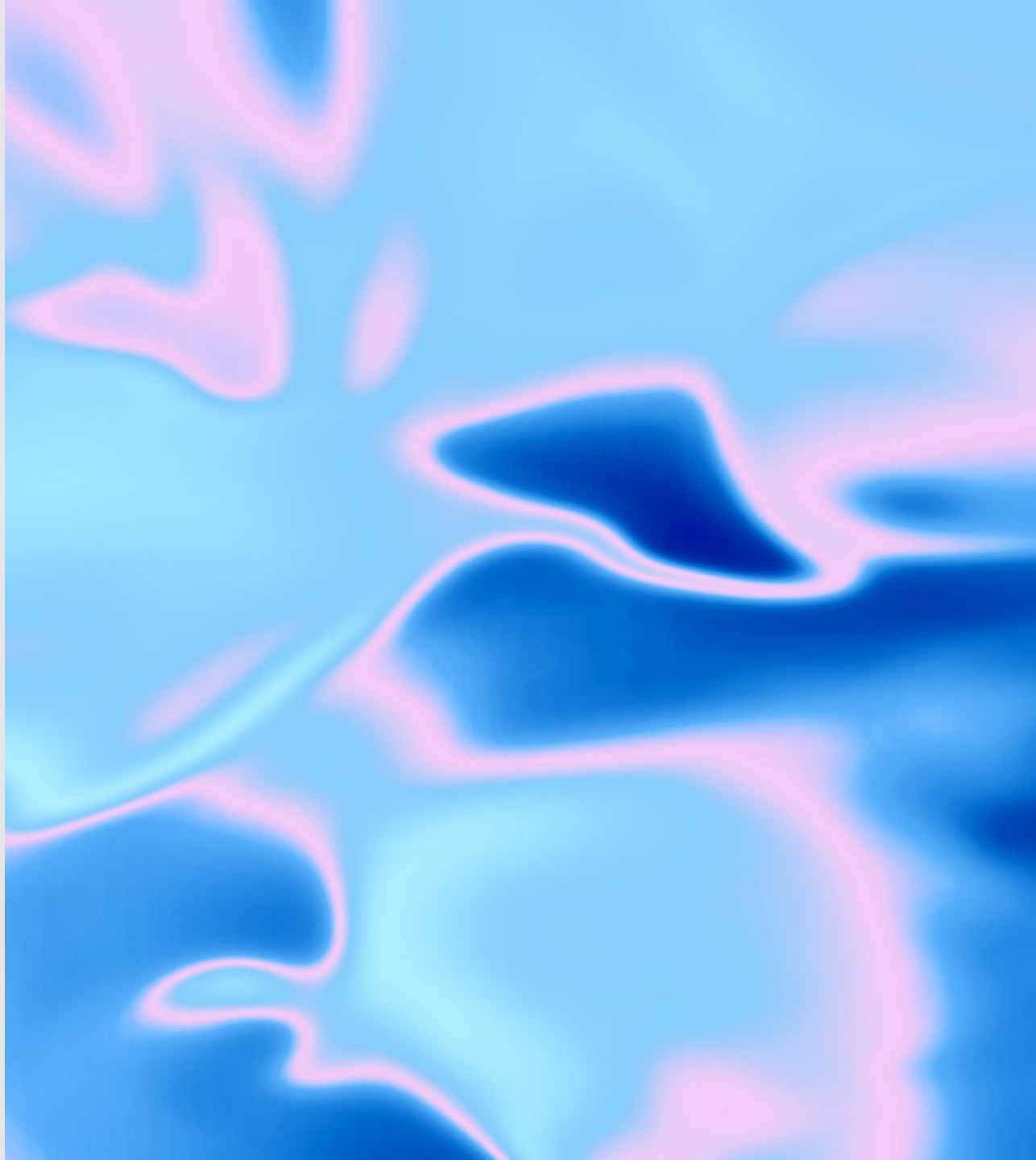
Microsoft Security

Prevent data leak to Shadow AI

Microsoft deployment blueprint



Problem summary

Shadow AI refers to the use of artificial intelligence tools - especially generative AI - by employees without the knowledge, approval, or governance of their organization's IT or security teams.

