hackerone

Securing the Future of AI

A Comprehensive Guide to Ethical and Security Risks

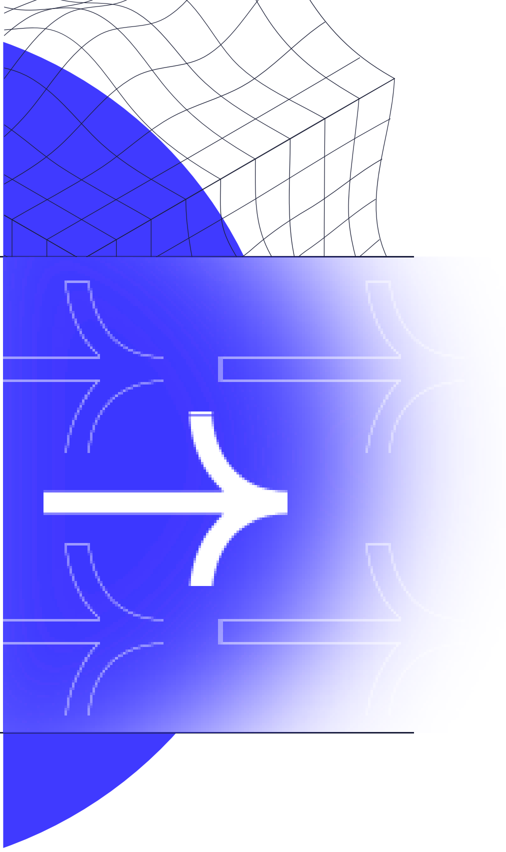


Table of Contents

More than half of security professionals admit that basic practices are overlooked in the rush to implement AI.[1]

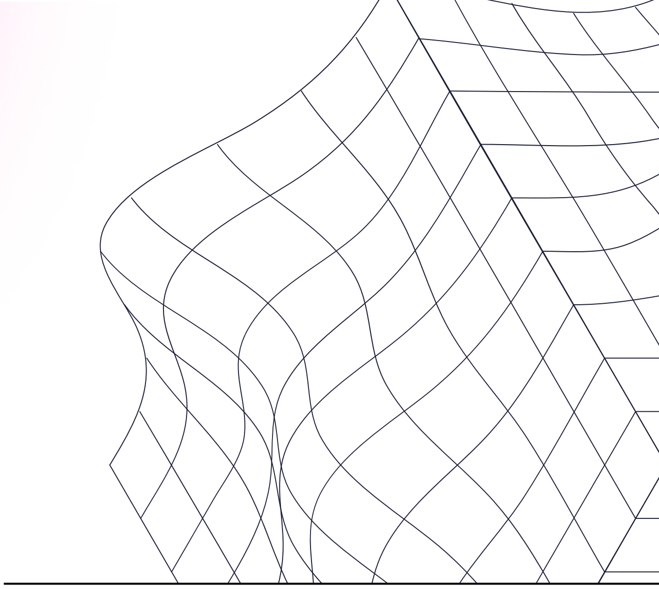
And as this technology continues to reshape business, security, trust, and compliance teams are struggling to stay ahead of emerging threats.

How can teams turn AI from a security risk into a security asset? At HackerOne, our direct customer engagements provide unique insights into AI evolution. This guide provides those insights so you can navigate this shifting landscape with confidence and clarity.

Key Takeaways

- Offensive AI is outpacing defensive AI, forcing organizations to adopt new security strategies to keep pace.
- Security researchers are crucial in securing AI by stress-testing models before deployment.
- AI red teaming is a strategic, proactive defense against AI-driven security threats.
- AI tools are available to streamline your vulnerability remediation processes and automate calculations to prove the value of your security program.

[1] HackerOne. Hacker-Powered Security Report. 8th Edition.



The Risks: An Evolving AI Threat Landscape

Developers of leading AI products, such as OpenAI, Grok, Google DeepMind, and our customer Anthropic, are at the forefront of creating foundational AI technologies—ranging from generative AI (GenAI) models and natural language processing systems to predictive AI tools that power recommendations, forecasts, and automated decision making. Meanwhile, integrators like our customers Snap, Instacart, CrowdStrike, Priceline, Cloudflare, X (Twitter), and Salesforce are deeply integrating these AI advancements into their offerings. Both of these groups pushing AI forward must take an offensive approach to prevent adversarial attacks, data leaks, and model manipulation that could compromise their success.

