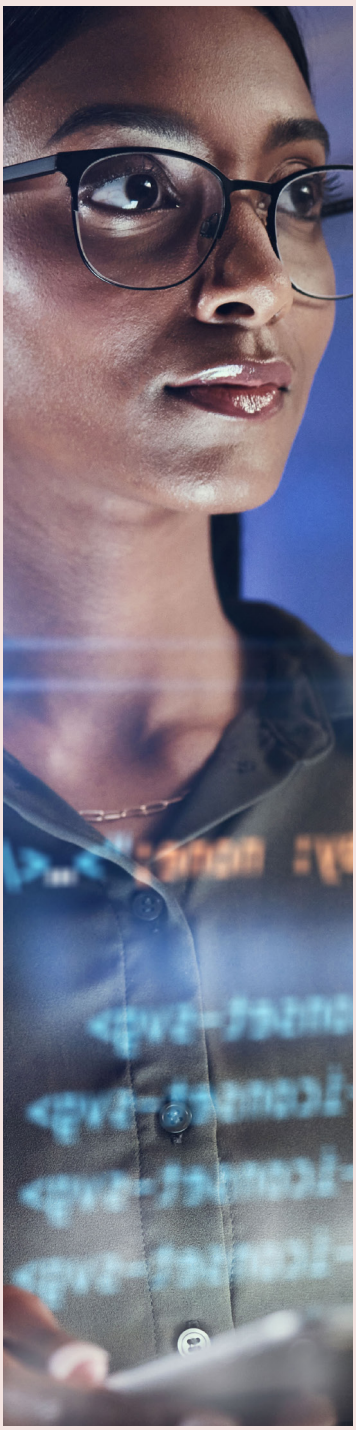
CASE STUDY



HACKING AWAY AT AI, ETHICALLY

Abstract

When Betty Parman, Senior Manager for AI Security and Governance at a global manufacturer of high-tech electronic devices, wanted to improve the security posture of its AI systems, she partnered with Infosys Responsible AI Office and Infosys BPM's specialist expertise. This case details how Infosys BPM's automated and manual red teaming exercises successfully identified and mitigated over 4,200 vulnerabilities, thus strengthening the systems' overall resilience.

Infosys®
Navigate your next

Wide open to attack

A leading global manufacturer of high-tech devices, which also develops the software for its products in-house, had realized the need to make its internal consumption of Artificial Intelligence (AI) more disciplined, robust, and secure. Its AI system, designed to serve diverse employee needs, was exposed to risks from code generation, compliance checks, and interactions with multiple third-party applications. As Senior Manager for AI Security and Governance, Betty Parman had charge of managing the security posture of the company's AI systems, coordinating risk assessments, and ensuring robust protection against evolving threats.

Betty's initial evaluations of the AI systems' prompt-responses had revealed glaring gaps in alignment and accuracy. What worried Betty even more was that because the AI models had not undergone rigorous security testing, they were vulnerable to sophisticated attacks and misuse. Though the situation highlighted the need for improved validation processes, it would be a significant challenge to ensure comprehensive security testing and vulnerability management for the complex, organization-wide AI platform.

However, considering the risks to the company's data integrity, compliance, and operational reliability, Betty moved quickly to engage an expert partner who could strengthen the organization's AI safety and security posture. Through a global RFP process, Betty found that Infosys Responsible AI Office, which brought deep governance, risk, and compliance capabilities, and Infosys BPM, with its strong technology and security execution expertise, matched her requirements perfectly, and soon brought a team of their security specialists, headed by team lead Prakash Jha, on board. After a series of meetings with Prakash and his team, to explain the AI safety and security gaps in the organization, Betty tasked them with conducting red teaming security assessments. Using this method, they would need to simulate real-world attacks on the organization to identify the cracks in its defences, mimicking the tactics used by actual adversaries.

Putting on the red team hats

Prakash, along with the Responsible AI Team, soon began their tenacious endeavours to breach the organization's cybersecurity defences.

First, they performed an initial assessment of the AI system architecture and its risk vectors. Then they curated a dataset of 60,000 prompts across 9 taxonomies and sourced over 40 prompt techniques to combine with the attack payloads. Then after fine-tuning an end-to-end production pipeline for the automated attack and response evaluation, they researched and mapped the prompts to proprietary taxonomy techniques. And soon, the automated attack simulations were underway.

Next up were the manual red teaming exercises in which Prakash and his team conducted rigorous testing and evaluation for every part of the system, exploring and utilizing over 16 distinct attack strategies. The team also conducted image guardrail testing, curating datasets of safe and unsafe images, developing a proof-of-concept, as well as a preprocessing pipeline for multimodal evaluation.