A story of data leakage via AI

Employee accessed sensitive data and inadvertently caused data breaches

Alex and Adele

Project Managers at Contoso

Contoso deployed Microsoft 365 Copilot (Copilot) to enhance employees' productivity	Alex and Adele are working on a confidential project called Project Obsidian, where all files are labeled as Confidential.	Alex asks Copilot to summarize Project Obsidian files.	Copilot provides him with a summary and automatically labels it as Confidential.
Adele wants to use ChatGPT instead	Adele pastes the content of the Project Obsidian files into ChatGPT and asked for a summary	The project details were prematurely revealed, resulting in data breaches.	Consequently, Contoso banned all consumer AI apps in the workplace

Impact

The information about Project Obsidian leaked to the public, resulting in bad PR for Contoso and significantly impacting Contoso's share price. Employees continue to find ways to use consumer AI via unmanaged devices and networks, leading to more hidden security incidents.

How can I protect my organization from Shadow AI?

Prevent data leak to shadow AI with Microsoft

1. Defender for Cloud Apps
2. Entra
3. Intune
4. Purview

Discover

Discover AI apps

Activities

- Discover the use of AI apps[1]
- Discover user interactions with AI apps[4]

Outcomes

Unknown AI app usage is discovered

Block access

Block user access to unsanctioned AI apps

Activities

- Block access to unsanctioned AI app for the org[1]
- Allow limited access to certain AI apps[1], and restrict specific users and groups from access[2] or block elevated risk user from access[2, 4]
- Block install of the AI app on devices[3]

Outcomes

Unsanctioned AI apps are blocked

Secure data[4]

Block sensitive data going to sanctioned AI apps

Activities

- Block pasting and uploading of sensitive info
- Block elevated risk users from submitting prompts in Microsoft Edge
- Block sensitive info from being sent in Microsoft Edge
- Block sensitive data shared through network service providers

Outcomes

Sensitive data is not sent to sanctioned AI apps

Govern data[4]

Govern data sent to AI app in Microsoft Edge

Activities

- Audit interactions
- Detect inappropriate behaviors in prompts
- Retain/delete prompts
- Investigate prompts

Outcomes

Prompts sent to AI apps in Microsoft Edge is retained for future investigations

Last updated: September 15, 2025

Discover AI apps

- Discover and monitor AI app usage with Defender for Cloud Apps with a risk assessment to help determine whether to sanction or un-sanction an app
- Discover user interactions and understand if sensitive data is sent to AI apps with Purview Data Security Posture Management (DSPM) for AI reports and activity explorer by creating these policies:
  - Get started – Extend your insights for data discovery
    - Detect when users visit AI sites – Purview IRM
    - Detect sensitive info shared in AI prompts in Microsoft Edge – Purview Browser Data Security
    - Detect sensitive info pasted or uploaded to AI sites – Purview Endpoint DLP
- Gain visibility into risky AI interactions in the browser or over the network with Purview Insider Risk Management

Block access to unsanctioned apps

- Block access for the whole organization to unsanctioned AI apps with Defender for Cloud Apps
- Allow limited access to certain AI apps, and restrict specific users and groups from accessing the unsanctioned AI apps with Entra internet access, or only block elevated risk user from accessing these AI app with Purview IRM and Entra conditional access
- Block install of unsanctioned AI apps on devices with Intune

Block sensitive data being sent to sanctioned AI apps

- Label content with Purview sensitivity labels
  - Other AI apps cannot decrypt content that’s encrypted by Purview
- Block pasting and uploading of sensitive info to AI apps with Purview Endpoint Data loss prevention (DLP)
- Block sensitive info in text prompts being sent to ChatGPT consumer, Microsoft Copilot (consumer version), DeepSeek, and Google Gemini in Microsoft Edge with Purview DLP or collection policies for Microsoft Edge.
- Detect sensitive data shared with cloud apps through non-Microsoft browsers, apps, APIs, add-ins, and more with Purview Network Data Security policies
- Monitor and protect Copilot use by detecting and responding to suspicious interactions with Copilot with Defender for Cloud Apps

Govern data sent to sanctioned AI apps

- Capture AI interactions with Purview Audit
- Detect inappropriate behaviors in prompts with Purview Communication Compliance
- Retain and/or delete prompts with Purview Data Lifecycle Management retention policies
- Preserve, collect, analyze, review, and export prompts with Purview eDiscovery by creating a case

Microsoft Security

Thank you

Microsoft Deployment models

Read the detailed guide for this model at https://aka.ms/PurviewDeploymentModels/ShadowAI