Offensive AI Is Outpacing Defensive AI

Cybersecurity has always been a cat-and-mouse game—but in the age of AI, offense is gaining the upper hand. Malicious actors are weaponizing AI faster than defenders can secure it, leveraging AI-generated malware, automated social engineering, and novel exploitation techniques at scale. The speed and efficiency with which AI can automate these processes amplify the scale of cyberthreats.

Since the rise of AI in 2022, phishing attacks have risen by

1,265 %[2]

Deloitte predicts that GenAI could enable fraud losses to grow 32% every year.[3]

[2] McKinsey & Company. The cybersecurity provider's next opportunity: Making AI safer.

[3] Deloitte Insights. Generative AI is expected to magnify the risk of deepfakes and other fraud in banking.

"The downside of AI is that it introduces more vulnerabilities. If a company uses it, we'll find bugs in it. AI is even hacking other AI models. It's going so fast and security is struggling to catch up."

Jasmin Landry, @jr0ch17
Security Researcher and HackerOne Pentester

Cybercriminals Are Using AI to Enhance Attacks

As businesses rush to integrate AI into their applications, attackers are discovering ways to manipulate AI models and exploit inherent weaknesses. Prompt injection—where adversaries craft malicious inputs that force AI models to disclose sensitive data or perform unintended actions—is already one of the most common AI-specific vulnerabilities. And deepfake-enabled fraud is rising, with attackers leveraging AI-powered voice cloning and synthetic media to execute convincing social engineering campaigns.

In 2024, the use of generative AI tactics such as deepfakes and deepaudio increased by 118%[4]

Bad actors—including those linked to nation-states and organized cybercrime groups—are leveraging GenAI tools to write malware, identify vulnerabilities, and conduct reconnaissance on targets. AI-powered cybercrime campaigns can now automate tasks that once required human expertise, significantly scaling up the volume and complexity of attacks.[5, 6]

The Power of AI Voice Cloning

Have you ever received a text from a random number claiming to be your CEO, asking you to buy 500 gift cards? While you're unlikely to fall for that trick, what if that phone call came from your CEO's phone number? What if it sounded exactly like your CEO, and the voice even responded to your questions in real time? That's the power of AI voice cloning. Security teams must implement voice authentication safeguards and user training programs to combat this growing threat.

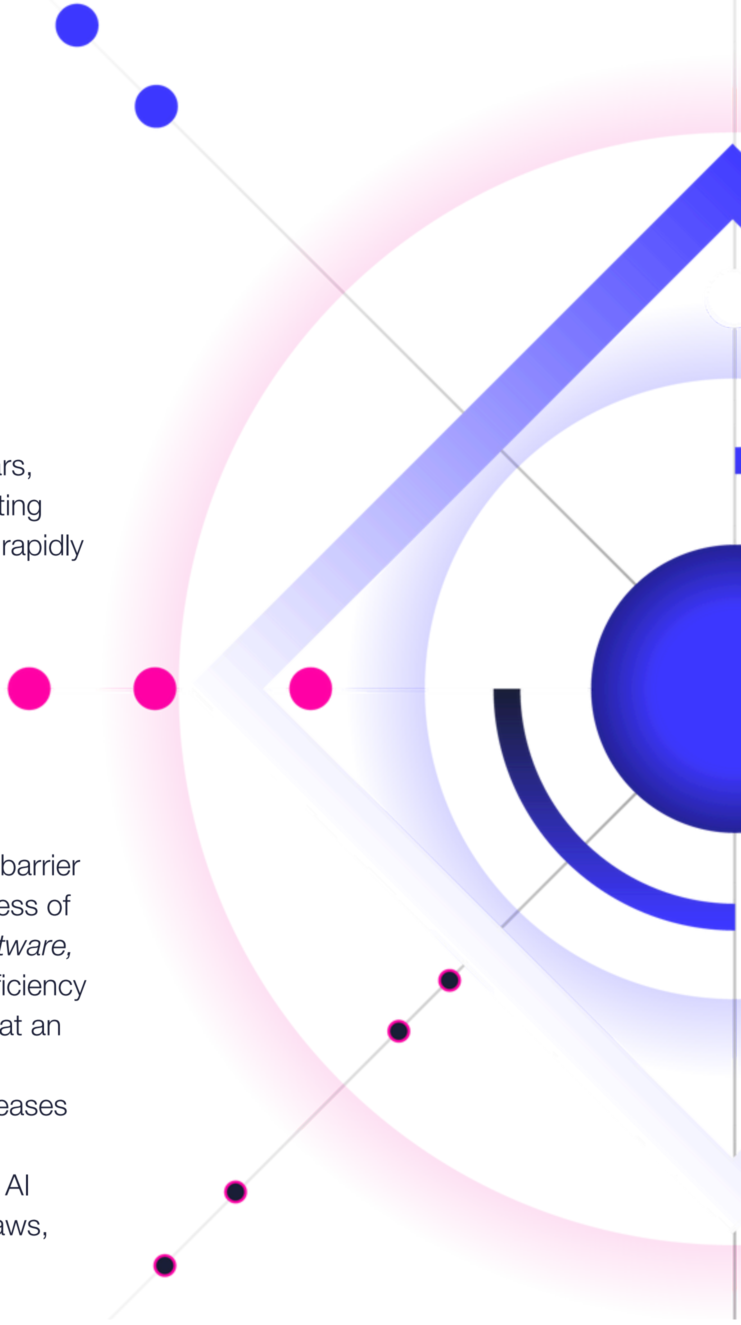
[4] Business Wire. Ninety Percent of U.S. Companies Experienced Cyber Fraud in 2024, According to New Trustpair Research.

