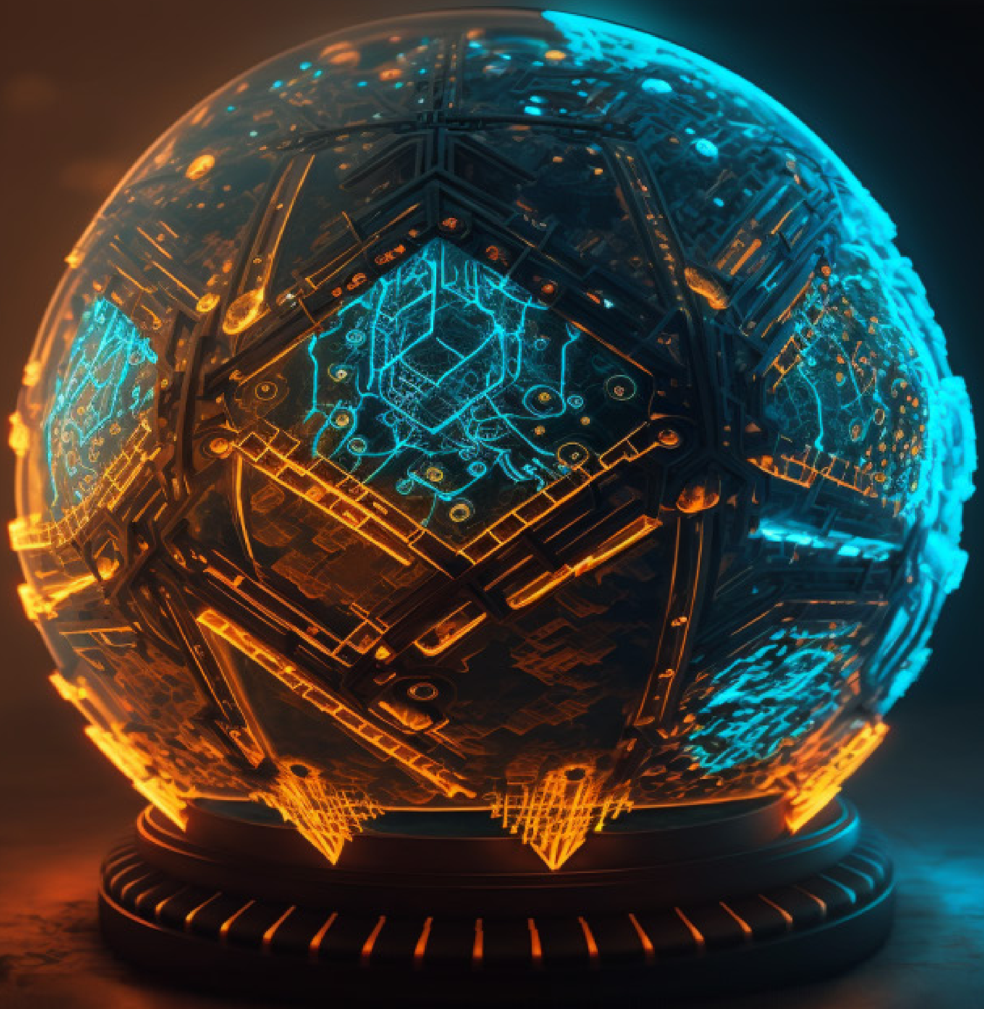
ChatGPT Security

DISCOVERING AND SECURING AI TOOLS

ASHUR KANOON



BANYAN SECURITY

THE PROBLEM

Today's cybersecurity threats have their roots in many places.

