ORNA™

Cybersecurity Advisory

We are publishing this report in response to a request made in regards to any cyber risk to the implementation of MS Co-Pilot.

Summary

Comparing MS Copilot to other generative AI tools such as ChatGPT we know that it has access to everything your organization has ever worked on in MS365. Copilot has the capability to instantly search and compile data from across your documents, presentations, email, calendar, notes, and contacts.

What we have learned as cyber security professionals over the short period of time AI has started to become commonplace within business operations is that AI systems are subject to novel security vulnerabilities, which we will describe briefly in the following section, that need to be considered *alongside* standard cyber security threats. Due to the high pace of development of AI, security can become a secondary consideration. However, security must be a core requirement, not just in the *development* phase of an AI system, but *throughout its entire lifecycle*. It is therefore crucial that those responsible for the use of AI systems within their business environments, including senior management, keep abreast of new developments as well as the evolving threat landscape.

What is MS Copilot?

Microsoft Copilot is an artificial intelligence (AI)-powered tool that is primarily meant to support Microsoft 365 users with automation features for Word, Excel, PowerPoint, Outlook and Teams.

Copilot has been designed to function as an assistant, the primary benefit of using Copilot is improved work-related productivity through the automation of repetitive tasks, such as writing repetitive emails and summarizing documents. In addition to productivity gains, Copilot has the potential to boost user creativity by suggesting new ideas, formats and content based on context and preferences, whilst improving communication by ensuring emails are sent and streamlines the workflow of Microsoft 365 applications.

Beyond workforce productivity, Copilot can enhance the decision-making process through data analytics, financial analysis, market research and project planning, suggesting next steps in a process based on context and past experience, reducing workload and fatigue.

Types of Cyber Security Attacks on AI Technologies

As we've just covered generative AI, including Copilot, and LLMs in particular are undoubtedly impressive in their ability to generate a huge range of content in different work based scenarios for multiple business functions. However, the content produced by these tools is only as good as the data they are trained on, and the technology contains some risks, including:

- it can get things wrong and present incorrect statements as facts (a flaw known as 'AI hallucination')
- it can be biased and is often gullible when responding to leading questions
- it can be coaxed into creating toxic content and is prone to 'prompt injection attacks'
- it can be corrupted by manipulating the data used to train the model (a technique known as 'data poisoning')

A summary of Prompt Injection Attacks and Data Poisoning is as follows:

Prompt injection attacks are one of the most widely reported weaknesses in LLMs. This is when an attacker creates an input designed to make the model behave in an unintended way. This could involve causing it to generate offensive content, or reveal confidential information, or trigger unintended consequences in a system that accepts unchecked input.

Data poisoning attacks occur when an attacker tampers with the data that an AI model is trained on to produce undesirable outcomes (both in terms of security and bias). As LLMs in particular are increasingly used to pass data to third-party applications and services, the risks from these attacks will grow over time.

Cyber Secure Implementation of MS Copilot

With collaboration tools, there is always an extreme tension between productivity and security. This following section is what we know from Microsoft and what you can manage to best implement MS Copilot into business operations without impacting organizational security.

What Microsoft handles for you:

- **Tenant isolation**. Copilot only uses data from the current user's M365 tenant. The AI tool will not surface data from other tenants that the user may be a guest, nor any tenants that might be set up with cross-tenant sync.
- **Training boundaries.** Copilot does not use any of your business data to train the foundational LLMs that Copilot uses for all tenants. You *shouldn't* have to worry about your proprietary data showing up in responses to other users in other tenants.

What you need to manage:

- **Permissions.** Copilot surfaces all organizational data to which individual users have at least view permissions.
- **Labels.** Copilot-generated content *will not* inherit the MPIP labels of the files Copilot sourced its response from.
- **Humans.** Copilot's responses aren't guaranteed to be 100% factual or safe; humans must take responsibility for reviewing AI-generated content.

Beyond making these control changes and implementing new AI-related security policies, the monitoring of network traffic in the MS365 environment to enable early detection of any abnormal activity, whilst in addition, running cyber simulation attack exercising specifically based on a MS365 Copilot cyber incident in order to exercise all relevant stakeholders in the face of this new threat vector.

Assessment

Generative AI technology is still relatively new, especially in the business function use case, and therefore the types of cyber attack are relatively low and unknown, however the Canadian NCSC, US CISA, and UK NCSC have released joint publications on the threat to businesses via AI over the past 12 months.

In the specific case of Copilot, the largest risk we have identified is the access to all sensitive information hosted within the MS365 environment, and with this, the ability for Copilot to rapidly generate new sensitive information that needs controls to protect the data. The ability to create and share data faster than the ability to create controls to protect it has often been the biggest challenge when looking at data breach trends. The implementation of generative AI into the business domain increases this risk exponentially.

There have already been real world breaches in which MS Copilot drafted a customer proposal which included sensitive data belonging to a completely different customer creating a privacy or data breach incident. We assess that an understanding of your data security posture be first understood before an MS Copilot rollout, including a form of audit as to where sensitive data sits across the organizations MS365 infrastructure that we know Copilot will have access to.

Recommendations

Recommendation #1: Sensitive Data Audit

Engage in the discovery and classification of all sensitive AI-generated data.

Recommendation #2: MS365 Controls Audit

Implement correct MPIP labels and enforce least privilege permissions if not already done.

Recommendation #3: Managed Monitoring, Detection & Response

Monitor network traffic in the MS365 environment and detect malicious communications that may be originating from suspicious locations.

Recommendation #3: Cyber Crisis Simulation Exercising

Conduct simulated attack scenarios to make sure that C-Suite, Senior Management and employees are well aware of how to respond to an AI-related cyber attack, and that they report the incident to the appropriate cybersecurity team(s).

References

1. Guidelines for secure AI system development:
   https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
2. Thinking about the security of AI systems:
   https://www.ncsc.gov.uk/blog-post/thinking-about-security-ai-systems

If you have any additional questions, please reach out to your ORNA representative at sme@orna.app.

Sincerely,
ORNA Team