BLACKSWAN CYBERSECURITY

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

Severity: High **CVSS:** 8.8 **CVE:** CVE-2026-24301
Affected service: Microsoft Copilot Personal / Copilot Web
Not Affected: Copilot Enterprise
Status: Patched by Microsoft on August 18, 2026

Executive Summary

Varonis Threat Labs disclosed a high-severity vulnerability chain in Microsoft Copilot Personal named **CoSnitch** and tracked as **CVE-2026-24301**. The vulnerability could allow an attacker to distribute a specially crafted Copilot link that automatically executes attacker-controlled instructions when opened. The injected instructions could cause Copilot to query information available through connected services and transmit the resulting data to an attacker-controlled external endpoint.

The demonstrated attack chain combined three distinct weaknesses:

1. Automatic execution of an attacker-supplied prompt through a specially constructed Copilot URL.
2. Collection and exfiltration of information from services connected to the victim's Copilot account.
3. Persistent manipulation of Copilot memory through hidden instructions embedded in a webpage submitted for summarization.

Accessible connected services included Gmail, Google Drive, Google Calendar, and OneDrive. Exfiltrated information could be encoded into an outbound URL and transmitted through Copilot's legitimate URL-fetching function, potentially causing the resulting traffic to resemble ordinary Copilot web activity.

Microsoft assigned CVE-2026-24301 a CVSS v3.1 score of **8.8 High**.

The published vector indicates network accessibility, low attack complexity, no privileges required, and required user interaction. Microsoft described the issue as improper neutralization of special command elements that could allow an unauthorized attacker to disclose information over a network.

Microsoft deployed the service-side fix on August 18, 2026. According to a Microsoft statement reported by Dark Reading, the issue affected Copilot Personal - enterprise customers were unaffected - and customers did not need to take action to receive the fix.

BLACKSWAN CYBERSECURITY

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

Organizational Risk Assessment

Risk to Microsoft 365 Copilot

The specific CoSnitch vulnerability reportedly **affected Copilot Personal rather than Microsoft 365 Copilot Enterprise**. Organizations should therefore avoid treating every Microsoft 365 Copilot Enterprise deployment as directly vulnerable to this CVE.

The vulnerability nevertheless demonstrates a broader security risk relevant to enterprise AI deployments: an AI assistant with access to email, files, calendars, cloud storage, or other connected data sources may become an indirect collection and exfiltration mechanism if it cannot reliably distinguish trusted user instructions from attacker-controlled content.

Potential Business Impact

If a similar attack succeeded against an exposed user, potential consequences could include:

- Disclosure of email, calendar, cloud-storage, or other information accessible through connected applications.
- Loss of confidentiality for personal information, authentication artifacts, financial records, business communications, or intellectual property present in connected services.
- Continued unauthorized AI behavior if malicious instructions were written into persistent Copilot memory.
- Incomplete remediation if responders revoked sessions or reset passwords but did not inspect the AI assistant's retained memory or personalization data.
- Reduced visibility because the collection and outbound requests could occur through an otherwise legitimate AI service workflow.

The first three impacts reflect the capabilities demonstrated by Varonis. The last two are defensive risk considerations derived from the reported persistence and exfiltration mechanisms rather than evidence of observed attacks.

Technical Analysis

Initial Access

An attacker would prepare a legitimate-looking Microsoft Copilot URL containing an attacker-controlled prompt. The demonstrated construction used Copilot's documented ?q= query parameter together with an undocumented parameter. When the victim opened the link, the supplied prompt could execute on page load without a separate confirmation or prompt-submission action.

Although the attack has been described as "one-click," the underlying prompt itself reportedly required no additional interaction after the victim opened the crafted link.

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

Collection

The automatically executed prompt could instruct Copilot to search services connected to the victim's account. The research specifically identified Gmail, Google Drive, Google Calendar, and OneDrive as examples of connected applications from which information could be retrieved.

