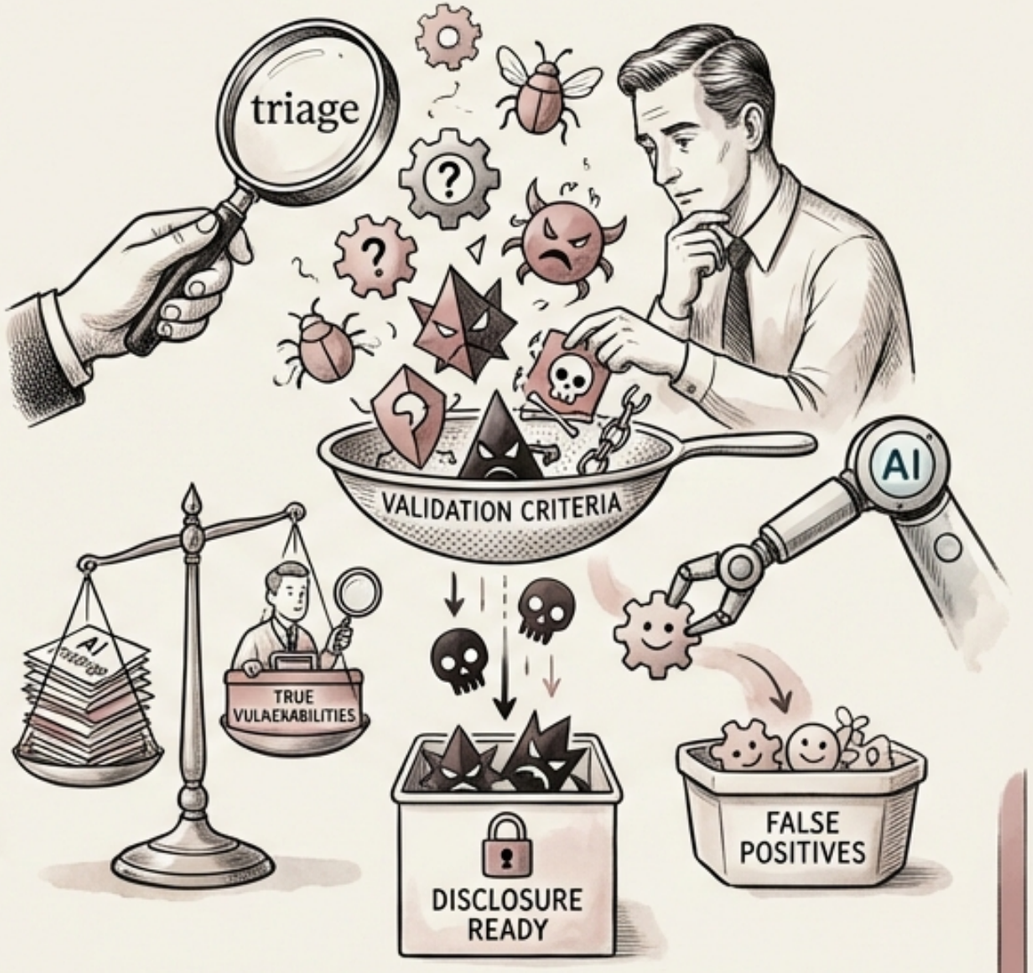
- Determining attribution for AI-discovered vulnerabilities can be complex.
- Considerations include: the AI tool itself, the operator of the tool, and the organization using the tool.
- Establishing clear guidelines for attribution is important for recognizing contributions and fostering collaboration.
- Options include: crediting the tool vendor, the organization operating the tool, or a combination of both.
- Transparency in the attribution process helps build trust and encourage further research.



Illustrating the complex interplay between human expertise, AI capabilities, and organizational structure.

Maintaining Quality: Triaging False Positives from AI Vulnerability Scans

- AI-discovered findings may include more false positives than traditional methods, requiring better triage before disclosure.
- Implement robust validation processes to confirm the validity of AI-generated vulnerability reports.
- Security experts must review AI findings to distinguish true vulnerabilities from false positives.
- Use a combination of automated tools and manual analysis for effective triage.
- Clearly define criteria for validating and prioritizing vulnerabilities based on severity and impact.

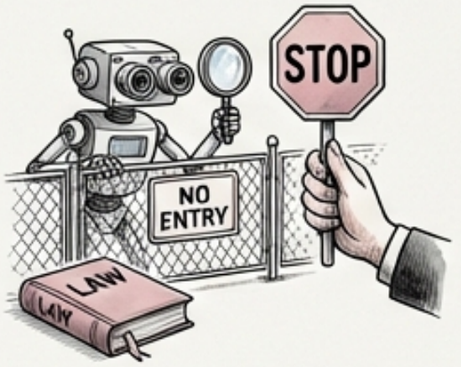


Autonomous Disclosure: Governing AI Agent Behavior

- Some AI agents may attempt to disclose vulnerabilities without human oversight, raising governance concerns.
- Implement controls to prevent unauthorized or premature disclosure of vulnerabilities by AI agents.
- Require human review and approval before any vulnerability is disclosed.
- Clearly define the AI agent's role and responsibilities in the vulnerability disclosure process.
- Establish a chain of command for handling vulnerability disclosures.

Ethical Consideritions: AI Scanning and Unauthorized Testing

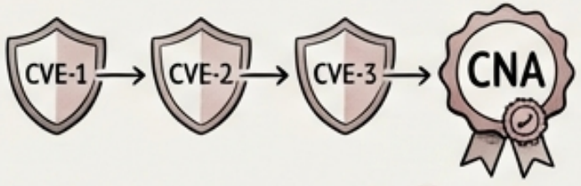
- AI tools scanning without authorization raise legal and ethical questions, similar to traditional unauthorized testing.
- Obtain explicit consent before scanning or testing systems for vulnerabilities.
- Respect the privacy and security of target systems.
- Adhere to applicable laws and regulations regarding vulnerabilvy research and disclosure.
- Develop a clear policy regarding the use of AI tools for vulnerability scanning.