[5] The Wall Street Journal. Chinese and Iranian Hackers Are Using U.S. AI Products to Bolster Cyberattacks.

[6] PCMag. After WormGPT, FraudGPT Emerges to Help Scammers Steal Your Data.

Check out this AI voice cloning Q&A with HackerOne innovations architect and AI security researcher Dane Sherrets.

Attack Surfaces Are Growing Exponentially



We're seeing an explosion in new attack surfaces. For years, defenders have focused on attack surface reduction—limiting the number of entry points attackers can exploit. But AI is rapidly expanding the attack surface instead.

As AI capabilities become embedded in more applications, workflows, and infrastructure, organizations are introducing entirely new avenues for exploitation—often faster than they can secure them.

The ability to generate code via AI dramatically lowers the barrier to entry for software development. Now, anyone—regardless of security knowledge—can ship code. The result? *More software, written faster, with less oversight.* And while speed and efficiency are great for innovation, they also introduce vulnerabilities at an unprecedented scale. AI-generated code does not always adhere to security guidelines or best practices, which increases the risk of introducing bugs, insecure dependencies, and misconfigurations. This also creates supply chain risks, as AI may pull in third-party components with hidden security flaws, further expanding the attack surface.

Beyond insecure code, GenAI runs on vast amounts of data—and attackers know it. The most powerful AI models are also the largest, trained on datasets that organizations are only beginning to understand how to secure. As AI adoption grows, so does the sheer scale of sensitive data being collected, stored, and inevitably targeted. The dark web economy for stolen data is booming, with cybercriminals eager to monetize AI training datasets, proprietary models, and the sensitive user data that fuels them.

Attack surface expansion doesn't stop there. Businesses are integrating AI into applications at record speed, often without fully understanding the security risks. With this rapid deployment comes a wave of novel attack vectors—from vulnerabilities specific to large language models (LLMs), such as prompt injection, to indirect attacks through AI-integrated plugins, APIs, and automation workflows.

The bottom line: AI is changing the game for both attackers and defenders. Organizations can no longer rely on traditional security playbooks. They must anticipate how AI is reshaping their attack surface—and prepare for threats that are evolving faster than ever.

The Regulatory Landscape and Business Imperatives Are Evolving

As AI adoption accelerates, regulatory bodies worldwide are mandating increased security measures, including red teaming, adversarial testing, and risk mitigation strategies. There is also growing momentum to adapt best practices from cybersecurity to evaluate AI, such as the implementation of vulnerability disclosure programs (VDPs) or bug bounty programs for AI flaws.

Stay informed on the global VDP policy landscape with our interactive map, and keep up with the latest news, policy developments, and regulatory updates by following HackerOne's public policy blog.

European Union's AI Act

The European Union has implemented the AI Act, aiming to establish a comprehensive regulatory framework for AI applications. The AI Act seeks to ensure that AI systems are safe, transparent, and respect fundamental rights. Notably, the Act requires adversarial testing, cybersecurity incident reporting, and risk mitigation for AI systems deemed high-risk, including those used with transportation, medical devices, and critical infrastructure.

