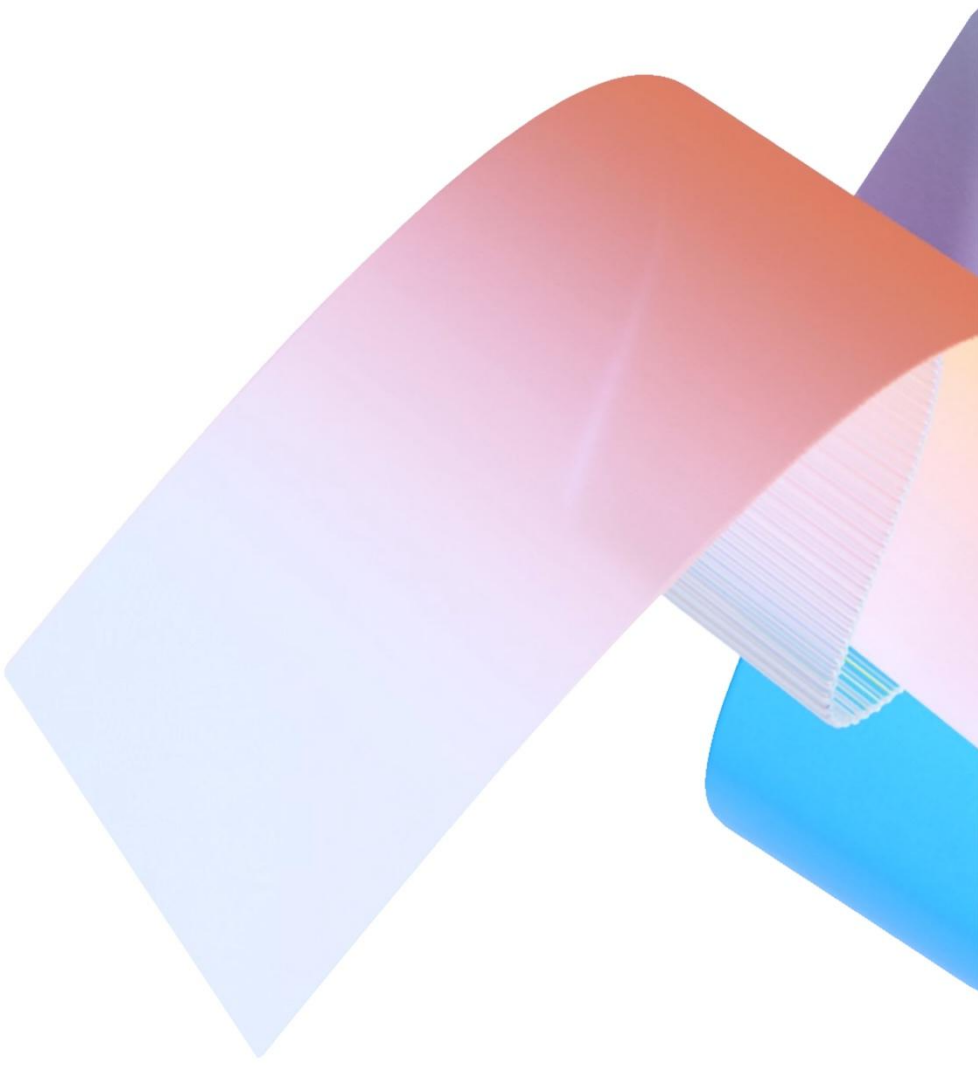
Microsoft

Agent Security Playbook



1 Introduction

1.1 Purpose

The purpose of this playbook is to provide clear, practical guidance for securing agents throughout their lifecycle—from inception and deployment to ongoing management—within your organization. By outlining recommended practices, critical considerations, and proven strategies, the playbook aims to empower teams to confidently navigate evolving security requirements, proactively address threats, and ensure compliance in dynamic digital environments.

