CASE STUDY

# Snap Inc. and HackerOne:

## Pioneering AI Red Teaming and Celebrating a Decade of Partnership

For over a decade, Snap Inc. has partnered with HackerOne to stay ahead in the evolving landscape of cybersecurity. From leading-edge AI red teaming exercises to reaching the impressive milestone of $1M in bounties paid to security researchers, this collaboration exemplifies the power of human creativity of the researcher community and the development of Snap's program to build a safer platform for their users.

**Leading in AI Red Teaming for Safety and Security**
Snap has been an early adopter of AI red teaming, working with HackerOne to test and refine strict safeguards for generative AI technologies. Together, they've developed innovative methodologies to surface previously unknown vulnerabilities in AI systems, contributing to a safer, more ethical digital landscape.

**AI Safety Red Teaming:** Focuses on preventing the generation of harmful content, such as offensive language or instructions for dangerous activities.

**AI Security Red Teaming:** Ensures bad actors can't exploit AI systems to compromise confidentiality, integrity, or availability.

> *AI red teaming allows us to explore the possibilities of what attackers might achieve—not just what's likely. Working with HackerOne has shown us that human ingenuity often outperforms adversarial datasets or AI-generated attacks.*
>
> ***Ilana Arbisser,***
> *Technical Lead, AI Safety at Snap Inc.*

### Key Milestones

**Global diversity:**
Snap selected 21 researchers from around the globe for their AI red teaming exercises, ensuring diverse perspectives in identifying harmful content.

**Safety innovations:**
Snap used over 100 flags in the first exercise, dynamically adjusting bounties to optimize researcher engagement. The second exercise focused on higher-value flags, pushing the boundaries of what the researchers could achieve.

**Adopting Hai, HackerOne's AI Copilot:**
During a private CTF hackathon, Snap leveraged HackerOne's AI tool Hai to translate submissions into seven European languages, making it easier to communicate with security researchers worldwide.

### Innovating with Capture-the-Flag (CTF) Exercises

Snap and HackerOne adopted a groundbreaking approach to stress-test AI models using CTF-style bug bounty programs. These exercises incentivized researchers to uncover vulnerabilities in Snap's generative AI products, such as the Lens and My AI Text2Image features, targeting harmful imagery like violence, self-harm, and inappropriate content.

This innovative approach provided valuable insights into AI models' behavior and informed Snap's safety benchmarks, which have become a blueprint for testing harmful content across the tech industry.

### Celebrating $1M in Bounties: A Decade of Collaboration

Reflecting on their 10-year partnership with HackerOne, Snap's Chief Information Security Officer, Jim Higgins, emphasized the transformative role of bug bounty programs in shaping their security and privacy strategies.

### Lessons Learned from the Researcher Community

**Fix low and medium bugs:**
Don't only focus on remediating the major bugs. When chained together, minor issues can lead to critical vulnerabilities.

**Build trust with the community:**
Treating researchers as allies fosters high-quality submissions.
Gamify engagement: Swag, live hacking events, and creative challenges keep researchers motivated.

> “
>
> *Hitting $1M in bounties is a badge of honor. It reflects our commitment to valuing the intelligent security researchers who help keep us safe. Bug bounty programs are notoriously difficult to build, but HackerOne's talented community provides us with the expertise and creativity we need to secure our platform. Our ultimate goal is to make Snap's bug bounty program a model for others to follow.*
>
> ***Jim Higgins,***
> *Chief Information Security Officer at Snap Inc.*

## **Looking Ahead:**
## The Future of AI Security and Safety

Snap plans to expand its bug bounty program to include hardware products like AR glasses and deepen its focus on AI security and safety. They aim to refine AI red teaming through simulated testing with LLM agents, offering scalable solutions for emerging vulnerabilities.

From pioneering AI red teaming methodologies to celebrating a decade of partnership, Snap is setting new standards for security in the digital age. They've demonstrated that human ingenuity, combined with innovative tools and strategies, is essential for navigating the challenges of tomorrow's technology.

## With over 4,000 customer programs, more companies trust HackerOne than any other vendor

gm Lufthansa ZEBRA zoom

Spotify citrix PayPal Uber HYATT

U.S. Department of Defense Google reddit Nintendo Adobe

A.S. Watson Group coinbase yahoo! priceline

shopify slack yelp salesforce TOYOTA

**Contact us**

**About HackerOne**

HackerOne pinpoints the most critical security flaws across an organization's attack surface with continual adversarial testing to outmatch cybercriminals. HackerOne's Attack Resistance Platform blends the security expertise of ethical hackers with asset discovery, continuous assessment, and process enhancement to reduce threat exposure and empower organizations to transform their businesses with confidence. Customers include Citrix, Coinbase, General Motors, GitHub, Goldman Sachs, Hyatt, PayPal, Singapore's Ministry of Defense, Slack, the U.S. Department of Defense, and Yahoo. In 2023, HackerOne was named a Best Workplace for Innovators by Fast Company.