The data exposed in a particular incident would depend on:

- Which applications the user connected to Copilot.
- The permissions granted to those connections.
- The data available to the affected identity.
- The content and scope of the injected prompt.

These are investigation variables, not published indicators of compromise.

Command-and-Control and Exfiltration

After collecting information, the malicious prompt could encode the results into a URL and cause Copilot's URL-fetch functionality to contact an attacker-controlled webhook or server. The primary research states that the exfiltration could occur through the assistant's built-in URL-fetching capability.

Secondary reporting states that the stolen information could be Base64-encoded and transmitted in outbound web requests that resembled normal Copilot browsing behavior.

Persistence Through Memory Poisoning

A separate component of the CoSnitch chain involved a malicious webpage containing hidden instructions. If a victim asked Copilot to summarize the page, the hidden instructions could reportedly be inserted into Copilot's persistent memory. Varonis reported that the resulting memory injection could survive password changes, session revocation, and device re-enrollment.

This behavior means that conventional identity-containment actions, standing alone, might not remove malicious instructions already stored in an assistant's memory. Responders investigating a suspected AI prompt-injection event should therefore include retained memory, personalization, and connected-application configurations within the scope of the investigation.

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

Attack Chain

The following two flows summarize the technical sequences explicitly described in the Varonis research.

Attacker prepares crafted Copilot URL
v
Victim opens a legitimate Microsoft-hosted link
v
Embedded prompt automatically executes
v
Copilot queries connected applications
v
Retrieved information is encoded into a URL
v
Copilot fetches attacker-controlled endpoint
v
Information reaches attacker infrastructure

An alternate persistence path demonstrated by the researchers was:

Attacker publishes webpage containing hidden AI instructions
v
Victim asks Copilot to summarize the webpage
v
Hidden instructions are processed as commands
v
Instructions are added to persistent Copilot memory
v
Future Copilot behavior may remain attacker-influenced

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

Tactics, Techniques, and Procedures

The following MITRE ATT&CK mappings are **defender-oriented analytical mappings**. MITRE has not necessarily assigned these techniques specifically to CoSnitch.

Tactic	Technique	CoSnitch Application
Initial Access	Spearphishing Link	A crafted Copilot link could be delivered by email, chat, social media, or another messaging channel. The delivery mechanism was not prescribed by the research.
Execution	User Execution: Malicious Link	The victim opens the attacker-prepared URL, triggering the attack chain.
Discovery	Cloud Service Discovery	An attacker may attempt to determine which cloud services or connectors are available to the assistant. This is a plausible hunting mapping rather than a specifically reported CoSnitch action.
Collection	Data from Cloud Storage	Copilot could retrieve data from connected cloud services such as Google Drive or OneDrive.
Collection	Email Collection	Connected Gmail information could potentially be queried by the injected prompt.
Collection	Data from Information Repositories	The assistant could collect information from services and repositories available through configured connections.
Command and Control	Application Layer Protocol: Web Protocols	The attack used web requests and Copilot's URL-fetching function to reach external infrastructure.
Exfiltration	Exfiltration Over Web Service	Collected information could be transmitted to an attacker-controlled webhook or server through web functionality.
Command and Control / Exfiltration	Data Encoding: Standard Encoding	Secondary reporting describes the use of Base64 encoding within outbound requests.
Persistence	Account Manipulation: Additional Cloud Roles or Permissions	This is **not a direct mapping** to the demonstrated memory-poisoning behavior. AI-memory persistence does not currently align cleanly with a traditional ATT&CK technique and should be tracked as a separate AI-specific behavior.

AI-Specific TTPs Outside Traditional ATT&CK Coverage

- **Indirect prompt injection:** Instructions are embedded in attacker-controlled content and interpreted by an AI assistant as commands.