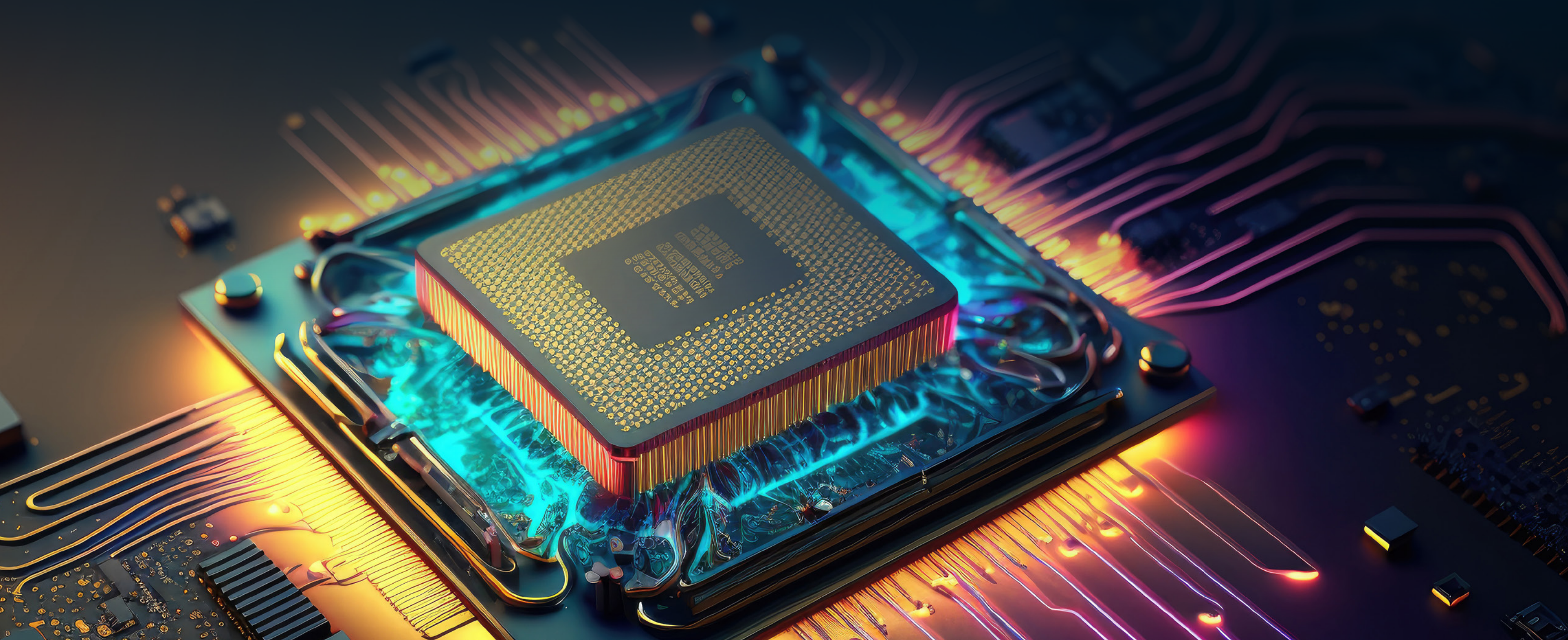
Some zero-day threats are from the software developers themselves, while others come from powerful tools used for bad purposes. ChatGPT is an example of a powerful tool often used nefariously. On the helpful side of the equation, generative AI and Large Language Models (LLMs) offer a wide range of benefits for companies across various industries: automating and streamlining processes saves valuable time and resources. These AI systems generate high-quality content, such as product descriptions, marketing copy, or even entire articles, reducing the need for manual content creation. This not only speeds up the production process, but also ensures consistent and engaging messaging. This productivity boost applies to the darker side of production as well: it's just as easy to ask ChatGPT or another LLM to generate complicated exploits that once took hackers years to perfect.

Additionally, generative AI and LLMs can assist with data analysis, uncovering patterns and insights that humans may overlook, thereby enhancing decision-making capabilities. By leveraging these technologies, companies can gain a competitive edge by quickly adapting to market trends and customer preferences. Moreover, generative AI and LLMs can improve customer experiences through personalized interactions and tailored recommendations, leading to increased customer satisfaction and loyalty.

Remember the darker side of the ChatGPT security story: employees are using these resources to do their jobs but may also be leaking protected corporate information by feeding schematics, statistics, instructions, and other intellectual property into the LLMs. ChatGPT security took center stage recently when Samsung employees leaked intellectual property into ChatGPT (including both confidential product information and meeting notes). Such risks are leading more and more organizations (such as Apple) to try to block these sites. As the number of generative AI and LLM tools and companies grows, the problem of protecting against them becomes more challenging.

Of course, these AI systems can facilitate research and development efforts by simulating and generating ideas, designs, and prototypes, expediting innovation cycles. Unfortunately, they also create a wide range of security issues for companies because of the behaviors noted above, in addition to attackers searching LLMs for carelessly shared company data.