U.S. Federal Guidance

Unlike the EU, the United States has so far taken a more flexible, non-legislative approach to AI regulation. While emphasis can shift over time, AI security and innovation remains a critical issue across administrations. One key initiative is the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), which provides a structured approach to identifying, assessing, and managing AI-related risks. This framework emphasizes securing AI systems throughout their lifecycle.

Global AI Security Efforts

- OWASP's Gen AI Security Project identifies top vulnerabilities in AI-powered applications, shaping security testing methodologies and equipping developers with best practices for mitigating AI-specific risks.

- The G7's Code of Conduct for AI emphasizes the need for independent security testing and ongoing AI risk evaluation. This initiative underscores a commitment to harmonizing AI ethics and security standards across major global economies.

- The United Kingdom's AI Cyber Security Code of Practice establishes voluntary baseline cybersecurity requirements across the AI lifecycle and is expected to inform changes to international standards through the European Telecommunications Standards Institute.

AI Safety vs. AI Security: What's the Difference?

Security leaders are trying to understand how to leverage GenAI technology while ensuring protection from inherent security issues and threats. This challenge includes staying ahead of adversaries who may discover and exploit malicious uses before organizations can address them. Alongside those AI security efforts, teams are faced with a new frontier: AI safety. These are two interconnected yet distinct domains that collectively ensure the responsible development and deployment of artificial intelligence.

AI safety is about protecting the outside world from the AI system.

It focuses on preventing AI systems from generating content that could pose reputational or legal risk to the companies that deploy them —from instructions for creating weapons to offensive language and inappropriate imagery. It aims to ensure responsible use of AI and adherence to ethical standards.

→ **55%**

of all AI vulnerabilities reported on the HackerOne Platform are AI safety issues, highlighting the growing attention to ensuring that AI behaves as intended, without causing unforeseen consequences.

AI safety issues often have a lower barrier to entry for valid reporting and present a different risk profile compared to traditional security vulnerabilities. However, that figure does not mean AI safety issues overwhelmingly outnumber AI security issues.

AI security is about protecting the AI system from threats in the outside world.

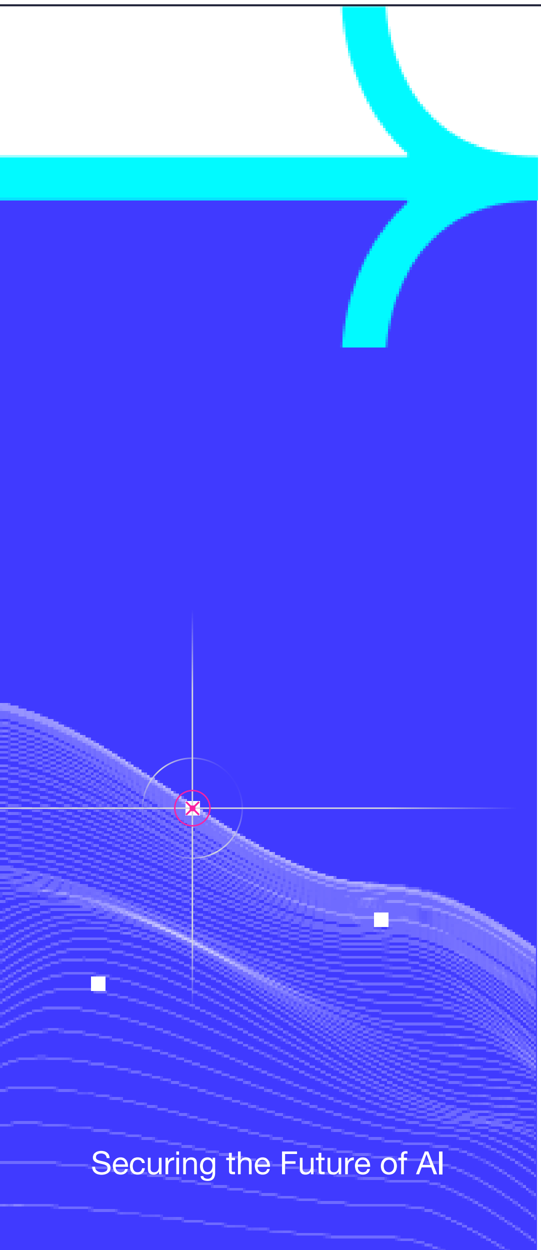
It involves testing AI systems with the goal of preventing bad actors from abusing the AI to, for example, compromise the confidentiality, integrity, or availability of the systems the AI is embedded in or interacts with.