THREAT ADVISORY
CoSnitch Microsoft Copilot Personal Information Disclosure (prior to 8/18/26)

August 26, 2026

- **Prompt auto-execution:** A crafted deep link supplies and executes an instruction without a second explicit user submission.
- **AI connector abuse:** The assistant's authorized access to external services is repurposed for unauthorized collection.
- **Trusted-function exfiltration:** Legitimate URL retrieval or browsing behavior is used to transmit collected information.
- **Persistent memory poisoning:** Malicious instructions are inserted into the assistant's retained memory or personalization state.
- **Meta-hacking:** Researchers progressively questioned the AI about why an attack would not work, causing it to disclose architectural and guardrail details that helped reveal the attack path.

Indicators of Compromise

Confirmed Static IOCs

No attacker-controlled domains, IP addresses, file hashes, email addresses, webhook identifiers, or complete exploit URLs were published in the reviewed sources.

This is important operationally. CVE-2026-24301 is a vulnerability identifier, not evidence that a system or account was compromised. Similarly, the presence of the normal Copilot ?q= parameter alone is not a reliable indicator because that parameter can be associated with legitimate functionality.

Indicator	Type	Confidence	Notes
CVE-2026-24301	Vulnerability identifier	High	Microsoft Copilot information-disclosure vulnerability.
CWE-77	Weakness classification	High	Improper neutralization of special elements used in a command.
Copilot links containing ?q= plus an additional unusual or undocumented parameter	Suspicious URL pattern	Medium	The complete malicious parameter combination was not included in the reviewed public summaries. Do not alert solely on ?q=.
Long encoded strings in Copilot-triggered outbound URLs	Behavioral observable	Medium	Could represent encoded exfiltrated data, but may also generate benign matches.
Outbound Copilot URL fetches to newly observed webhook or data-collection services	Behavioral observable	Medium to High	Higher confidence when immediately preceded by a crafted Copilot link and connected-app access.

Indicator	Type	Confidence	Notes
Unexpected instructions in Copilot memory or personalization	Host/account artifact	High	Particularly suspicious if instructions reference external URLs, data collection, secrecy, encoding, or recurring actions.
AI summarization of a webpage followed by an unexplained memory change	Behavioral sequence	High	Consistent with the described persistent memory-poisoning chain.

Detection and Hunting Recommendations

Priority 1: Identify Exposed Copilot Personal Usage

1. Identify users who accessed Copilot Personal or Copilot Web around or before the August 18, 2026 fix.
2. Determine whether users connected personal or organizational Gmail, Google Drive, Google Calendar, OneDrive, or other data services to Copilot.
3. Prioritize accounts with access to sensitive business, customer, financial, legal, or security information.
4. Identify shadow-AI use outside sanctioned enterprise AI services.

Microsoft stated that customers were already protected after the service-side update, so these steps are retrospective exposure-assessment recommendations rather than patch-deployment requirements.

Priority 2: Hunt for Suspicious Copilot Links

Search relevant email, proxy, browser, endpoint, messaging, and secure web gateway records for:

- Copilot URLs containing a populated q parameter plus additional unusual query parameters.
- Copilot links delivered from newly registered, external, or unexpected senders.
- URL-encoded prompt text containing terms associated with:
    - Reading or searching connected applications.
    - Collecting emails, files, calendars, tokens, or account data.
    - Encoding results.
    - Fetching, opening, calling, or visiting an external URL.
    - Concealing actions or suppressing user-visible output.
- Copilot links immediately followed by access to an unfamiliar external domain.

Because the undisclosed parameter was not published in the reviewed sources, defenders should avoid depending on a single exact URL signature.

Priority 3: Hunt for Connected-Service Collection

Review cloud and identity telemetry for:

- Unusual bursts of Gmail, Google Drive, Google Calendar, or OneDrive access associated with a user who had just visited Copilot.
- Queries covering unusually broad date ranges, folders, mailboxes, or repositories.
- Access to data outside the user's normal activity profile.
- AI-mediated access to multiple connected services in a short sequence.
- Access followed immediately by an outbound request to a previously unseen destination.