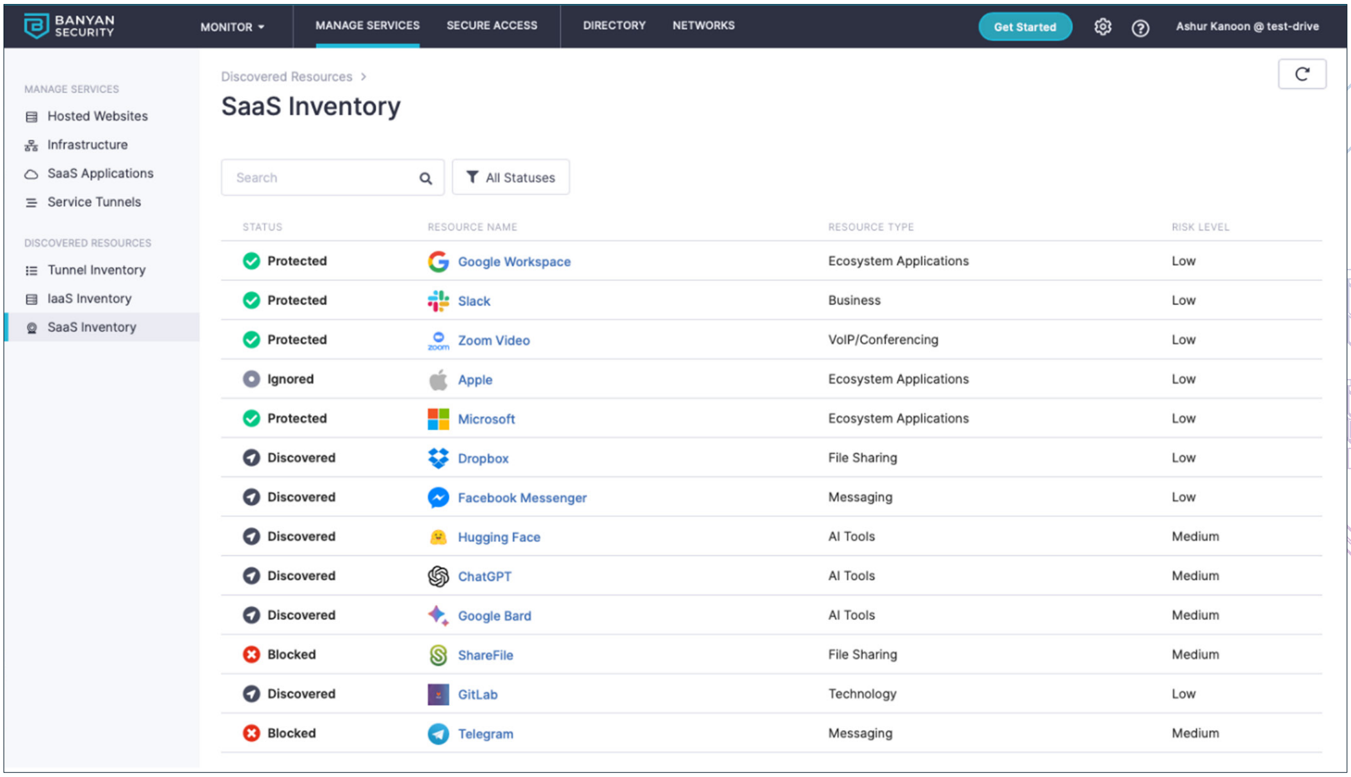
There is a profound danger in reactive security without strategy, and much opportunity for overcorrection. Some of the solutions for ChatGPT security include blocking access by directing all traffic over a VPN, and then using an outbound security stack to inspect traffic. Eventually, though, employees find new ways to get around some of these blocks, or hunt for other tools that aren't blocked.



THE SOLUTION

Discovery is important and a crucial step to combating data exfiltration.

Security for AI should be able to detect quickly, then categorize accurately where the data is going. With new AI tools popping up daily, this isn’t always so easy. Banyan’s solution looks at all DNS transactions and its real-time categorization engine assesses a range of information. Our security for AI also inspects traffic for sensitive data, such as PII, PHI, Secrets and Keys, PCI, using a modern cloud-based Data Loss Prevention (DLP) engine. The administrator sees where users are going, when, from what device this traffic originates, as well as what type of data is being sent.

