

Data Protection Considerations for Generative AI

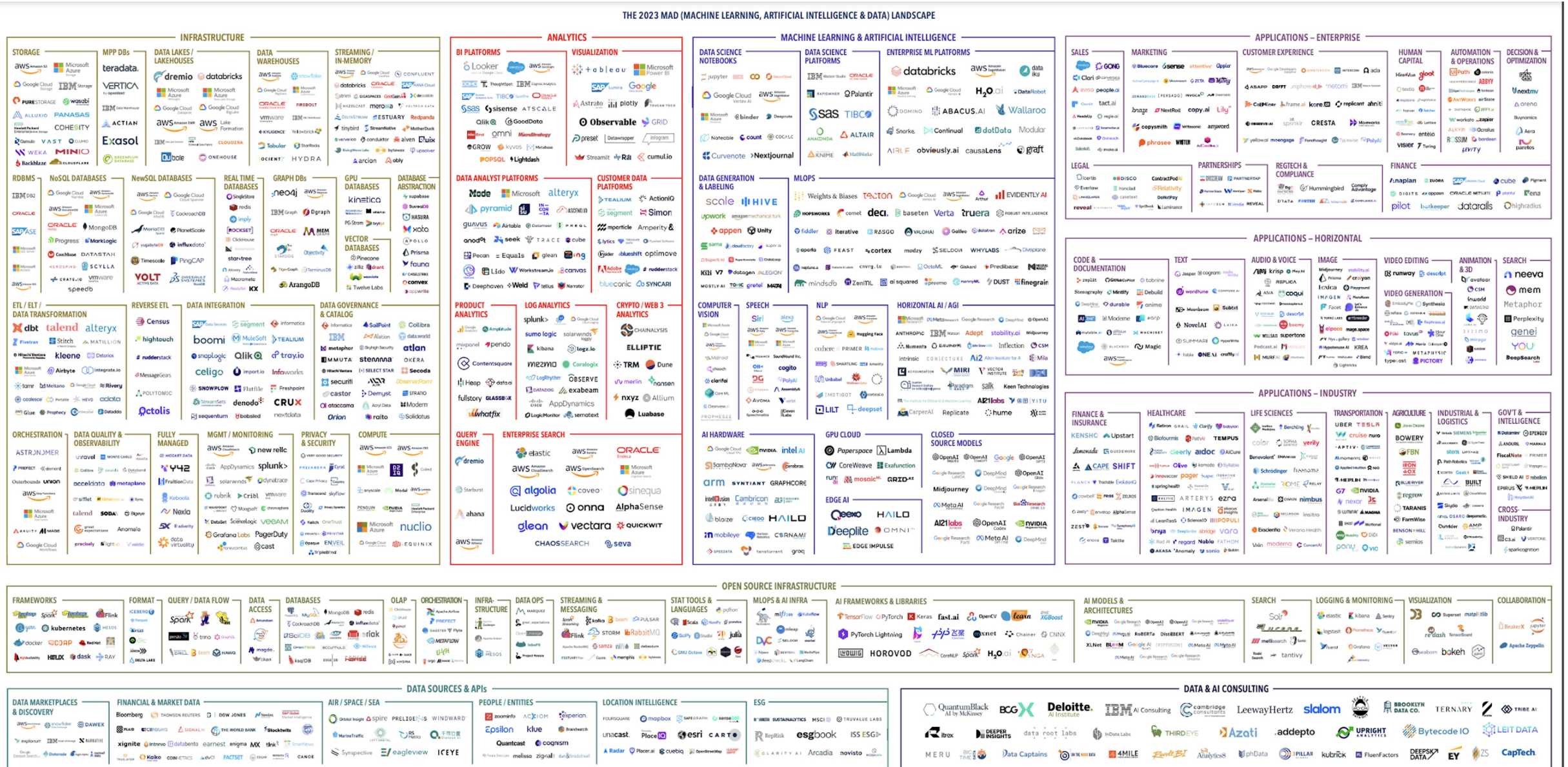
Julya Rebstock, CIPT
Information Governance Practice Manager

Aug 18, 2023

Agenda

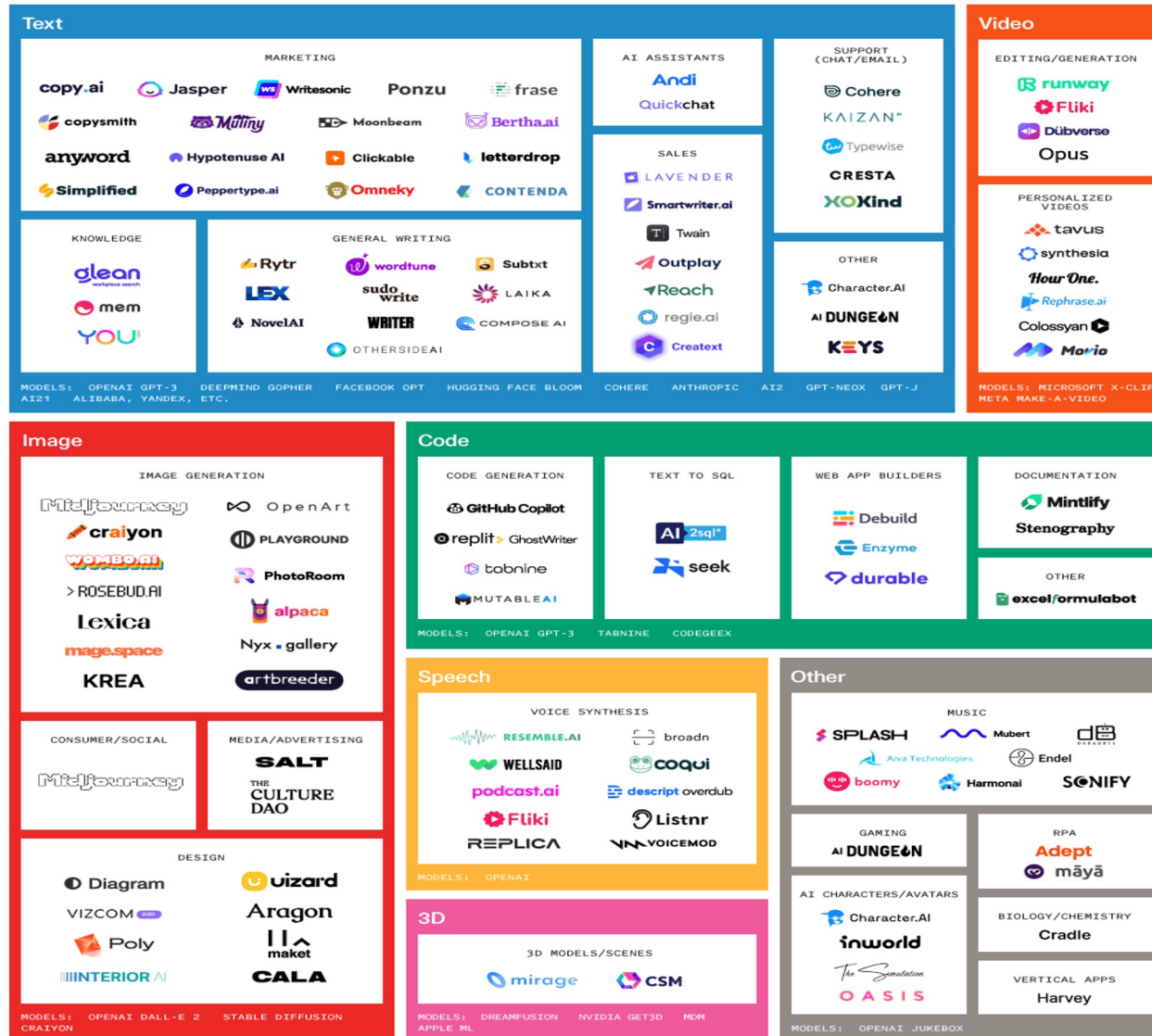
- What is Generative AI?
- What are the Risks?
 - Security, Privacy, Legal, Regulatory
- What are others doing?
- What can you do?
- What is Symantec doing?
- Q&A

<https://mad.firstmark.com/>



The Generative AI Application Landscape ^{v2}

A work in progress



What are you looking for?