1.2 Scope

This playbook focuses on governing agents within the Microsoft 365 ecosystem, including:

- **SharePoint** – Use SharePoint to build agents based on content stored in SharePoint sites and libraries. Learn more at Get started with SharePoint agents.
- **Copilot Studio Agent Builder** – Use the Copilot Studio Agent Builder tool directly within Microsoft 365 Copilot to create conversational templates tailored to specific tasks or business needs. Learn more at Use Copilot Studio Agent Builder to Build Agents.
- **Copilot Studio** – Use triggers, advanced logic, and connections to other Microsoft services or third-party platforms to create agents with full Copilot Studio. Learn more at Overview - Microsoft Copilot Studio.
- **Pro Developer Tools** – Use tools and services like Team Toolkit and Azure AI Foundry to build fully-customized agents with the model and orchestration engine of your choice.

1.3 Target Audience

The target audience for this playbook is Administrators and Architects in Small to Medium Businesses (SMB) and Large Enterprises who are responsible for their organization's agent security management and strategy.

1.4 How to use this playbook

This playbook offers an overview and practical guidance for agent security. Use it as a comprehensive guide or refer to sections as needed. It helps you:

- Identify key security challenges in agent management
- Review features and tools for risk mitigation
- See which agents and scenarios are affected
- Compare unique benefits of each security solution
- Find references and resources for deeper exploration

2 Overview

Agents are intelligent software tools that automate tasks, boost productivity, and support decision-making within organizations. They can be created using Copilot Studio Agent Builder, Copilot Studio, or pro developer tools like Teams Toolkit and Azure AI Foundry, offering varying levels of customization and integration with Microsoft 365.

The governance of agents in Microsoft 365 is organized into zones, each reflecting a stage of maturity. These zones help organizations progress from basic controls to advanced, integrated governance. Zone 1 covers personal productivity, Zone 2 adds departmental collaboration, and Zone 3 focuses on enterprise management with enhanced security and analytics. This tiered approach ensures organizations can scale and adapt their agent management and security as they grow.

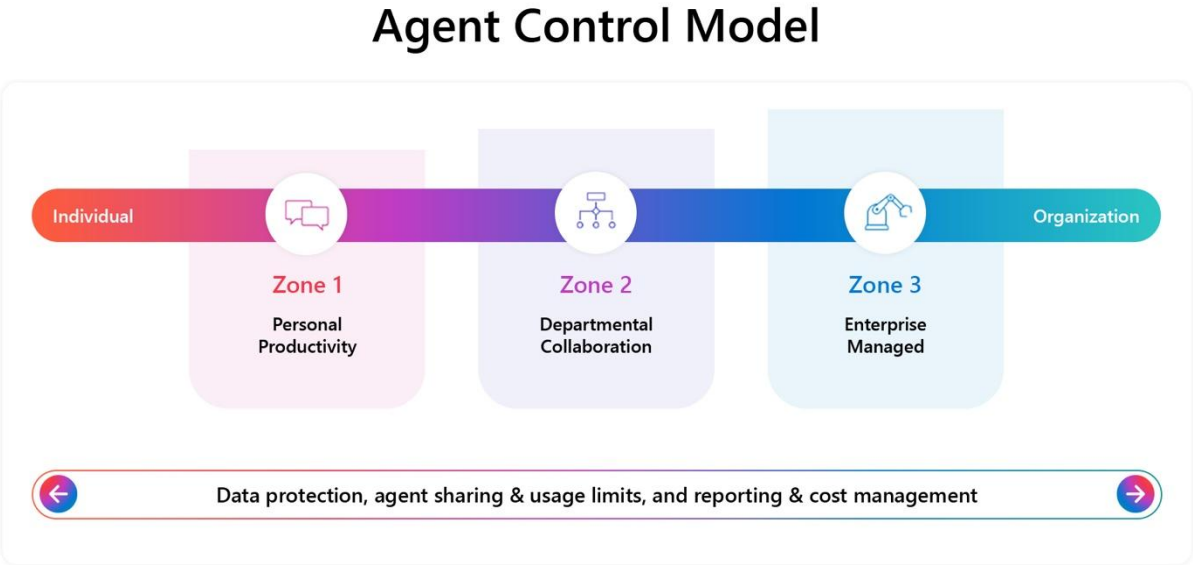


Figure 1 Agent Zones

Effective governance within the Microsoft 365 ecosystem also requires a structured and tiered approach to security, management, and operational strategies. By segmenting governance into distinct pillars, organizations can address their evolving needs while maintaining scalability and adaptability.