AI safety risks can result in:

- Spread of biased or unethical decision-making
- Legal and regulatory penalties for non-compliance with ethical standards
- Erosion of public trust in AI technologies and the organizations that deploy them
- Unintended consequences that could harm individuals or society

AI security risks can result in:

- Disclosing sensitive or private information
- Providing access and functionality to unauthorized users
- Compromising a model's security, effectiveness, and ethical behavior
- Doing extensive financial and reputational damage



AI Safety and Security Best Practices

- **Establish continuous testing, evaluation, verification, and validation throughout the AI model lifecycle.** Provide regular executive metrics and updates on AI model functionality, security, reliability, and robustness. Regularly scan and update the underlying infrastructure and software for vulnerabilities.
- **Determine country-, state-, or government-specific AI compliance requirements.** Some regulations exist around specific AI features, such as facial recognition and employment-related systems. Establish an AI governance framework outlining roles, responsibilities, and ethical considerations, including incident response planning and risk management.
- **Train all users on ethics, responsibility, legal issues, AI security risks, and best practices** such as warranty, license, and copyright. Establish a culture of open and transparent communication on the organization's use of predictive or generative AI.

For more, download this detailed AI safety and security checklist.

ANTHROP\C

Real-world AI safety testing: Anthropic's jailbreak challenge

To test the robustness of its Claude Sonnet GenAI model, Anthropic partnered with HackerOne to conduct an AI red teaming challenge aimed at stress-testing its Constitutional Classifiers—a system designed to block harmful queries, particularly those related to chemical, biological, radiological, and nuclear topics.

The seven-day challenge invited researchers to bypass Claude's safety measures using a universal jailbreak—a single input that could consistently defeat the model's defenses across multiple scenarios.

The system successfully resisted jailbreaking attempts for five days—but then one participant discovered a true universal jailbreak, while another developed a borderline-universal jailbreak that required some manual modifications to succeed. Two more participants found significant vulnerabilities.

The most effective jailbreaking strategies included:
Encoding tricks, such as using ciphers to evade content filters Role-playing prompts that subtly manipulated the model Keyword substitution, such as replacing restricted terms with innocuous ones. Prompt injection attacks to override model safeguards

The challenge generated over 300,000 chat interactions and a total of $55,000 in bounty rewards.

"This challenge demonstrated the high return on investment for collaborative efforts. Delivering large language models (LLMs) in a safe and aligned manner is a significant challenge.... As these models get smarter, our strategies for testing must evolve to stay ahead of potential risks."

Salia Asanova (@saltyn)
Security Researcher

Real-world AI security testing: Google Gemini's extensions

Security researchers Joseph "rez0" Thacker, Johann "wunderwuzzi23" Rehberger, and Kai Greshake collaborated to strengthen Google's AI red teaming by hacking its GenAI assistant, Bard—now called Gemini.

The launch of Bard's Extensions AI feature provided Bard with access to Google Drive, Google Docs, and Gmail. This meant Bard would have access to personally identifiable information and could even read emails and access documents and locations. The security researchers identified that Bard analyzed untrusted data and could be susceptible to insecure direct object reference (IDOR) and data injection attacks, which can be delivered to users without their consent.

In less than 24 hours from the launch of Bard Extensions, the security researchers were able to demonstrate that:

- Google Bard was vulnerable to IDOR and data injection attacks via data from Extensions.
- Malicious image markdown injection instructions will exploit the vulnerability.
- A prompt injection payload could exfiltrate victims' emails.

With such a powerful impact as the exfiltration of personal emails, the security researchers promptly reported this vulnerability to Google, which resulted in a $20,000 bounty award.

And that was just the start. Read about another recent Google Gemini vulnerability that Johann Rehberger discovered.

Discover the OWASP Top 10 for LLM Applications

The Open Web Application Security Project (OWASP) has unveiled its "Top 10 for LLM Applications 2025," high-lighting the most pressing security risks to large language model applications. This essential guide is designed to help organizations stay ahead of evolving AI vulnerabilities. For an in-depth look, explore our perspective on how to address these challenges and enhance your AI security strategy.