Purpose built for...

- Text
- Video
- Images
- Code
- Speech
- 3D
- Other

AI Apps

Purpose built on...

- AWS AI services
- Google AI Platform
- Azure Machine Learning
- IBM Watson

AI as a Service

<https://research-assets.cbinsights.com/2022/11/14115447/SC-genAI-map.png>



There are now 13 generative AI unicorns

Generative AI startups with \$1B+ valuations (as of 05/08/2023)



<https://www.cbinsights.com/research/generative-ai-unicorns-valuations-revenues-headcount/>

We're hiring!

THERE'S AN AI FOR THAT

5,040 AIs for 1,422 tasks. Updated daily. Sponsored by [1-Click SEO Article Generation](#).

310,884 searches today

Find AIs using AI

⚡ The largest AI aggregator. Used by over 2,000,000 humans each month. ⚡

Sign in to save AIs

New Most saved Timeline Get featured Jobs Community Subscribe

Just launched

<https://theresanaiforthat.com/>

How Generative AI impacts your cybersecurity?

Exfiltration:

Users input sensitive data into Generative AI apps impacting privacy and data security

Infiltration:

Generative AI used by bad actors to design better attacks

- phishing email copy
- fake Gen AI portals
- malware design
- deep fakes

Symantec Security Center

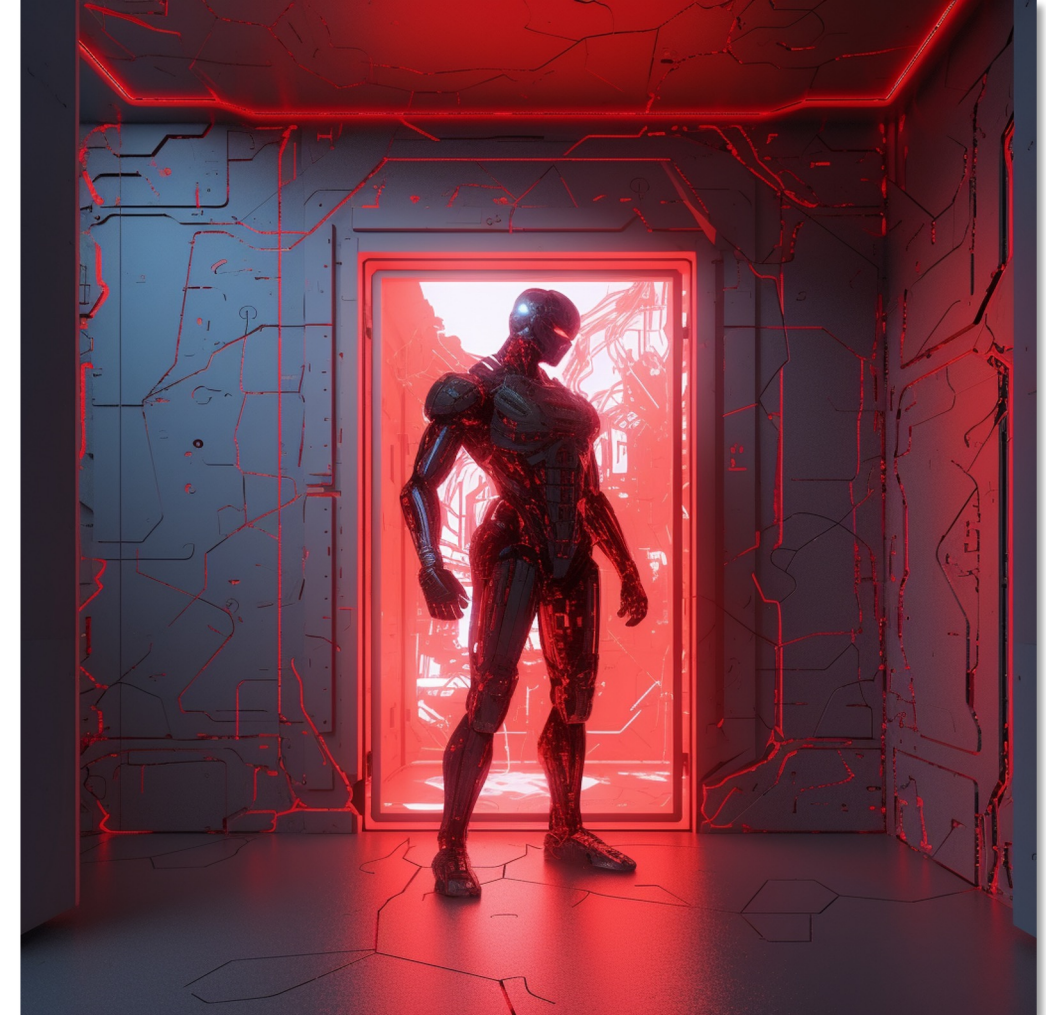
Stay ahead of tomorrow's threats and security incidents with the latest information from the global leader in cyber security.

Symantec Security Center

Risks of Artificial Intelligence in Cyber Security

- **AI-driven attacks**
 - Automated spear-phishing
 - Intelligent malware
 - Advanced persistent threats (APTs) and AI
- **Exploitation of AI systems**
 - Adversarial machine learning
 - Model poisoning attacks
 - Data privacy breaches
- **Generative AI misuse**
 - Creation of realistic digital forgeries
 - Privacy, Copyright, Lawsuits
 - Fraud and identity theft

There are many more...



How do I protect against data loss to generative AI? It depends...

The request:

“Build policies to protect my data from being consumed by ChatGPT:

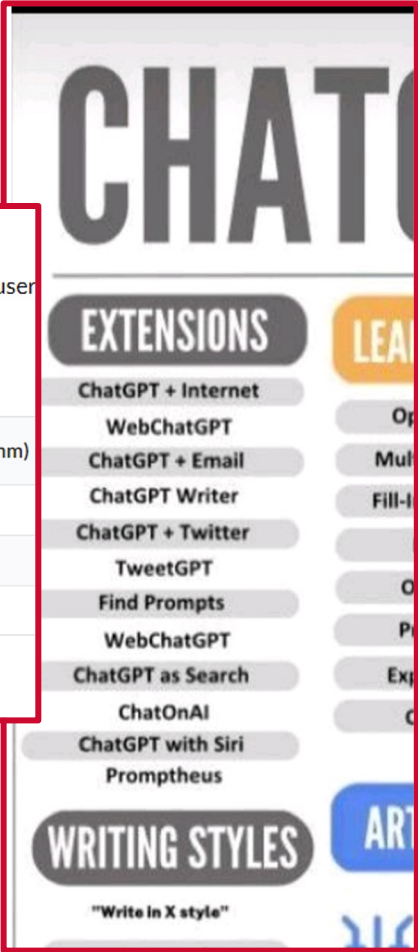
- Unstructured and narrative data*
- Code writing and debugging*
- General business letter writing*
- Strategic planning and writing”*

The Problem:

This assumes access to the sites and permission to share data outside of the customer’s control is an accepted **RISK**

Now things are not so simple...