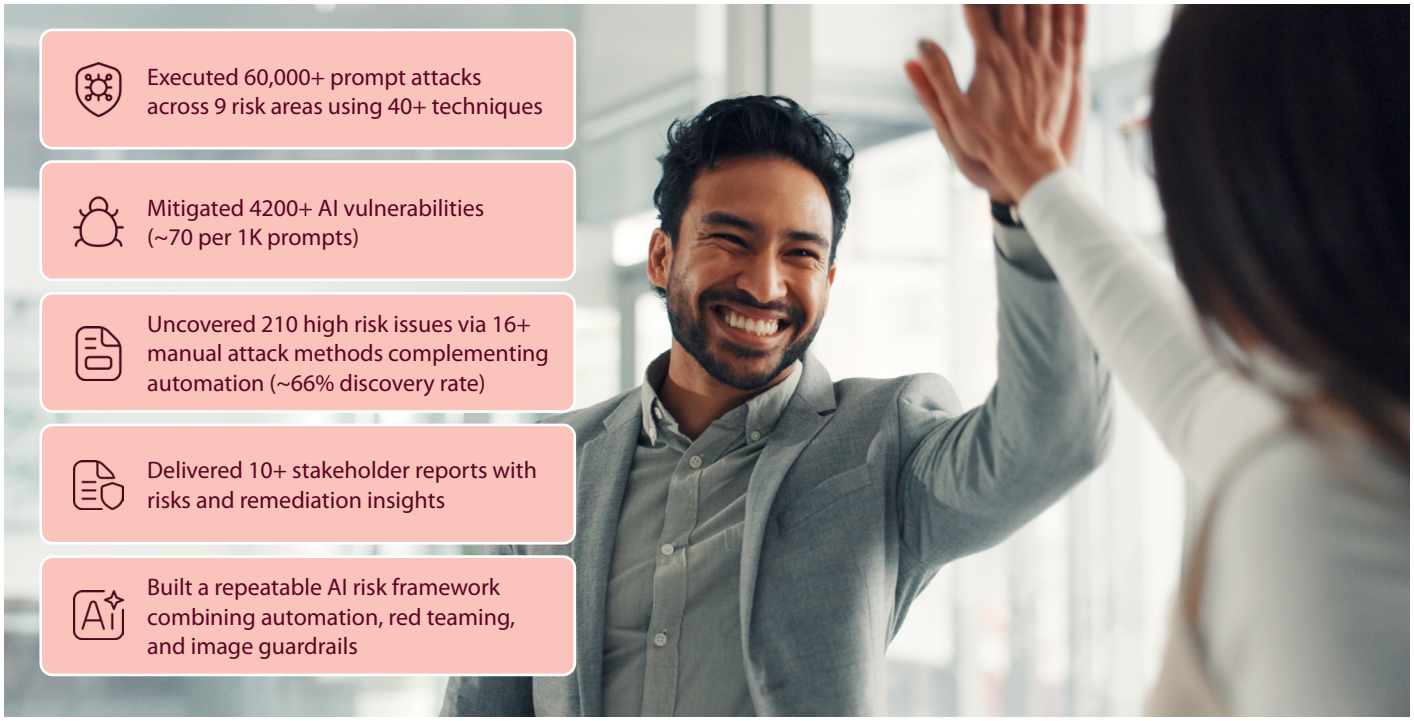
Throughout the red teaming assessments, Prakash undertook all the needed steps to eliminate any potential disruption to the company's regular operations. He used a phased implementation approach, kept all the stakeholders engaged through transparent communications, and provided comprehensive reporting. Later, he also directed his security experts to provide remediation of the identified vulnerabilities and continuous improvement of the AI system's security controls.

Approach summary

Everyone wins when the red team wins

When Betty sat down to review the outcomes of the red teaming exercises, she was completely satisfied. Prakash's team had successfully navigated the scale and diversity of the company's AI system, identifying its gaps and model vulnerabilities, integrating automated and manual testing pipelines, and ensuring comprehensive coverage of attack vectors. They had also mitigated potential risks such as missed vulnerabilities or delayed remediation through iterative testing and stakeholder collaboration.

Key benefits



In all, Prakash's automated red teaming assessments delivered valuable insights through over 10 detailed reports for stakeholders, explaining attack success rates and metrics for 60,000 prompt attacks. The Responsible AI team successfully identified and mitigated over 4,200 vulnerabilities, strengthening the system's overall resilience. Moreover, their manual red testing further uncovered and documented 210 unique vulnerabilities out of 320 attack attempts, providing even more actionable insights for remediation. The team's efforts had also improved the organization's prompt-response evaluation accuracy across multiple models and prompt types.

Thus, Betty's proactive initiative gave the company critical insights into potential vulnerabilities in its AI systems before they could be exploited. As Prakash's team worked on mitigating these risks, they ensured robust AI security and compliance. The improved risk posture, over time, led to greater stakeholder confidence, through strengthening trust in AI deployments, reducing the likelihood of regulatory breaches, and positioning the organization as a leader in secure AI adoption.

**Names have been altered to preserve the identities of the people involved.*

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next