Priority 4: Hunt for Exfiltration Behavior

Examine proxy, DNS, network, and cloud access logs for:

- Requests from a Copilot workflow to webhook, paste, request-capture, tunneling, or other user-controlled endpoints.
- Long URL paths or query strings containing Base64-like data.
- High-entropy query parameters following connected-application access.
- Repeated outbound requests containing similarly structured encoded values.
- External destinations first observed immediately after a Copilot deep-link event.
- Connections to destinations with no established business relationship.

Do not automatically classify every long or Base64-like URL as malicious. Correlate it with Copilot access, connected-service retrieval, user identity, destination reputation, and timing.

Priority 5: Inspect Memory and Personalization

For users considered potentially exposed:

- Review available Copilot memory and personalization entries.
- Look for instructions the user does not recognize.
- Look for entries directing Copilot to contact an external URL, encode information, suppress responses, or repeat future actions.
- Document suspicious entries before removing them if an investigation or legal hold applies.
- Validate that the unwanted behavior no longer occurs after memory remediation.

Recommended Mitigations

Immediate

1. **Verify scope.** Confirm whether the organization permits Copilot Personal and distinguish it from Microsoft 365 Copilot Enterprise.
2. **Confirm service protection.** Microsoft stated that the service-side fix was deployed on August 18, 2026 and no customer action was required.
3. **Inventory AI connectors.** Identify data sources connected to personal and enterprise AI assistants.

4. **Remove unnecessary connections.** Disconnect applications and repositories that do not have an active business need.
5. **Apply least privilege.** Limit each assistant to the minimum information required for authorized use.
6. **Warn users about AI deep links.** Treat unexpected links that open an AI assistant with a pre-populated prompt as potentially malicious.
7. **Enhance outbound monitoring.** Monitor AI-mediated requests to new, rare, or unapproved external destinations.
8. **Add AI state to incident response.** Include memory, personalization, connectors, conversation history, and retained instructions in AI-related investigation checklists.

Varonis specifically recommended mapping organizational AI usage, auditing connector configurations, disconnecting unnecessary applications, minimizing access, monitoring AI-originated access, and exercising caution with links that open an AI tool.

Strategic

- Establish a formal inventory of approved AI assistants, owners, connectors, permissions, and associated data classifications.
- Require security review before enabling AI integrations with email, cloud storage, calendars, source-code repositories, CRM systems, or ticketing platforms.
- Create detections that correlate AI access, connected-service collection, and outbound network activity.
- Test whether existing DLP, CASB, SSE, and SIEM controls can distinguish authorized AI activity from AI-mediated data exfiltration.
- Incorporate indirect prompt injection and memory poisoning into AI red-team exercises.
- Require AI vendors to document memory controls, connector authorization boundaries, audit telemetry, and incident-response procedures.
- Treat AI assistants with broad access as privileged identities rather than ordinary productivity applications.

Incident Response Guidance

If suspicious CoSnitch-like activity is identified:

1. Preserve the original email, message, or webpage containing the suspected link.
2. Capture the complete URL without opening it in an active user session.
3. Preserve browser history, proxy events, DNS logs, Copilot conversation history, cloud audit records, and identity telemetry.
4. Identify every application connected to the affected Copilot account.
5. Determine what information those connections could access.

6. Inspect outbound requests to identify attacker-controlled collection infrastructure.
7. Review Copilot memory and personalization for injected instructions.
8. Remove unauthorized memories or retained instructions after evidence preservation.
9. Revoke sessions, rotate potentially exposed credentials, and reassess application consent where justified by evidence.
10. Confirm that unexpected AI behavior and outbound communication have stopped.
11. Assess whether regulated, confidential, or customer information was accessed or transmitted.
12. Escalate to legal, privacy, compliance, and executive stakeholders when the affected information warrants it.

Conclusion

CoSnitch is significant, less because it introduced a conventional software exploit and more because it demonstrated how multiple AI-native weaknesses could be chained together: deep-link prompt execution, abuse of trusted connectors, exfiltration through normal AI web behavior, and persistent memory poisoning. The vulnerability also demonstrated "meta-hacking," where the AI's explanations about its own controls helped researchers identify the vulnerable behavior.