ChatGPT Users	
ChatGPT set a record for the fastest app to reach 100 million users which took one year and two months to reach that milestone.	
ChatGPT Users December 2022 to February 2023 (mm)	
Date	Users (mm)
December 2022	1
January 2023	57
February 2023	100
Sources: CBS , Reuters	



<https://maxrascher.gumroad.com/l/cheatsheet>

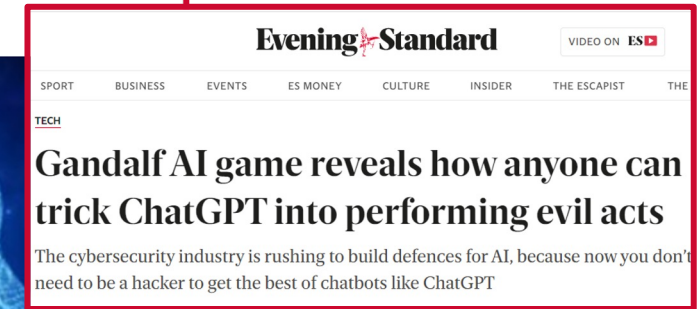
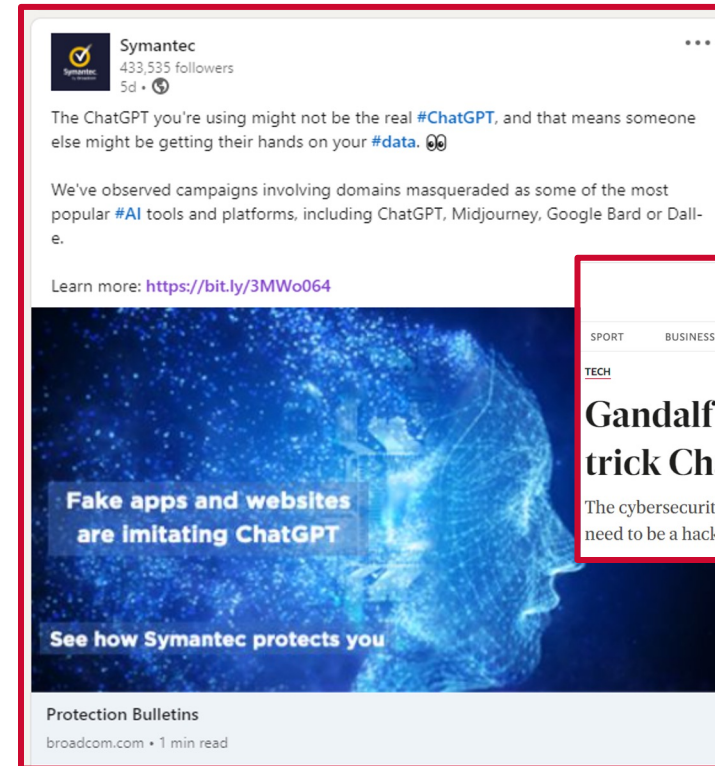
Consider these Questions:

- Is accessing these services allowed?
 - *Should* you allow access to the services?
- How will it be used and by whom?
- What are the potential risks?
 - Security
 - Privacy
 - Legal
 - Regulatory
- What is the value to my business?
- What can we do?



Generative AI: Security Risks

- Already being used to create:
 - Phishing campaigns
 - Key loggers
 - Malicious code, prompt engineering
 - “Deep fakes”
 - A virtual machine “inside” of ChatGPT
 - Undetectable steganography malware
 - Per FTC AI is “turbocharging fraud”
- [Broadcom Protection Bulletins](#)
- ChatGPT has already had more than one data breach (due to an open source vulnerability that revealed other users chat histories)



Group-IB Discovers 100K+ Compromised ChatGPT Accounts on Dark Web Marketplaces; Asia-Pacific region tops the list

Generative AI: Privacy Concerns

- Any data presented to it becomes its property and can be reviewed by staffers.
 - API submitted data not used to train the LLM but still retained 30 days
 - Public site data IS used to train the LLM, users have a specific Opt-Out form to complete if they don't want their data used
- Lack of Transparency
 - How was data collected?
 - How is it processed?
 - How are decisions made?
 - “Automated discrimination”
- How do you comply with “*right to be forgotten*” with a learning model?
- Data scraped from internet is regulated differently across privacy jurisdictions globally and may not meet “legal basis” required for collection, storage, and processing.
- AI used to process data is typically considered “3rd party processing”



Generative AI: Legal Concerns

- Do AI companies have the rights to the training data?
 - Liability risk for web scraping technologies
 - “Posting publicly is not implied consent”
<https://jolt.law.harvard.edu/assets/articlePDFs/v34/6.-Xiao-Bad-Bots-Regulating-the-Scraping-of-Public-Personal-Information.pdf>
- Is the output of generative AI new and original content, or a derivative of the data sets it trained on?
 - Can you guarantee the data returned wasn’t copyrighted?
 - How do you attribute source?
 - “Lawsuits are piling up from creative industries about copyrighted work co-opted or used by AI”
- CNBC
- Can you trust the information you are seeing?
 - AI “hallucinations” are an issue.
 - What if the output is misleading or dangerous?
 - First defamation notice issued to OpenAI
- Could be considered “3rd party software” but YOUR company may not have a license

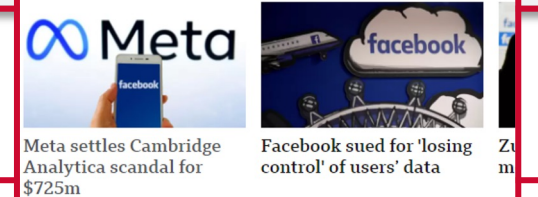
COPYRIGHT —

Stable Diffusion copyright lawsuits could be legal earthquake for AI

Experts say generative AI is in uncharted legal territory

TIMOTHY B. LEE - 4/3/2023, 5:45 AM

Facebook-Cambridge Analytica scandal



HOME > TECH

Clearview AI scraped 30 billion images from Facebook and other social media sites, says police. It puts everyone into a police line-up'

Katherine Tangelakis-Lippert Apr 2, 2023, 9:18 PM CDT

Simon Willison's Weblog

Lawyer cites fake cases invented by ChatGPT, judge is not amused

Legal Twitter is having [tremendous fun](#) right now reviewing the latest documents from the case [Mata v. Avianca, Inc. \(1:22-cv-01461\)](#). Here's a neat summary:

So, wait. They file a brief that cites cases fabricated by ChatGPT. The court asks them to file copies of the opinions. And then they go back to ChatGPT and ask it to write the opinions, and then they file them?

[Beth Wilensky](#), May 26 2023

Generative AI: Regulatory Landscape

Policy makers and regulators struggling to catch up:

- [EU Artificial Intelligence Act](#)
 - New API restriction drafted
 - Seeking “*binding rules for “high-risk” uses of AI in sectors like health care, transport, or criminal justice*”
 - Proposed criteria: safety, require human oversight (particularly for decisions), transparency in how data collected and processed, require traceability, enforce non-discrimination, right to file complaints and receive explanations of decisions
 - Proposed bans: biometric surveillance, emotion recognition, predictive policing, social scoring
- US Proposed: [Equitable AI Playbook](#)
- U.S. National Institute of Standards and Technology [AI Risk Management Framework](#)
- White House Office of Science & Technology Policy published “[Blueprint for an AI Bill of Rights](#)”
- Congressional: [Defense Auth Act for 2019](#) (Sec 238, 1051), 10 U.S.C. §2358; [National AI Initiative Act 2020](#); Advancing American AI Act; AI Training Act; [Federal Digital Platform Commission](#)
- Multiple Federal Agencies
 - FTC “watching intently”, “[algorithmic decision making](#)” proposed rule

AI Warning: ChatGPT Blocked for Data Laws Breach

The future development and use of AI systems is under fire as an EU regulator blocks ChatGPT for GDPR breaches.

