Google Cloud

AI in SecOps: Red teams

Previously, you learned about the different components of security operations management, including the role of red teams in vulnerability management. A **red team** is a group of ethical hackers who mimic potential adversaries in order to examine the security defenses of an organization. The emergence of generative artificial intelligence (GenAI) means organizations and red teams have more attack vectors to take into consideration.

In this reading, you'll learn about the red team's role in artificial intelligence (AI) security, and explore examples of simulated attacks against AI technology.

AI and red teams

Red teams are an important part of an organization's security strategy. By testing for vulnerabilities in the various components of a system, red teams help build a stronger line of defense against adversarial parties.

The introduction of GenAI brings new complexities into systems, so red teams must increasingly be aware of how these complexities impact their system's security. Performing simulated attacks against AI systems is one way organizations can better prepare for modern technological threats.

Some organizations, like Google, create specialized AI red teams to specifically test the security of their AI deployments. AI red teams have additional expertise in AI subject matter so they can perform these more advanced attacks. Organizations that don't have the resources to employ a dedicated AI red team should encourage traditional red teams to collaborate with AI subject matter experts.

AI red teams in action

Similar to conducting tests on traditional systems, AI red teams use different types of simulated attacks against AI systems. These attacks can take a variety of forms.

One way a red team can simulate an attack is to execute a prompt attack. Prompt attacks are a way for attackers to bypass AI security features. A prompt attack works by instructing the AI model to follow untrusted input. For example, an organization uses AI to label suspicious emails as "spam" or "phishing." An attacker can trick the AI model by adding undetectable instructions to the email. They might carry out this maneuver by formatting the malicious input in white text that instructs the AI model to label the email as legitimate. Since the white text blends in with

the background of the email, the untrusted input goes undetected by the recipient, and the email isn't flagged as spam. The recipient may unintentionally fall for the phishing scheme.

Another way red teams can evaluate the integrity of a GenAI model is to focus on how an AI model is reading and converting information. For example, images can have embedded text that AI models scan. The text can instruct the AI model to ignore certain parts of an image, or to mislabel what's shown in the image. This is one way threat actors can manipulate data, leading to skewed datasets. This process is called data poisoning, where adding incorrect data to the model skews its accuracy. If the model doesn't detect the incorrect data, then it should be retrained with a better training dataset. Research indicates that poisoning .01% of a dataset can affect the entire model, making it essential for organizations to monitor and maintain the training data.

Other security considerations

Organizations that adopt GenAI must take into account certain security considerations. GenAI's ability to generate several content types presents opportunities for risk. As part of their tests, red teams might inspect the AI system for data leakage and an abuse of rights.

A data leak is when sensitive information is unintentionally disclosed. When data is leaked, unauthorized users are able to access the data, which can lead to a serious security breach. Data leaks can expose sensitive information like employee credentials. Applying inadequate security boundaries or using internal documentation to train an AI model are examples of how AI can leak data. Security analysts that work with GenAI systems have to ensure that the GenAI model:

- Hasn't been fed unapproved data
- Hasn't been granted access to too much information that could skew the generated content
- Has clearly defined security boundaries
- Can absorb real-time data without leaking information or skewing the model's intended functionality
- Follows best practices outlined for making application programming interface (API) calls that it has been programmed to do

Another important security consideration is monitoring for abuse of rights. An abuse of rights can occur when too many privileges are granted to a user or application. Organizations that deploy GenAI-integrated models or tools must be diligent in validating the access those tools are granted. For example, an abuse of rights can happen with failed role-based access controls (RBACs). RBAC, a subset of identity and access management (IAM), is a method of controlling access to resources based on the roles assigned to users, services, and groups. If a user or tool is given a role with too many permissions, the GenAI gains access to that privileged information, potentially leaking sensitive data into the dataset. Red teams can test for abuse of

rights by gaining access to the system, finding faulty RBACs, and escalating their privileges to retrieve data.

Key takeaways

Red teams are a critical part of helping bolster an organization's security posture. With organizations increasingly adopting AI models, red teams have additional security elements to consider. Exercises like simulating attacks, and best practices like monitoring RBAC, are ways that red teams can help continuously test, and help secure how systems integrate with GenAI.