Read HackerOne's Analysis of the OWASP Top 10 for LLMs

The Opportunity: Collaboration With Security Researchers to Secure & Safeguard Your AI Tools

Security researchers have been at the forefront of experimenting with AI systems since the introduction of models like ChatGPT. Their innate curiosity and expertise make them invaluable allies for organizations aiming to implement AI both swiftly and securely. More than two-thirds (68%) of security leaders believe that an external, unbiased review of AI implementations is the most effective way to uncover AI safety and security issues.[7] In a third-party survey of HackerOne customers, 91% agree or strongly agree that "Hackers provide more impactful and valuable vulnerability reports than AI or scanning solutions."

[7] HackerOne. Hacker-Powered Security Report. 8th Edition.

> *"Human ingenuity is more effective than consistently using adversarial prompt datasets or LLM-written attacks."*
>
> **Ilana Arbisser,**
> Technical Lead, AI Safety at Snap Inc.

> *"The program with HackerOne has surfaced the most interesting results across all of our AI testing and is by far the most cost-effective."*
>
> **Security Leader,**
> San Francisco–Based AI Research and Development Company

What Security Researchers Think About Top AI Risks

While the OWASP Top 10 for LLMs is a comprehensive study of the types of vulnerabilities that can affect GenAI models, we also maintain an ongoing conversation with the security researchers to learn what they encounter most often and which vulnerabilities organizations need to look out for. Here are a few of their insights.

"We've almost forgotten the last 30 years of cybersecurity lessons in developing some of this software."

> **Gavin Klondike**
Security Researcher

"There are now suddenly a whole host of attack vectors for AI-powered applications that weren't possible before—tricking the AI into doing or revealing something it shouldn't."

"If an attacker uses prompt injection to take control of the context for the LLM function call, they can exfiltrate data by calling the web browser feature and moving the data that are exfiltrated to the attacker's side. Or, an attacker could email a prompt injection payload to an LLM tasked with reading and replying to emails."

"LLMs are as good as their data. The most useful data is often private data."

> **Joseph Thacker, aka @rez0**
Security Researcher

"What AI does remarkably well is use existing information and content to deliver something creative. Something I'm worried about is data that currently doesn't seem valuable or exploitable gains far more value when combined with AI."

> **Herman Satkauskas, aka oxalis**
Security Researcher

"ChatGPT hallucinates library names, which threat actors can then take advantage of by reverse-engineering the fake libraries."

> **Roni Carta, aka @arsene_lupin**
Security Researcher

"This is a great opportunity to take a step back and bake some security in as this is developing, and not bolting on security 10 years later."

"As we see the technology mature and grow in complexity, there will be more ways to break it. We're already seeing vulnerabilities specific to AI systems, such as prompt injection or getting the AI model to recall training data or poison the data. We need AI and human intelligence to overcome these security challenges."

> **Katie Paxton-Fear, aka @InsiderPhD**
Security Researcher

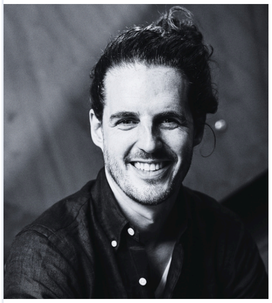
"I love that the bug bounty program gives me visibility into things that I'm not aware of...the things my scanners or even my SDLC are not looking after. The human factor is very difficult to overcome, and with the addition of AI, they're going to be able to find crazy vulnerabilities that we probably would spend a year looking for."

> **Christopher Von Hessert,**
VP, Security at Polygon Labs

Tom Anthony has direct experience with the change in how security researchers (aka hackers) approach processes with AI, noting that AI will significantly uplevel the reading of source code:

"Anywhere that companies are exposing source code, there will be systems reading, analyzing, and reporting in an automated fashion." He also shared that, "At a Live Hacking Event with Zoom, there were easter eggs for hackers to find—and the hacker who solved them used LLMs to crack it. Hackers can use AI to speed up their processes by, for example, rapidly extending the word lists when trying to brute-force systems."