By Rafi Azim-Khan

Generative AI: Integrations

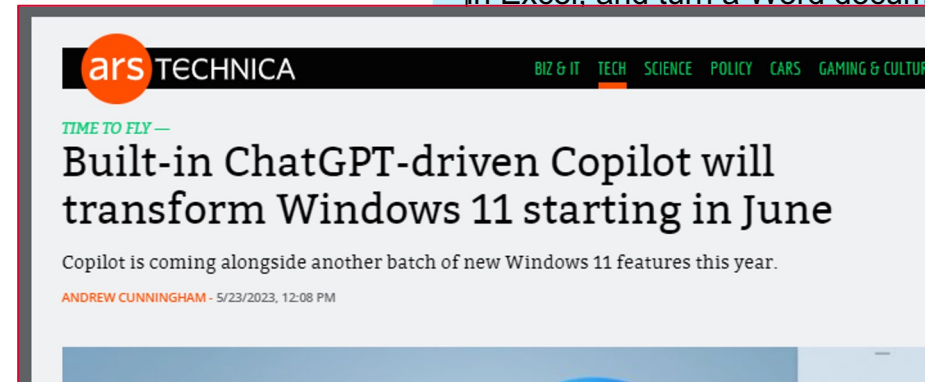
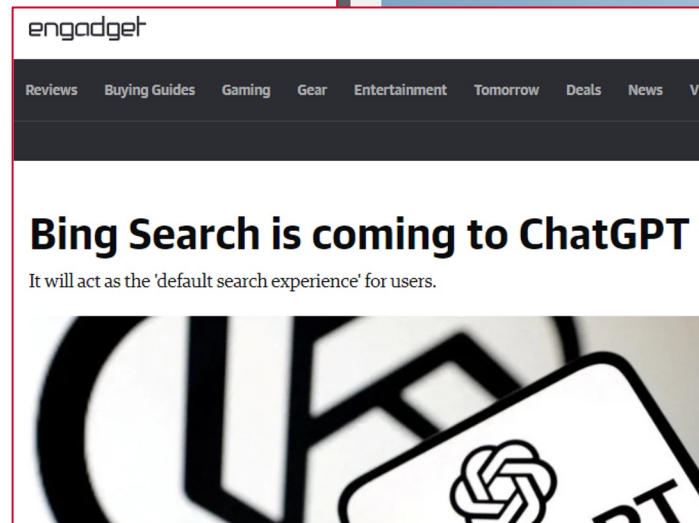
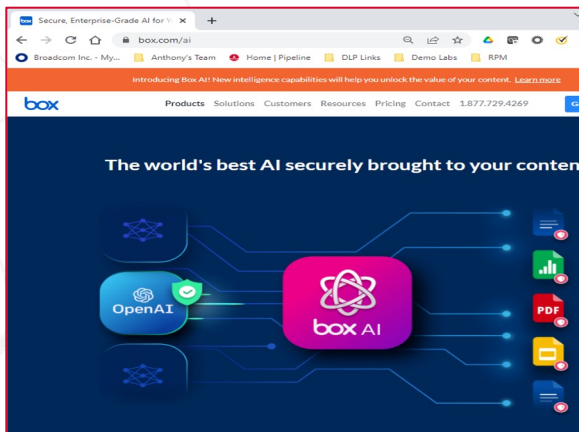
Microsoft

- Microsoft announced integration into **Windows 11 and O365** for “co-pilot which will help edit, summarize, create, and compare documents.”
- Google adding AI to Google Cloud, MakerSuite, and Workspace.
- LinkedIn, Slack, Teams already using ChatGPT
- Box integrated ChatGPT

“With the new features, users will be able to transcribe meeting notes during a Skype call, summarize long email threads to quickly draft suggested replies, request to create a specific chart in Excel, and turn a Word document into a presentation.”

Concept called Copilot essentially rides along with users to understand their data. The Copilot can read email and on their documents and presentations they’ve made during meetings, and on the Teams platform, users can then ask Copilot to write a status report on documents across the organization and then draft an email with an update.” -

[microsoft-365/index.html](https://www.microsoft-365/index.html)



OpenAI - “One stop shop for the world’s data”

As of: Aug 16, 2023 - <https://openai.com/blog/chatgpt-plugins>

- Microsoft Windows 11 OS, O365 “Copilot”, Bing & Edge
- Box
- LinkedIn
- Salesforce
- Slack “Einstein”
- Snapchat
- Koo (Twitter rival)
- Quizlet
- Instacart
- Shopify
- iOS app
- + Plugins: [web browser](#), [code interpreter](#), [retrieval](#)
- Duolingo
- Expedia
- FiscalNote
- Ghost
- KAYAK
- Klarna Shopping
- Milo Family
- OpenTable
- Speak
- Wolfram
- Zapier (interacts w/>5k apps)