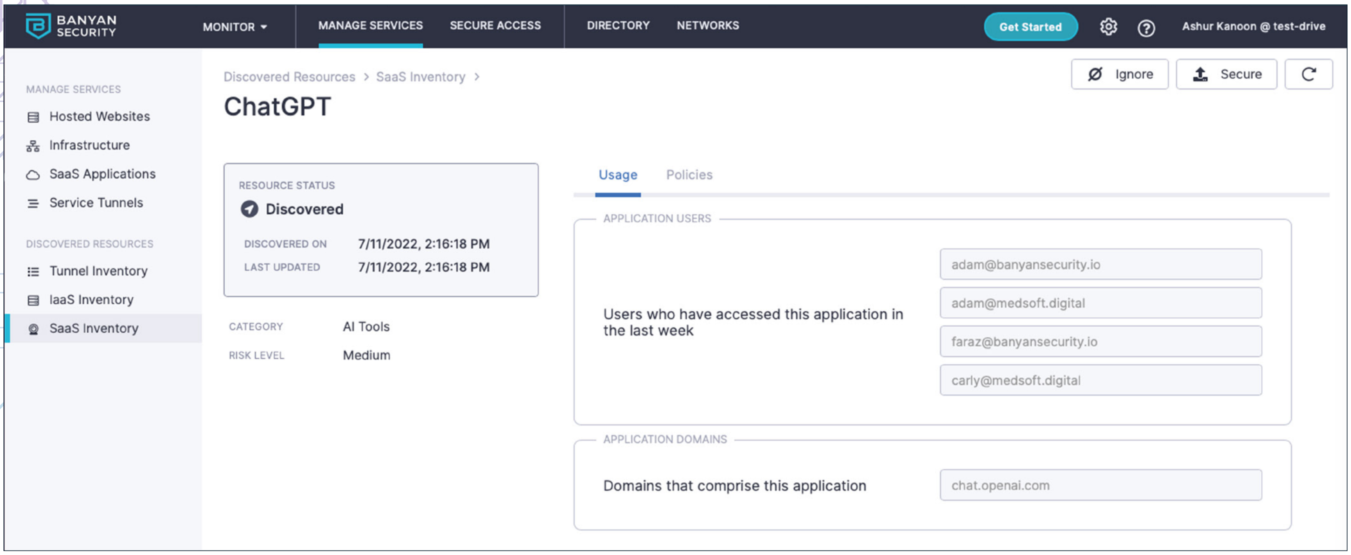


Discovered Resources >

SaaS Inventory

STATUS	RESOURCE NAME	RESOURCE TYPE	RISK LEVEL
Protected	Google Workspace	Ecosystem Applications	Low
Protected	Slack	Business	Low
Protected	Zoom Video	VoIP/Conferencing	Low
Ignored	Apple	Ecosystem Applications	Low
Protected	Microsoft	Ecosystem Applications	Low
Discovered	Dropbox	File Sharing	Low
Discovered	Facebook Messenger	Messaging	Low
Discovered	Hugging Face	AI Tools	Medium
Discovered	ChatGPT	AI Tools	Medium
Discovered	Google Bard	AI Tools	Medium
Blocked	ShareFile	File Sharing	Medium
Discovered	GitLab	Technology	Low
Blocked	Telegram	Messaging	Medium

It is worth noting that our solution is always-on, so end users will benefit from the protection without having to do anything. Administrators also gain visibility without needing to configure anything extra, or have their users do anything. As soon as the first user visits the first website, the administrator gets actionable insights. These are presented as applications and categories: much easier to use and create policies for those rather than just configuring policies around domains. A single SaaS application can have hundreds or thousands of domains, so being able to quickly find a SaaS application (and how it’s been used) is the first step to creating a comprehensive policy.



Discovered Resources > SaaS Inventory >

ChatGPT

RESOURCE STATUS
Discovered

DISCOVERED ON	7/11/2022, 2:16:18 PM
LAST UPDATED	7/11/2022, 2:16:18 PM

CATEGORY	AI Tools
RISK LEVEL	Medium

APPLICATION USERS

Users who have accessed this application in the last week

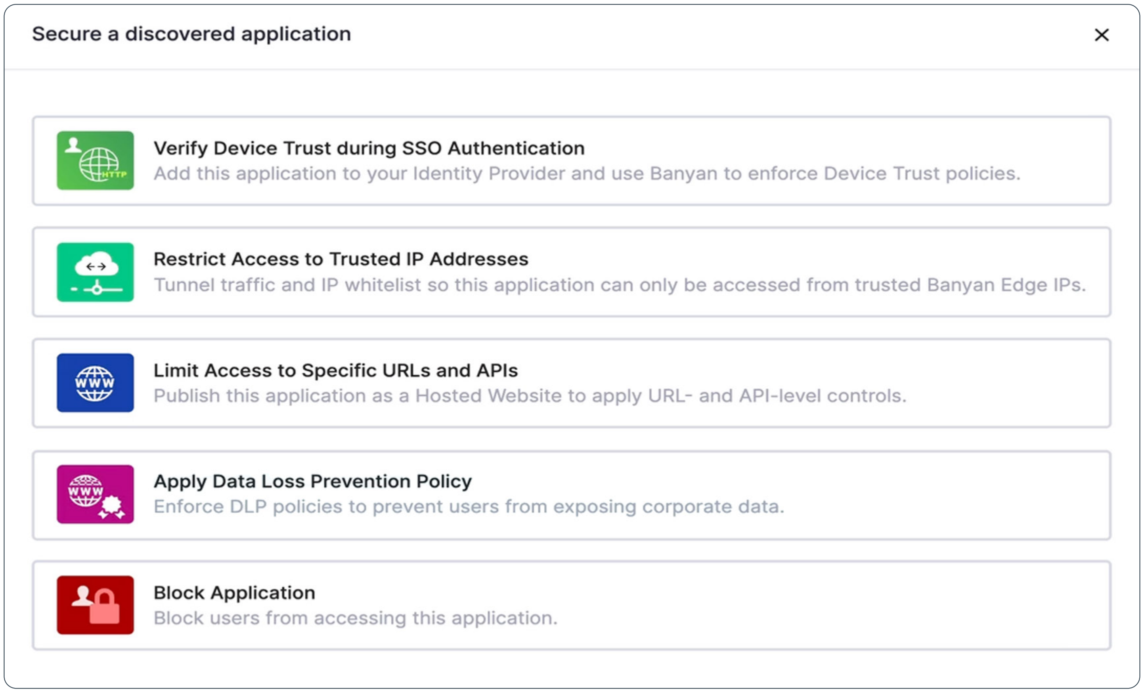
adam@banyansecurity.io
adam@medsoft.digital
faraz@banyansecurity.io
carly@medsoft.digital

APPLICATION DOMAINS

Domains that comprise this application

chat.openai.com

Once you've Discovered Resources, you have options on what to do next. The most restrictive option: to completely block these types of sites (along with new domains and proxies) that may be used to circumvent blocks. Less restrictive options including proxying or tunneling the traffic to be able to further inspect or enable URL filtering.