Tom Anthony,
Security Researcher

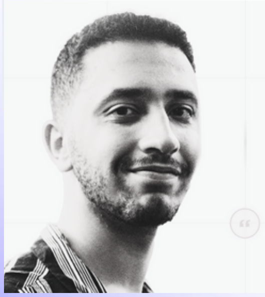


Security researcher Jonathan Bouman uses ChatGPT—and collaboration with his researcher peers—to enrich his own expertise and uncover critical vulnerabilities:

"I can hack web applications but not break new coding languages, which was the challenge at one Live Hacking Event. I copied and pasted all the documentation provided (removing all references to the company), gave it all the structures, and asked it, 'Where would you start?' It took a few prompts to ensure it wasn't hallucinating, and it did provide a few low-level bugs. Because I was in a room with 50 ethical hackers, I was able to share my findings with a wider team, and we escalated two of those bugs into critical vulnerabilities. I couldn't have done it without ChatGPT, but I couldn't have made the impact I did without the hacking community."

Johnathan Bouman
Security Researcher

So, how do you engage these creative security researchers to make your AI implementations safe and secure?
AI red teaming.

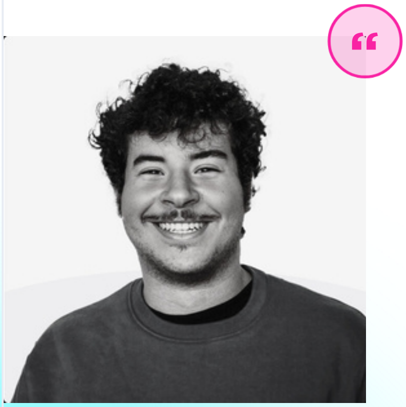


"I leverage AI-powered vulnerability scanners to quickly identify potential weak points in a system, allowing me to focus on more complex and nuanced aspects of security testing. I also use AI for reporting. Previously, I spent 30-40 minutes writing reports to ensure all details were included, the tone was appropriate, and there were no grammatical mistakes. AI has streamlined this process, reducing the time to an average of 7-10 minutes per report."

Hazem Elsayed, @hacktus
Security Researcher

How Security Research Is Evolving in the AI Age

AI tools are now a key part of security researchers' bag of tricks, and adoption is growing rapidly. 33% of security researchers on the HackerOne Platform use AI to summarize information and write reports. At the end of 2024, 20% of researchers considered tools such as “hackbots” an essential part of vulnerability discovery—up from 14% in 2023. Many researchers also engage with online games designed to help people understand prompt injection.—such as by working through levels to get a chat model to divulge secrets.



“When pentesting, I use AI to automate repetitive and time-consuming tasks so I can concentrate on finding security issues. I also use AI to summarize documentation when I want a general overview of a new technology. When I do content discovery before my pentest, AI allows me to generate customized wordlists to find niche content that can fly under the radar of commonly used wordlists."

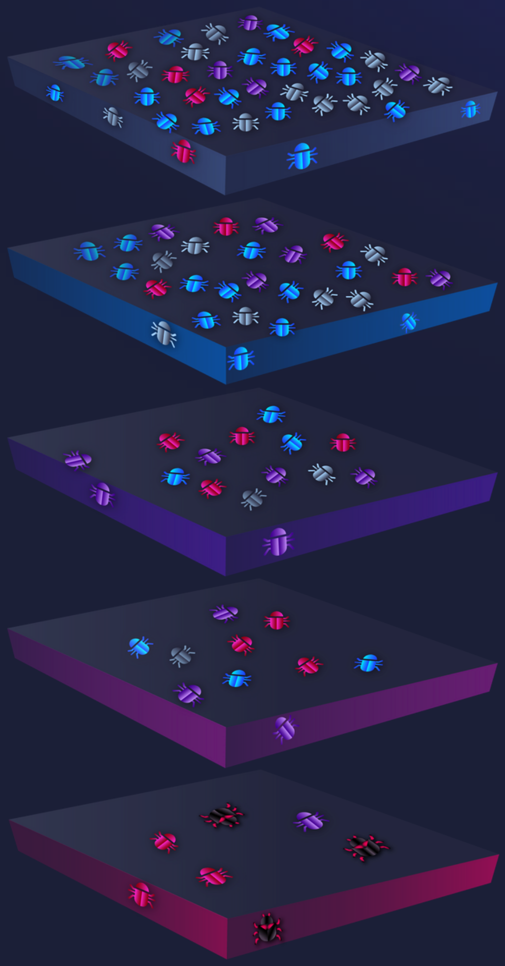
Adam Deziri, @a_d_a_m
Security Researcher

The Solution: A Defense in Depth Approach to AI

As AI systems become increasingly central to business operations, a *defense in depth* approach ensures that security scales in parallel with innovation—not behind it. At its core, defense in depth means building overlapping layers of security across the entire AI development and deployment lifecycle, ensuring that if one control fails, others are in place to prevent compromise. This strategy not only protects against known vulnerabilities, but also adapts to AI-specific threats such as prompt injection, model manipulation, and training data poisoning.