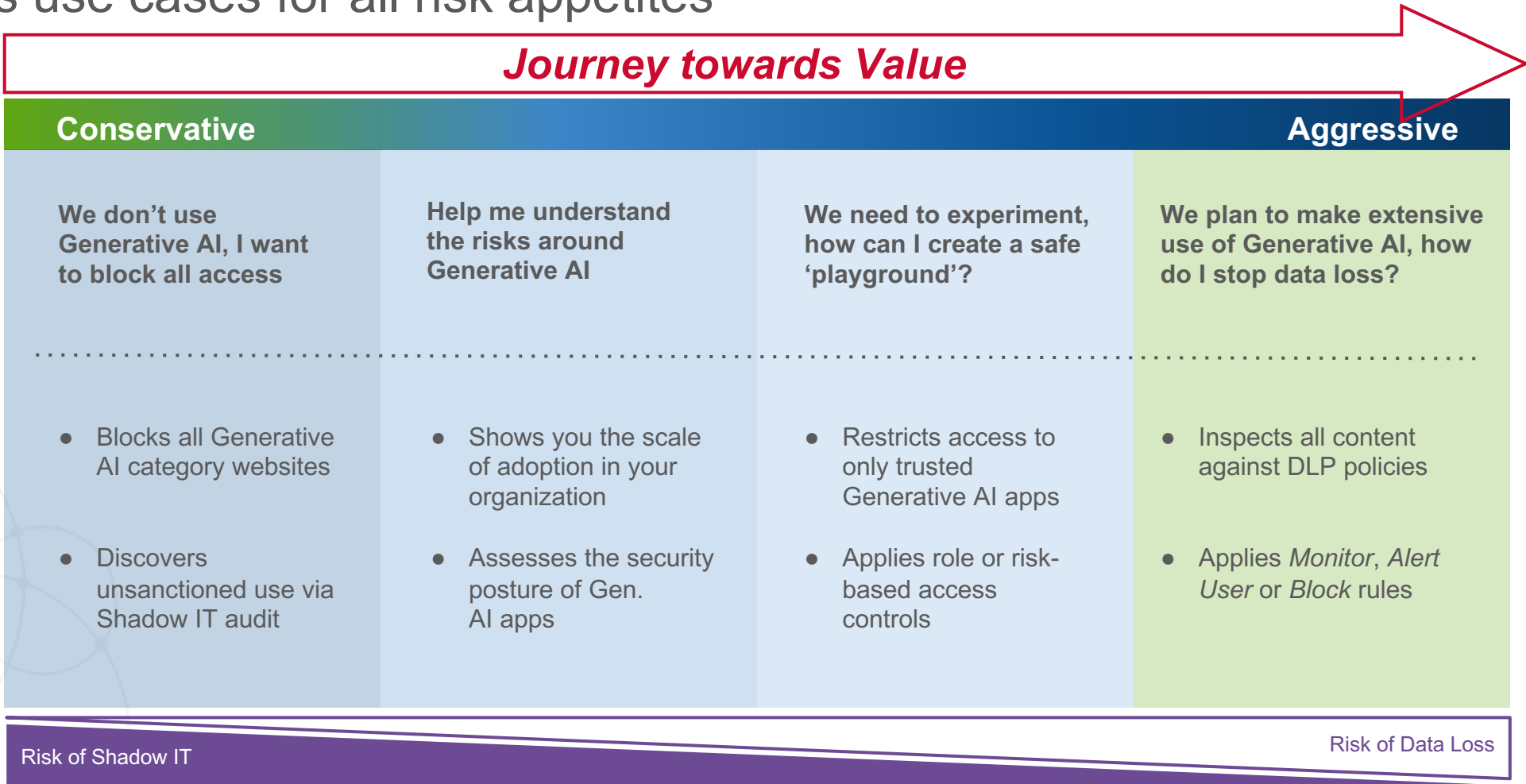


Generative AI Protection can help everyone

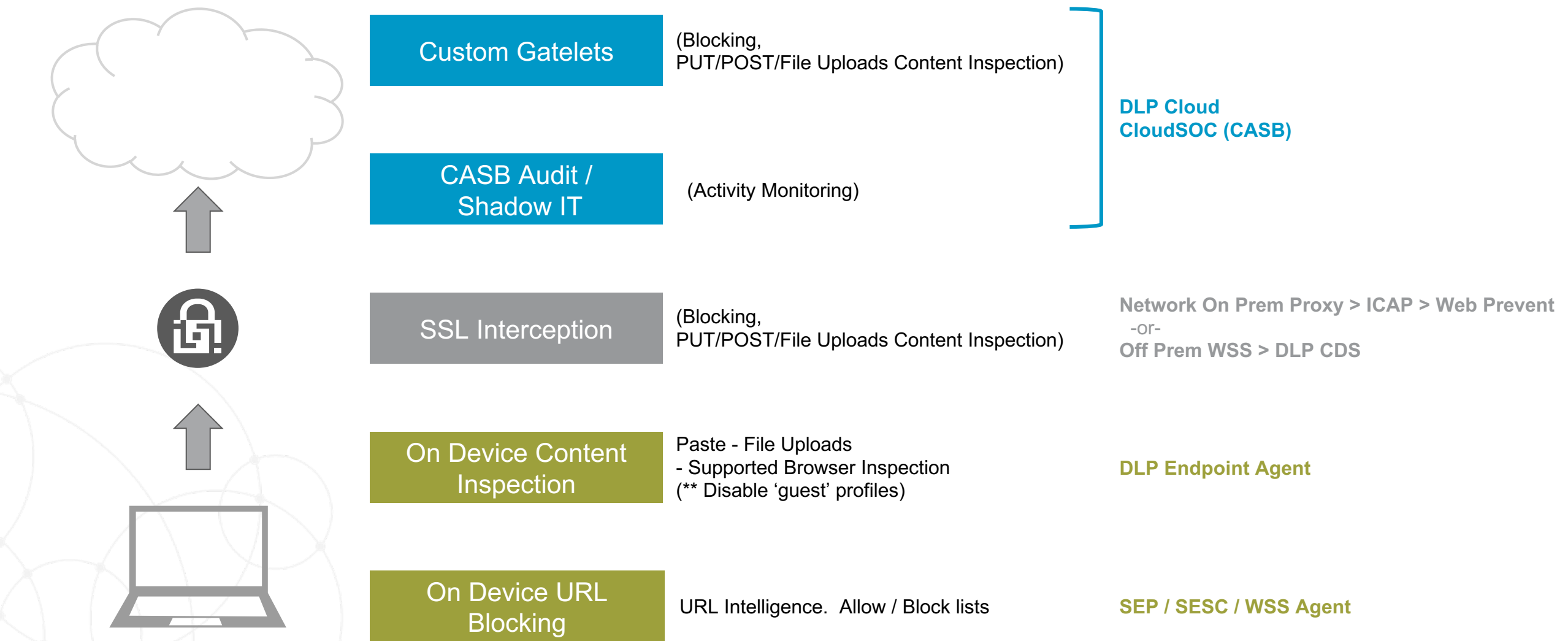
Addresses use cases for all risk appetites

Company approach to using Generative AI

What Symantec Generative AI Protection does



Layered Data Inspection and Protection



Use cases and Solution – DLP Endpoint

Full visibility in Generative AI ChatBots usage (e.g. chatGPT)

- DLP endpoint **Clipboard Copy Paste** channels
- DLP endpoint **Chrome, Firefox, Safari, Edge** channels
(Disable guest profiles!)
- Response Actions: Block, Notify, User cancel
- Incident contains: Source & Destination URL, User, Data Pasted

Protecting against copy-paste structured content (e.g. copy-paste-upload credit card, SSN)

Detection Policy Options:

- DCM - Regex, DI
- EMDI - exact data validator

Protecting against unstructured content (e.g. copy-paste-upload design document)

Detection Policy Options:

- Use **MS MIP** classification for File Uploads
- DCM - keywords
- Layer DI & EMDI if structured data is commonly found
- IDM for strategic docs

User education on Generative AI ChatBots usage policy

- DLP Popup (user cancel, notify, block)
- Popup text can be customized based on channel type

Use cases and Solution – DLP Web Prevent/DLP CDS

Full visibility in Generative AI ChatBots usage (e.g. chatGPT)

- “Content Generation” category available to **ProxySG w/BCIS**
- Proxy intercept via ICAP to DLP Web Prevent -or- **WSS** intercept to DLP CDS
- Response Actions: Block, Notify
- Incident contains: Source & Destination URL, User, Data Posted

Protecting against copy-paste structured content (e.g. POST/upload credit card, SSN)

Detection Policy Options:

- DCM technology - Regex, DI, Structured DI
- EDM/EMDI - indexed exact match
- Form Recognition

Protecting against unstructured content (e.g. POST/upload design document)

Detection Policy Options:

- Use **MS MIP** classification
- DCM (keywords, patterns)
- IDM - indexed documents
- OCR - forms and images