Secure a discovered application

Verify Device Trust during SSO Authentication
Add this application to your Identity Provider and use Banyan to enforce Device Trust policies.

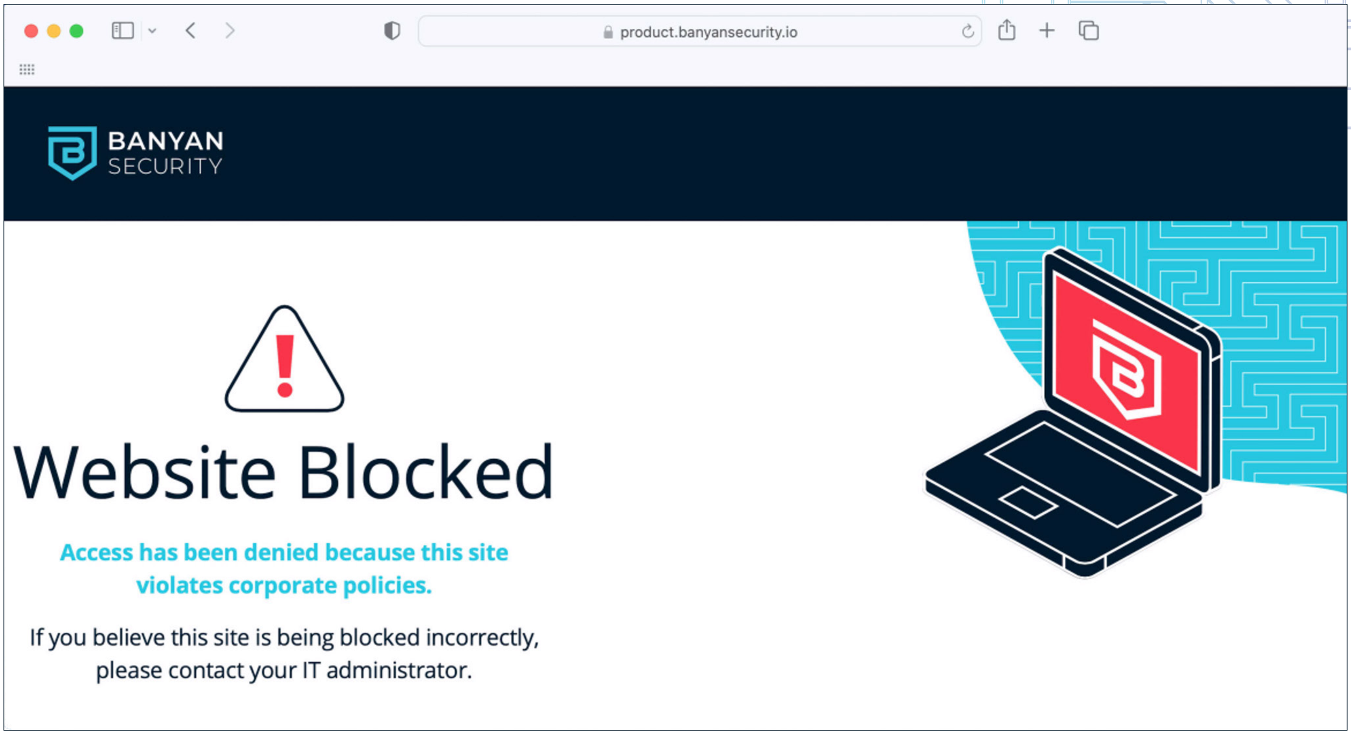
Restrict Access to Trusted IP Addresses
Tunnel traffic and IP whitelist so this application can only be accessed from trusted Banyan Edge IPs.

Limit Access to Specific URLs and APIs
Publish this application as a Hosted Website to apply URL- and API-level controls.

Apply Data Loss Prevention Policy
Enforce DLP policies to prevent users from exposing corporate data.

Block Application
Block users from accessing this application.

As you can see, the end user is made aware of why access was denied, and is not being blackholed, which may lead to a call to IT's Helpdesk and degraded productivity.