Here are the pillars of a defense in depth approach:

- **“Secure by design” principles:** Integrate AI-assisted security checks early in development—within pull requests or IDEs—to catch vulnerabilities before they reach production.
- **Automated + human testing:** Combine AI-powered scanning with human validation to improve accuracy, reduce false positives, and prioritize real risks.
- **Human-driven penetration testing, enhanced by AI:** Call on human testers to bring creativity and contextual awareness, while AI helps streamline repetitive tasks and enrich findings with tailored insights.
- **Adversarial testing & AI red teaming:** Simulate real-world attacks to identify gaps in AI model behavior and system integration that other methods may miss.



Snap's Challenge: AI Safety for Text-to-Image Technology

Snap is constantly refining its AI-powered functionality to enhance its users' creativity, and wanted to stress-test the guardrails of its Generative AI Lens and Text2Image products to prevent the generation of harmful content. The company sought to go beyond traditional safety methods: instead of analyzing patterns in user behavior, it needed to assess the AI model itself to uncover edge cases of inappropriate content that could emerge from model flaws.

Using a Bug Bounty Model to Incentivize Researchers

Snap collaborated with HackerOne to conduct a Capture the Flag (CTF)–style red teaming exercise, adapting bug bounty methodologies for AI safety testing. This engagement focused on testing for violence, self-harm, explicit content, and other harmful imagery. The findings from this initiative would not only improve the specific products tested but also contribute to Snap's AI safety benchmark dataset—a crucial tool for enhancing AI model safety across the platform.

A key component of the program was HackerOne Clear, which enabled Snap to select vetted, age-appropriate researchers for participation.

Out of a deep pool of talented researchers, 21 experts from across the globe were selected to participate. Global diversity was crucial for covering harmful imagery across different cultures, and the researchers' mindset was key for breaking the models.

> *"We use AI red teaming to determine qualitative safety aspects—what's possible, not necessarily what's likely. We're also constantly surprised by what's possible—we try to keep an open mind while designing exercises."*
>
> **Ilana Arbisser,**
> Technical Lead, AI Safety at Snap Inc.

The Result: Stronger AI Safety Benchmarks

Through this engagement, Snap successfully identified specific vulnerabilities in its AI models, leading to targeted mitigations. The AI red teaming exercise helped set new benchmarks for AI safety testing by demonstrating a scalable, systematic approach to stress-testing generative AI products. And those benchmarks can help other social media companies, which can use the same flags to test for content safety.

> *“Snap has helped HackerOne refine its playbook for AI red teaming, from understanding how to price this type of testing to recognizing the wider impact the findings can deliver to the entire GenAI ecosystem. We’re continuing to onboard customers onto similar programs who recognize that a creative, exhaustive human approach is the most effective modality to combat harm.”*
>
> **Dane Sherrets,**
> Senior Solutions Architect at HackerOne

> *“As time goes on, these areas will become less novel, and we will be able to rely more on automation and existing datasets for testing. But human ingenuity is crucial for understanding potential problems in novel areas.”*
>
> **Ilana Arbisser,**
> Technical Lead, AI Safety at Snap Inc.

Let’s Strengthen AI Security & Safety Together

Emerging technologies are often developed with trust, safety, and security as afterthoughts. In collaboration with our customers, HackerOne is changing the status quo.

We are committed to enhancing security through safe, secure, and confidential AI, tightly coupled with strong human oversight. Our goal is to provide organizations with the tools they need to achieve security outcomes beyond what has been possible before—and to do it without compromise.

As the demand for secure and safe AI grows, HackerOne remains dedicated to facilitating a present and future where technology enhances our lives while upholding security and trust.

To learn more about how to strengthen your AI safety and security with AI red reaming, contact the AI red teaming experts at HackerOne.