User education on Generative AI ChatBots usage policy

- DLP Notify options

CASB Audit Category *Content Generation* for Generative AI Apps

What

- Discover generative AI application that my organization is using

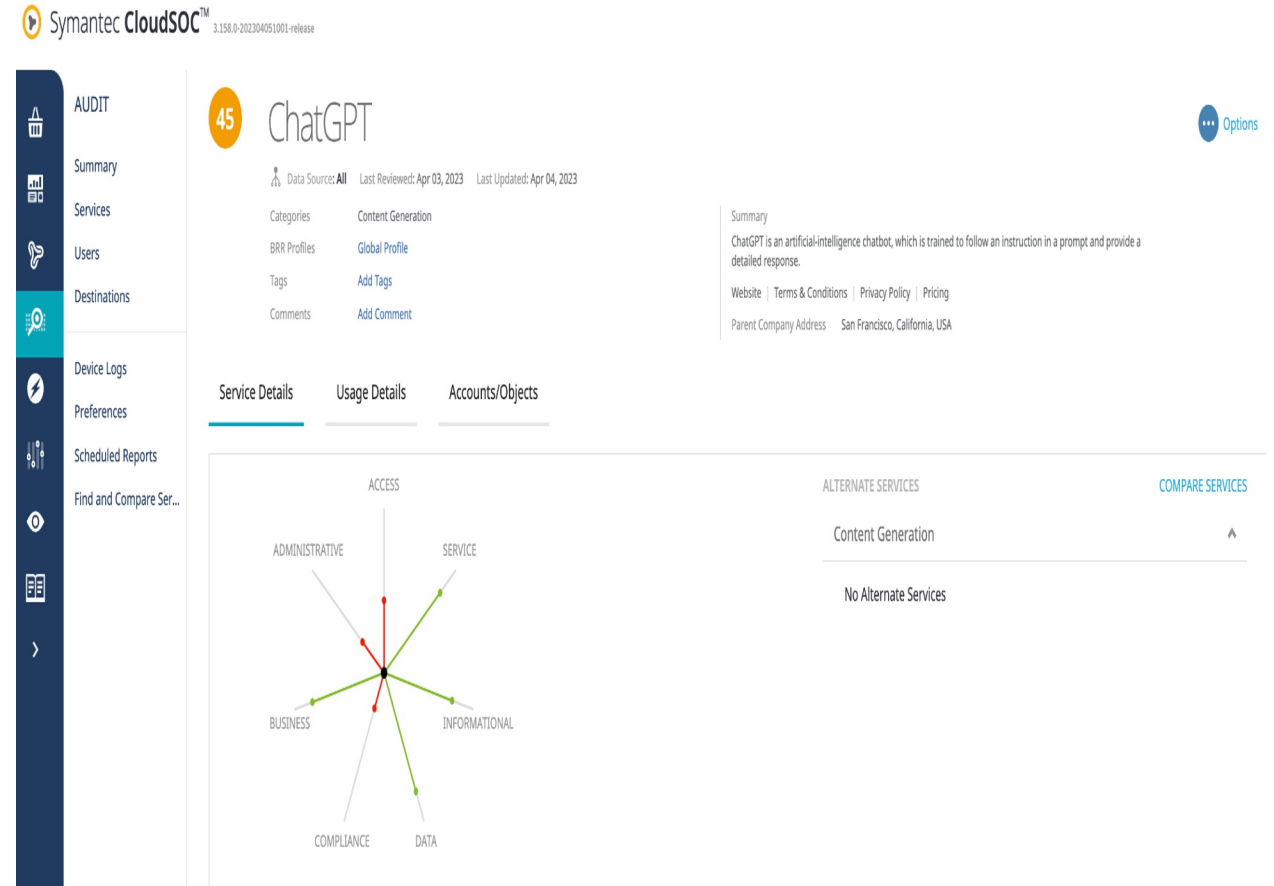
How

- Leverage CASB Audit's new category to identify Generative AI apps using "Content Generation"

Benefits

- Ability to identify users using Generative AI
- Ability to use this CASB Audit category within Edge- and Cloud SWG to block or gradually allow sanctioned apps
- Ability to generate Custom Gatelets to drive secure adoption

Secure Adoption of Generative AI Apps



How do I protect against data loss to generative AI? The method...

Build DLP policies to protect:

- Unstructured or narrative data
 - *Pull common phrases and keywords used*
 - *Layer EDM/EMDI and/or Structured Data Identifiers if structured data is also commonly found*
 - *Known strategic documents: IDM, OCR, Form Recognition, custom file types*
- Code writing and debugging
 - *Create IDMs of code pieces, avoiding open source or non-proprietary segments*
 - *VML*
 - *Source Code template policy*
 - *Future: Source Code classifier*
- Imaging and Forms
 - *OCR for images with text extraction*
 - *IDM for Exact Match of images*
 - *Form recognition and matching*

Employees: Communicate Communicate Communicate!



- Information Governance, Data Governance, Privacy Governance!
 - Resource: [IAPP Artificial Intelligence Center](#)
 - Report: Privacy and AI Governance by IAPP – Leveraging Privacy Governance as a foundation for AI Governance https://iapp.org/media/pdf/resource_center/privacy_ai_governance_report.pdf



- Corporate Policies to review and link:
 - Data Classification
 - Acceptable Use
 - Information Security
 - Data Retention, Data Handling
 - HIPAA Policies and Procedures

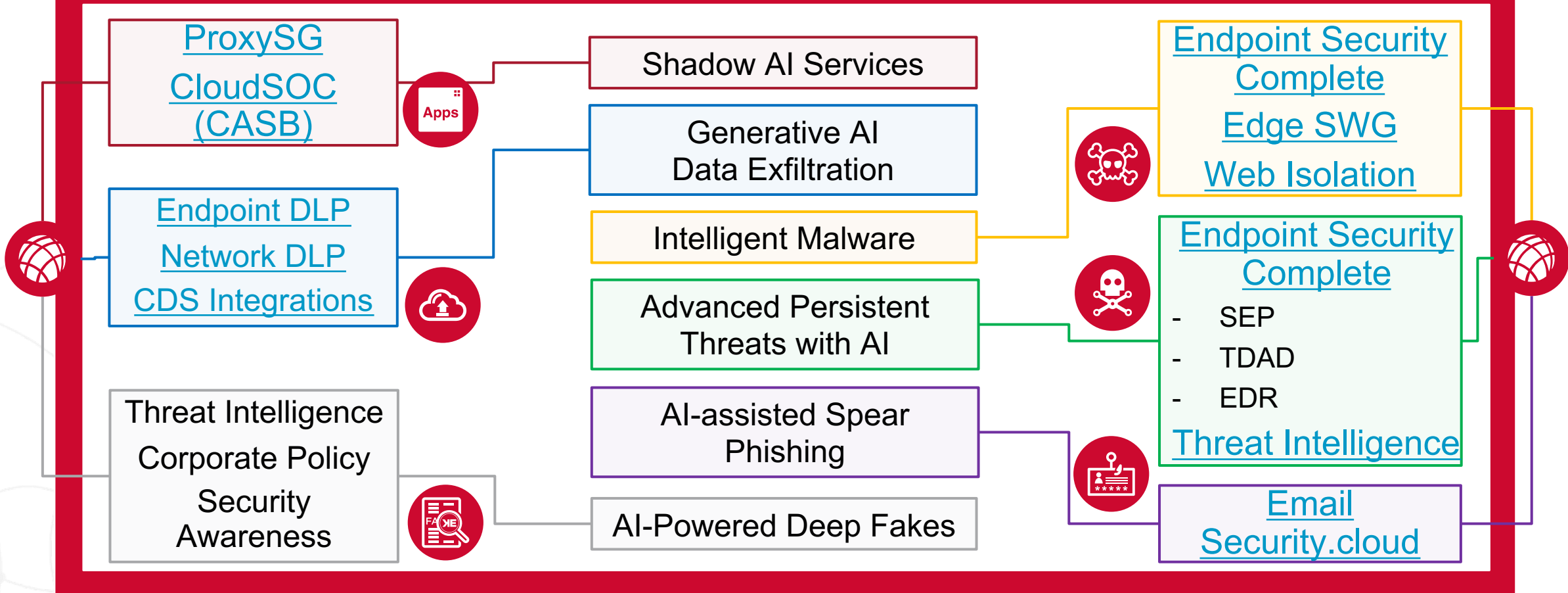


- Enforce References/Citations if AI is used
- Do not use AI as the only decision point - “yet” - not defensible
- Remind that generated products may not be eligible for copyright/patent protection