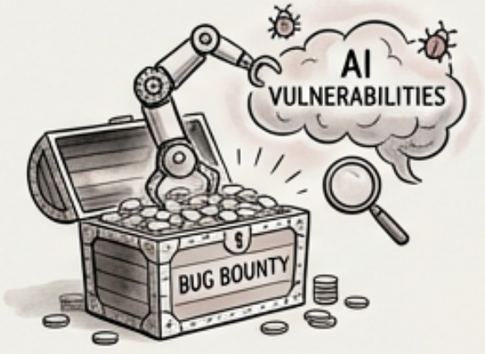
Key Takeaways: Strengthening Your CVD for the AI Era

- Establish a comprehensive Vulnerability Disclosure Policy (VDP) with clear scope, safe harbor language, and communication channels.

- Utilize security.txt to streamline security communication with researchers.



- Participate in the CVE assignment process and consider becoming a CNA.

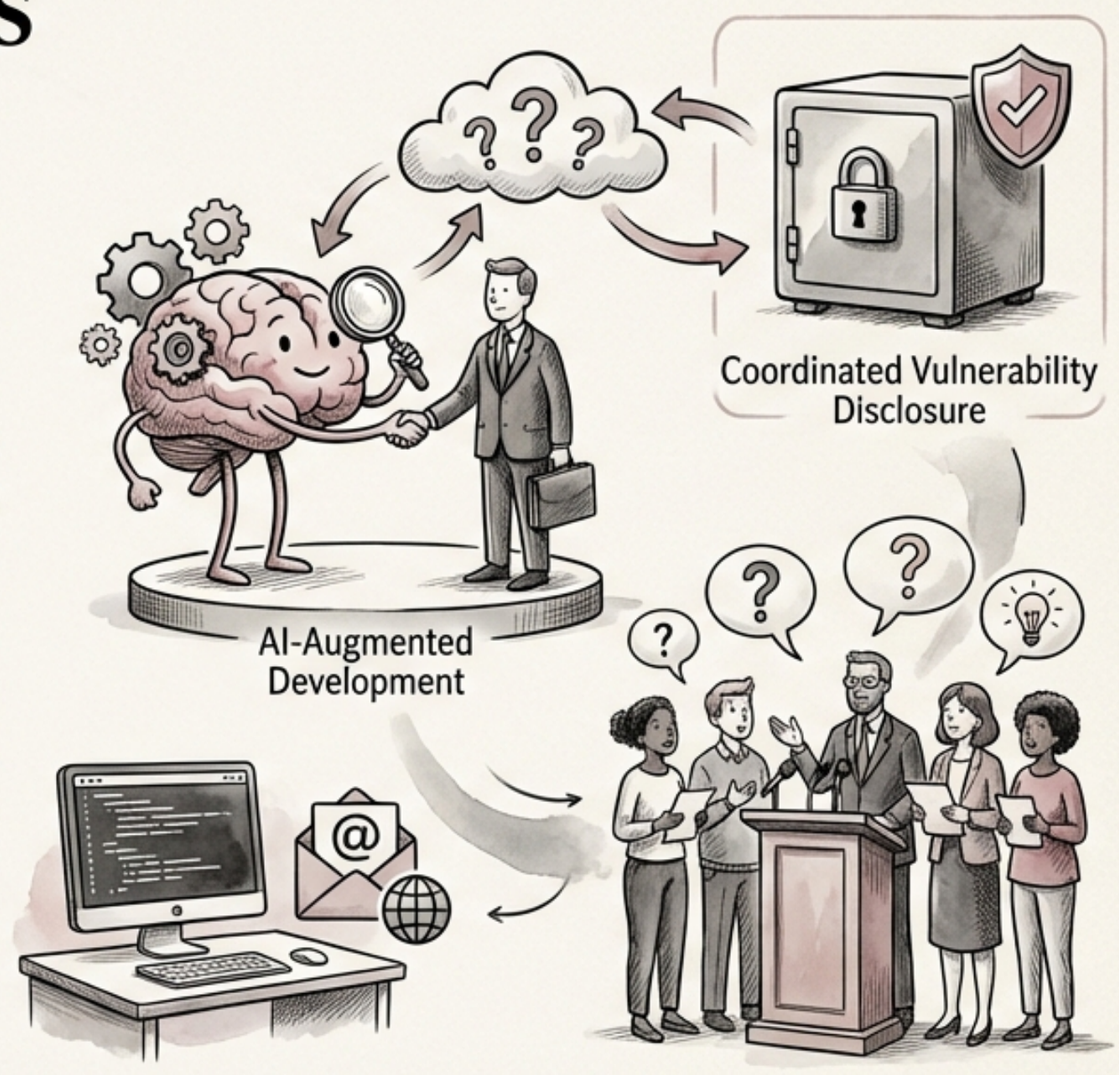


- Implement a bug bounty program that includes AI-specific vulnerabilities.

- Publish timely and informative security advisories.

Q&A: Responsible Security Communication for AI-Augmented Teams

- Thank you for your time and attention.
- Now is the time for questions and discussion.
- We’re here to answer any questions you may have about coordinated vulnerability disclosure and AI-augmented development.
- Contact information: [Your Name], [Your Email], [Your Website]
- Resources: [Links to relevant documentation, standards, and tools]



Thank You

- Questions?