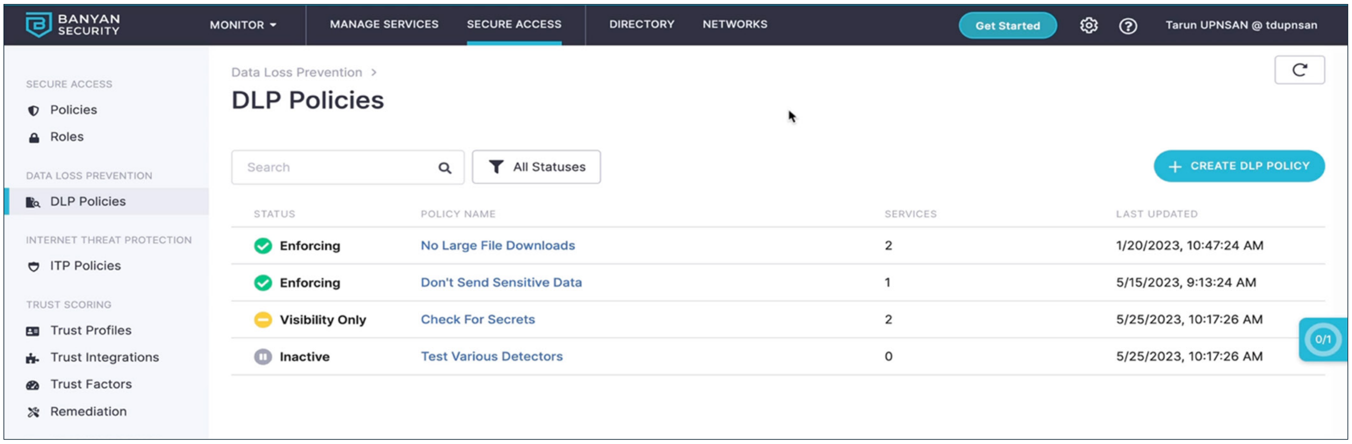


Website Blocked

Access has been denied because this site violates corporate policies.

If you believe this site is being blocked incorrectly, please contact your IT administrator.

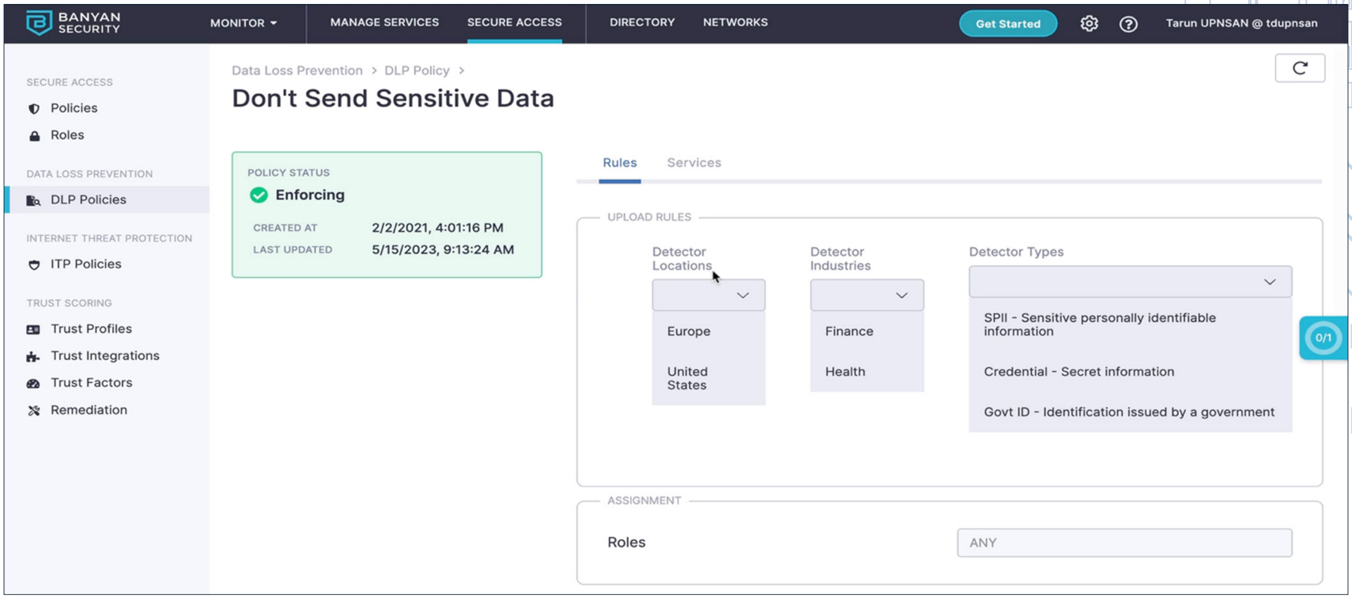
The administrator also has the option to apply a Data Loss Prevention (DLP) policy. The policies may include blocking downloads or restricting sensitive data uploads.



DLP Policies

STATUS	POLICY NAME	SERVICES	LAST UPDATED
Enforcing	No Large File Downloads	2	1/20/2023, 10:47:24 AM
Enforcing	Don't Send Sensitive Data	1	5/15/2023, 9:13:24 AM
Visibility Only	Check For Secrets	2	5/25/2023, 10:17:26 AM
Inactive	Test Various Detectors	0	5/25/2023, 10:17:26 AM

Sensitive data inspection is based on known patterns across multiple regions and countries.