Symantec Protecting Against AI-based Threats



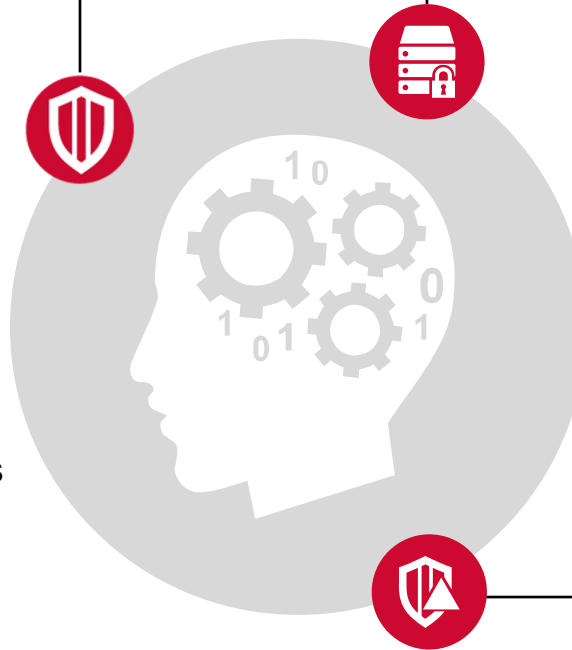
Symantec Global Intelligence Network



Endpoint Security – AI Technologies

Endpoint Security

- Advanced Machine Learning (AML) used to determine if a file is good or bad
 - Static Data Scanner engine, Download Insight, SONAR, virus and spyware scans
- Symantec Insight (Deep Learning) determines the risk level of files and websites
- Antivirus engine (STARGate) uses machine learning, deep learning, NLP
 - Leveraged by 21 products and services
- SONAR monitors application behaviors
 - 1.2B applications, 1,400 behavioral attributes
- Adaptive Protection uses AML to generate and recommend adaptations
- Intensive Protection uses AML to provide aggressive detection capabilities



Endpoint Detection and Response

- Cynic uses machine learning to examine files in a cloud-based sandbox environment
- Targeted Attack Analytics uses Threat Hunter analytics, which combines local and global telemetry and supervised machine learning
- Criterion machine learning engine detects files in the gray region between known good and bad
- Sapient machine learning engine detects malware based on static attributes

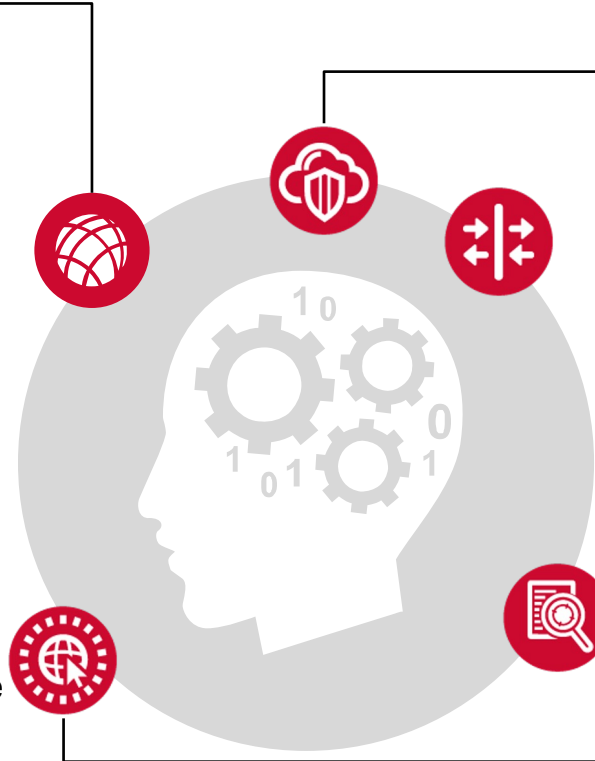
Threat Defense for Active Directory

- TDAD uses AI and Natural Language Processing to create decoy environments

Network Security – AI Technologies

Global Intelligence Network

- Includes supervised and unsupervised machine learning, deep learning, neural networks, and natural language processing
- Uses different machine learning models to create intelligence from vast amounts of data:
 - 175M+ endpoints
 - 80M+ web proxies
 - 160M+ email accounts
 - 126M+ attack sensors
 - 240K+ network sensors in 157 countries
- Machine learning models/services such as Insight, SONAR, STARGate, Cynic, etc. reside here
- Analyzes the reputation of files, URLs, IPs, and other network resources through threat intelligence and user-generated feedback
- Analyzes historical data to provide predictive analytics based on patterns and trends
- What analysts did two weeks now takes minutes



Cloud SWG / Edge SWG

- Leverages intelligence from GIN
- Content analysis uses malware scanning, predictive analysis, and sandbox analysis from Endpoint Protection AV, which uses AML
- Isolation functionality leverages the same machine-learning technology used in Endpoint Protection AV
- Dynamic categorization service for real-time categorization of URLs for filtering
- Predictive Analysis reports

Web Isolation

- Symantec Cynic examines files in a cloud-based sandbox environment and uses machine learning to compare results to known bad results
- Uses Symantec AV as an anti-malware sanitizer

Information Security – AI Technologies

CloudSOC

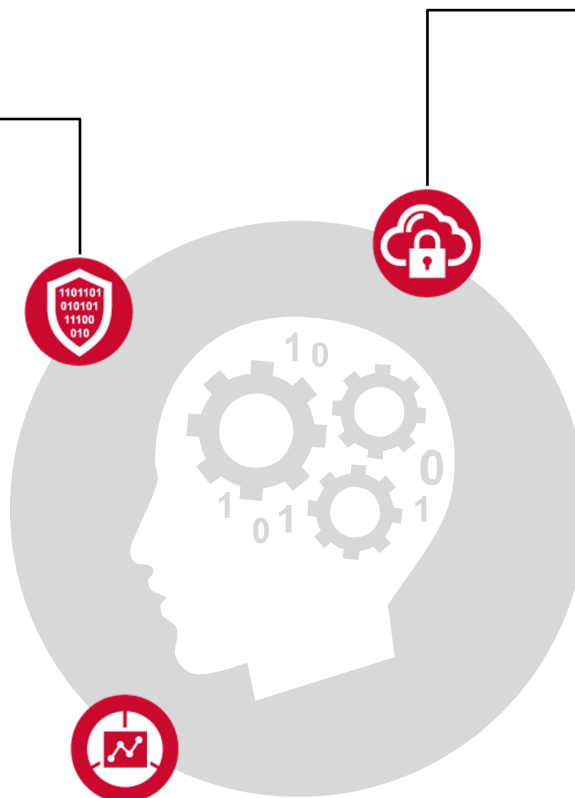
Data Loss Prevention

- Vector Machine Learning for statistical analysis of unstructured data
- Industry's first machine learning technology designed to simplify the decision of hard-to-find intellectual property
- Machine learning techniques used to rapidly scan and score data-in-motion, including email and text messages, documents, and associated attachments
- Multi-lingual Natural Language Processing

Information Centric Analytics

- Uses supervised and unsupervised machine learning to analyze user behavior
- Machine learning used for continuous analysis and peer-based user risk across platforms

- Maintains intelligence on 1000's of applications through supervised machine learning
- Audit uses machine learning for in-depth analysis of a service's strengths and weaknesses and identifies probable app data breaches
- Securlets use machine learning and data science to analyze real-time user traffic
- Detect uses machine learning to build behavioral models



Email Security – AI Technologies

Email Security.cloud

- Skeptic - Heuristic scanning technology and machine learning to block previously unseen threats, potential spam, and zero-day threats
- Cynic sends copies of files of interest to a secure sandbox and mimics end-user behavior to trigger suspected malware and correlates data with the Global Intelligence Network
- Synapse correlates events recorded by Endpoint Security, EDR, and Email Security.cloud
- Email threat campaign correlations are determined using machine learning



GCP is Ready to Expand Broadcom's AI Capabilities

Debanja Saha, vice president and general manager for data analytics for Google, said in the longer term, GCP also provides the level of scale that will be required to infuse additional artificial intelligence (AI) capabilities, based on machine learning algorithms, into the Broadcom portfolio.