Governance in Microsoft 365 is organized into structured zones to ensure security, management, and operational effectiveness as organizations mature. Key elements include:

- **Security Controls:** Safeguard data and systems through technical, administrative, and physical measures.
- **Management Controls:** Provide policies and procedures for overseeing agent lifecycle, ensuring compliance and integration across the IT environment.
- **Agent Reporting:** Systematically collect and analyze agent activity data to improve transparency, accountability, and decision-making.

Together, these pillars create a robust, adaptable governance framework for managing agents within Microsoft 365.

This playbook primarily addresses Security Controls. Additional information regarding Management Controls and Agent Reporting can be found in section 5 Appendix.

3 Security Controls

Security controls are measures, policies, and procedures designed to safeguard an organization's assets, data, and systems from unauthorized access, misuse, or exploitation. These controls typically encompass technical, administrative, and physical elements that work in concert to mitigate risks and ensure compliance with regulatory and organizational standards. From encryption protocols and access management to monitoring tools and incident response strategies, security controls form the backbone of a robust governance framework. Their implementation ensures that data integrity, confidentiality, and availability are maintained across all zones and operational layers.

In today's digital landscape, organizations face an evolving array of security challenges that demand robust, adaptable solutions. To effectively protect sensitive information, maintain compliance, and support secure collaboration, it's essential to address a range of security controls within your agent ecosystem.

The following sections provide an overview of common security problems encountered in agent management and the strategic features that resolve them.

For each security problem it dives into the details outlining the nature of the risk, recommended approaches, and detailed descriptions of relevant features or tools. Where applicable, you'll also find guidance on implementation, value to your organization, and references for additional information. This structured approach ensures you have a clear roadmap for strengthening your organization's security posture across agent environments.

3.1 Prevent data exfiltration and misuse of sensitive data

3.1.1 Background

Data exfiltration refers to the unauthorized transfer or movement of data from within an organization to an external location or party. This can occur through various means, such as malicious insiders, compromised accounts, or security vulnerabilities that allow attackers to send confidential information outside the organization's control. Misuse of

sensitive data, on the other hand, encompasses any improper handling, access, or sharing of confidential or protected information—whether intentional or accidental—that violates policies, regulations, or privacy expectations. Such misuse can expose organizations to legal risk, reputational harm, and financial loss, making robust prevention mechanisms critical in modern security strategies.

3.1.2 What to do

To **prevent data exfiltration and misuse of sensitive data**, organizations can enforce **Data Loss Protection (DLP) Policies** for **all agents** using **Power Platform Admin Center** (PPAC) and **Microsoft Purview**.

3.1.3 More Information

Both the Power Platform Admin Center (PPAC) and Microsoft Purview play pivotal roles in establishing and managing Data Loss Protection (DLP) Policies to safeguard organizational data.

Within the Power Platform Admin Center, administrators can define DLP policies that govern which connectors—both standard and custom—can be used together within Power Apps and Power Automate environments. By grouping connectors into "business" and "non-business" categories, the platform ensures that sensitive data remains confined to trusted channels and isn't inadvertently shared with less secure or external services. For example, a policy can prevent data from being transferred from a secure SharePoint Online environment to an external social media connector, thus mitigating the risk of accidental or intentional data leakage. The process includes selecting environments, categorizing connectors, and enforcing rules that align with the organization’s data handling standards.

Microsoft Purview, on the other hand, delivers a comprehensive suite for information governance and compliance management, extending DLP capabilities across the broader Microsoft 365 ecosystem. Administrators can craft granular DLP policies that monitor and control data both in transit and at rest—whether in emails, documents, or collaboration platforms like Teams and SharePoint. Purview leverages advanced content inspection, pattern recognition, and policy templates tailored to regulatory requirements (such as GDPR or HIPAA) to automatically detect and restrict activities that might expose sensitive information, such as sharing credit card numbers, health records, or intellectual

property beyond approved boundaries. Policies are customizable, allowing organizations to set specific actions—alerts, automatic encryption, blocking, or user notifications—based on the context and risk level of each data interaction.