Data Loss Prevention > DLP Policy >

Don't Send Sensitive Data

POLICY STATUS
Enforcing

CREATED AT 2/2/2021, 4:01:16 PM
LAST UPDATED 5/15/2023, 9:13:24 AM

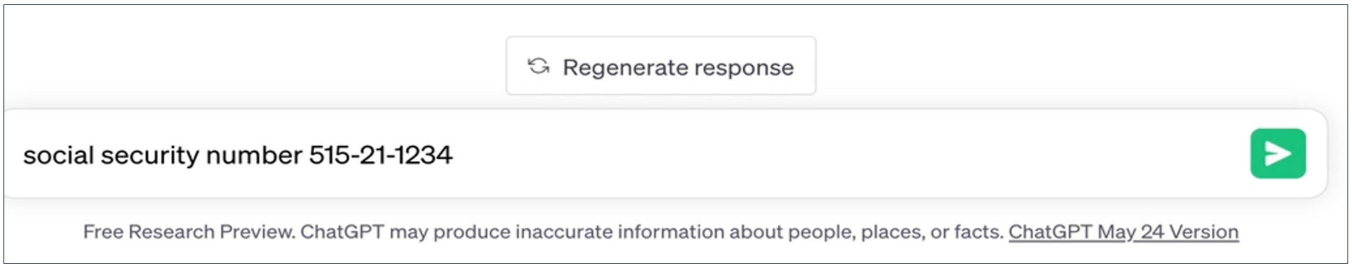
UPLOAD RULES

Detector Locations	Detector Industries	Detector Types
Europe	Finance	SPII - Sensitive personally identifiable information
United States	Health	Credential - Secret information
		Govt ID - Identification issued by a government

ASSIGNMENT

Roles: ANY

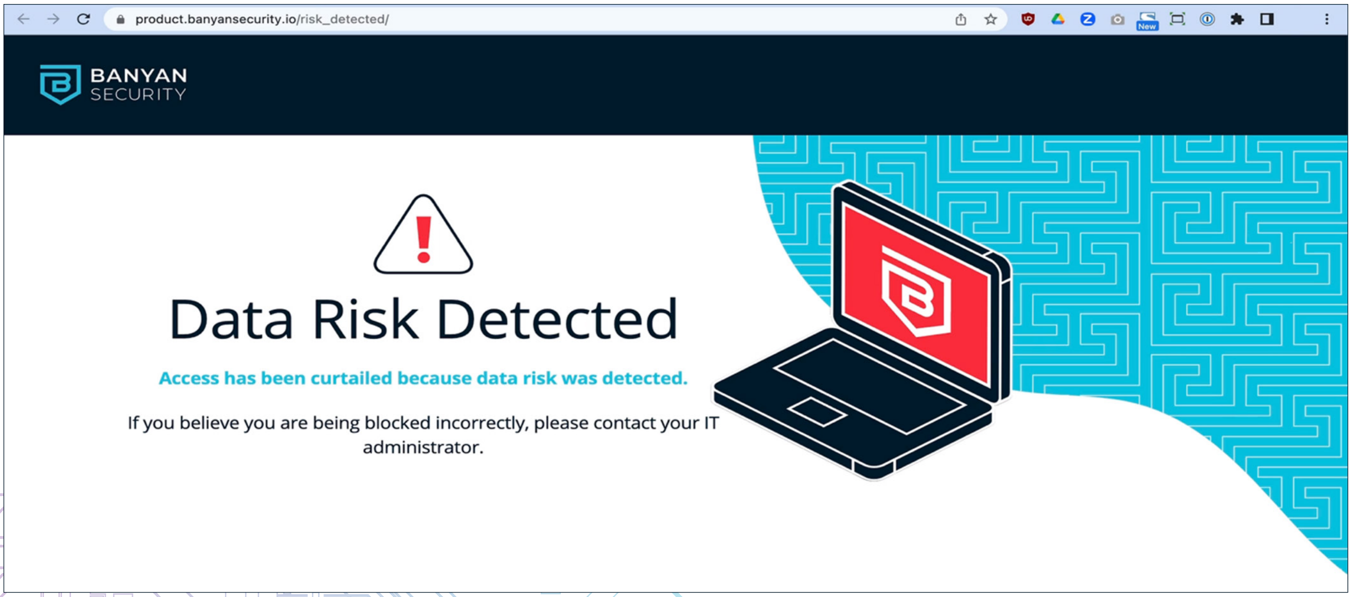
In this example, we try uploading a social security number to ChatGPT. All other non-sensitive information interactions with ChatGPT, and other AI tools, are allowed.



social security number 515-21-1234

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

End user is notified that the specific action is not allowed, and the interaction is blocked.



product.banyansecurity.io/risk_detected/

Data Risk Detected

Access has been curtailed because data risk was detected.