Protecting from AI Misuse: Tools and Strategies

Policy and Training – Back to the Basics

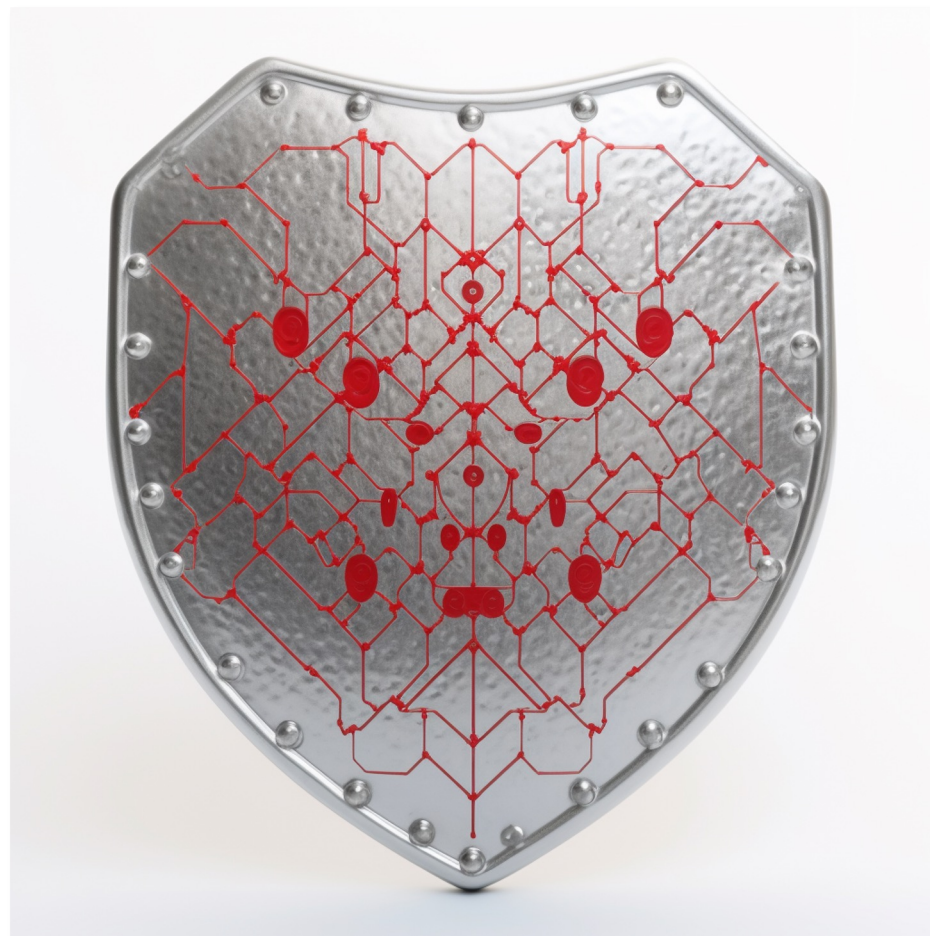
- Clear AI Use policy (and keep it updated)
- Clear Cyber Security Use policy (keep it updated too)
- Regular employee training

Access Controls and Data Protection

Access Controls	CASB
Network Segmentation	Data Loss Prevention
Data Encryption	Web Proxy
MFA, Biometrics, Risk-Based	Web Isolation

Security Measures and Maintenance

Secure Development Practices (if relevant)	Threat Detection & Response
Update & Patch Software	Threat Intelligence
Endpoint Security	Regular Threat Assessments



OPENAI / JOBS

Killswitch Engineer

San Francisco, California, United States

\$300,000-\$500,000 per year



About the Role

Listen, we just need someone to stand by the servers all day and unplug them if this thing turns on us. You'll receive extensive training on "the code word" which we will shout if GPT goes off the deep end and starts overthrowing countries.

We expect you to:

- Be patient.
- Know how to unplug things. Bonus points if you can throw a bucket of water on the servers, too. Just in case.
- Be excited about OpenAI's approach to research

Next Steps...



Use the resources from this presentation and the Appendix



Keep up to date on Symantec news:

[Symantec Community](#)

[Symantec Enterprise Blogs](#)

[Symantec eLibrary Courses](#)

[How to access eLibrary](#)



Register for our next AI Focused webinar

[Demonstration: Symantec Generative AI Protection](#)

When: July 11 @ 11am CDT

Presenter: Mike Winslow



Contact your Account SE to dig deeper on any product or feature.

Thank you



APPENDIX



Symantec protects against AI / ML generated threats...

- On the Device | [Symantec Endpoint Security Complete](#)
- Over the Network | [Web Protection Suite](#)
- In the Cloud | [DLP Cloud \(CASB\)](#)
- Over Email | [Email Security Complete](#)

To Get Details of a CloudSOC Category

AUDIT

Summary

Services

Users

Destinations

Device Logs

Preferences

Scheduled Reports

Find and Compare Ser...

Symantec CloudSOC™

Find and Compare Services

Service Availability: Configure

☒ Include partially researched services

category: content

CATEGORY: Content Distribution

CATEGORY: Content Generation

CATEGORY: Content Discovery

☐ Online Shopping

229

☐ Point of Sale

229

☐ API Provider

228

☐ Managed Hosting

222

☐ Business Process Man...

217

☐ Online Bookings

217

☐ Block@NWDevice

4

☐ Ignore

3

☐ Untagged Servi...

36906

☐ UnsanctionedD...

3

☐ Engineering

3

☐ Exclude From BRR

2

☐ DontCareMFA

2

☐ AWS

1

☐ DisruptMFA

1

Save Filters

Load

Risk

☐ High

10976

☐ Medium

25543

☐ Low

425

☐ Unresearched

4

Add Additional Filters

Select all 27 services

Actions

Rating

Name

☐

40

Unbounce

Content Generation

☐

60

Grammarly

Content Generation

☐

Pivotal Tracker

Gatelet

Securlet

Actions

x

x

✓

x

Step 1: Log into CloudSOC
Go to Find&Compare

Step 2: Expand Filters

Step 3: Search for the following:
Category: Content Generation

Step 4: check the # of services in the BRR library
Updated every Friday (but may not have NEW apps in this category)
Depends on Research Completed by the Team

To ADD a Service to CloudSOC

 Symantec CloudSOC™
3.159.0-202305231954-patch



AUDIT



Summary



Services



Users



Destinations



Device Logs



Preferences



Scheduled Reports



Find and Compare Ser...



>

Find and Compare Services

Service Visibility: Configure

☒ Include partially researched services

☐ Select all 36,949 services

Rating	Name
☐ 67	Workday ERP
☐ 40	Bitcasa File Sharing, Storage
☐ 45	Pivotal Tracker Project Management
☐ 82	G Suite Online Productivity Suite

Recommend to Add Service

Note: Please make your service request below. The request will be reviewed and considered for addition to the list of services.

Service Name *

Service URL *

Comments

400 characters left

Priority: ☐ High ☐ Medium ☒ Low

☒ Receive Email Notifications for Tracking

Submit Recommendation

Filters

Options

Create Custom App

Recommend Addition

Requested Services

	Gatelet	Securlet	Actions
	✓	✓	▼
	✗	✗	▼
	✗	✗	▼
	✓	✗	▼

35 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

 Symantec
by Broadcom