If you believe you are being blocked incorrectly, please contact your IT administrator.

SUMMARY

Generative AI introduces new cybersecurity threats by enabling the creation of highly sophisticated and realistic phishing attacks, capable of tricking even the most vigilant users.

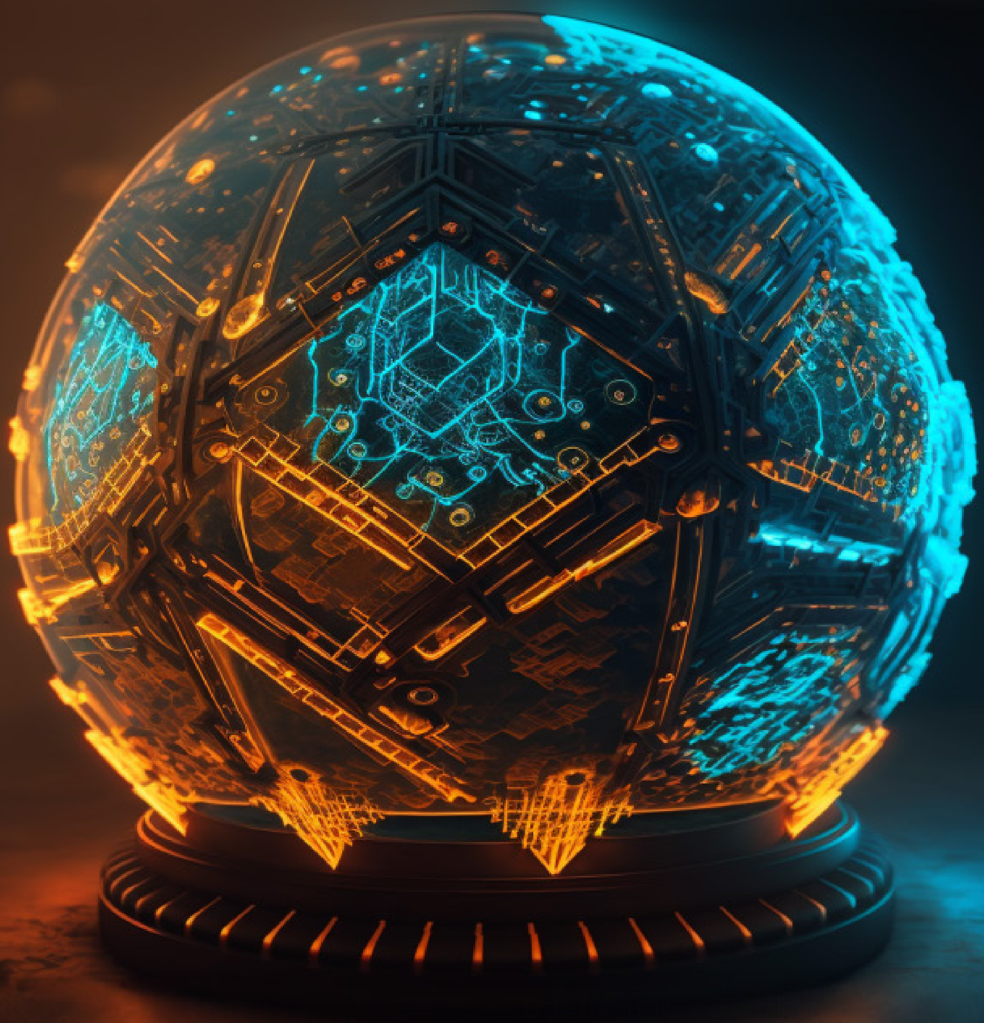
Additionally, malicious actors can leverage generative AI to automate the creation of advanced malware, making it harder for traditional security solutions to detect and mitigate these evolving threats. Employees are also leaking valuable corporate intellectual property in the hopes of getting work done quickly and easily. Effective security for AI must effectively address all of these facets.

In closing, focus on solutions that give the ability to block access to generative AI sites and tools effectively. By leveraging advanced web filtering capabilities and DLP inspection, SWGs like the Banyan SWG can detect and prevent users from accessing websites or tools specifically designed for generative AI. These solutions analyze and categorize web content based on predefined policies, allowing administrators to create rules that identify, then block sites related to generative AI. SWGs employ a combination of URL filtering, content inspection, and machine learning algorithms to accurately identify and categorize websites and tools associated with generative AI.

By blocking access to these resources, organizations can mitigate potential risks and prevent unauthorized or inappropriate use of generative AI technologies within their networks. SWGs provide a robust defense against potential security threats, ensuring that employees are unable to access generative AI sites or tools that may compromise data integrity, violate privacy regulations, or infringe upon intellectual property rights. In summary, SWGs offer an effective solution to block access to generative AI sites and tools, helping organizations maintain control and security over their network environments.

Learn more about ChatGPT security through Banyan SSE.

Schedule a Custom Demo Today



BANYAN
SECURITY