By harnessing both the Power Platform Admin Center for app-level controls and Microsoft Purview for ecosystem-wide governance, organizations can implement layered, adaptive DLP measures that prevent data exfiltration, enforce compliance, and provide visibility into data flows. This integrated approach not only curtails misuse or unauthorized sharing of sensitive data but also provides audit trails and reporting tools essential for demonstrating regulatory adherence and maintaining stakeholder trust.

3.1.4 Learn More

Learn more about **Data Loss Protection (DLP) Policies:**

- Data loss prevention (DLP) policies - Power Platform | Microsoft Learn
- Manage data policies - Power Platform | Microsoft Learn
- Configure data loss prevention policies for agents - Microsoft Copilot Studio | Microsoft Learn
- Learn about Microsoft Purview | Microsoft Learn
- Learn about data loss prevention | Microsoft Learn
- Data Loss Prevention policy reference | Microsoft Learn
- Learn about the Microsoft 365 Copilot location (preview) | Microsoft Learn

3.2 Ensure consistent data classification and compliance

3.2.1 Background

Consistent data classification is fundamental to effective compliance and risk management. By accurately labeling and categorizing information according to sensitivity and regulatory requirements, organizations can ensure that data receives appropriate protection throughout its lifecycle. This precision not only helps prevent accidental exposure or mishandling of confidential information but also streamlines audits and demonstrates adherence to legal and industry standards. Moreover, well-

defined classification policies empower teams to apply relevant access controls, automate policy enforcement, and swiftly respond to evolving regulatory landscapes, ultimately reducing the risk of costly breaches or penalties.

3.2.2 What to do

To **ensure consistent data classification and compliance**, organizations can enforce **Sensitivity Labels for Knowledge and Connectors** for **Copilot Studio Agents** using **Microsoft Purview**.

3.2.3 More Information

Sensitivity Labels for Knowledge and Connectors enable organizations to systematically tag and categorize data based on their level of sensitivity and regulatory requirements. Leveraging Microsoft Purview, these labels allow for automated enforcement of data protection policies, ensuring that sensitive information is consistently secured throughout its lifecycle—from creation to sharing and storage. This approach helps prevent accidental leaks, restricts unauthorized access, and simplifies compliance reporting. By applying these labels across knowledge assets and data connectors, organizations gain granular control over who can view, modify, or distribute confidential content, making it easier to adapt to new regulations and audit demands while reducing the risk of data breaches or compliance failures.

3.2.4 Learn More

Learn more about **Sensitivity Labels for Knowledge and Connectors:**

- Learn about sensitivity labels | Microsoft Learn
- View sensitivity labels in connectors | Microsoft Learn
- Manage agents with embedded file content as a knowledge source in the Microsoft 365 admin center - Microsoft 365 admin | Microsoft Learn
- Learn about the Microsoft 365 Copilot location (preview) | Microsoft Learn
- View sensitivity labels for SharePoint data sources - Microsoft Copilot Studio | Microsoft Learn

3.3 Enable secure, enterprise-grade user validation

3.3.1 Background

In today's interconnected business environment, verifying the identity of every user accessing sensitive systems is essential to maintaining the integrity and confidentiality of enterprise data. Secure, enterprise-grade user validation—such as manual authentication with certificates—ensures that only authorized personnel can interact with critical resources, minimizing the risk of impersonation, credential theft, or unauthorized access. This level of validation is foundational for upholding trust, safeguarding intellectual property, and complying with regulatory mandates. By enforcing rigorous user verification, organizations can more confidently defend against sophisticated cyber threats and insider risks, delivering peace of mind to stakeholders and demonstrating a proactive commitment to security best practices.

3.3.2 What to do

To **enable secure, enterprise-grade user validation**, organizations can use **Manual Authentication with Certificates** for **Copilot Studio Agents** using **Copilot Studio**.

3.3.3 More Information

By implementing Manual Authentication with Certificates for Copilot Studio Agents in Copilot Studio, organizations introduce a robust identity-verification layer that goes beyond basic username and password credentials. Manual certificate-based authentication ensures that only users who possess a valid digital certificate—issued and managed by the organization—can access Copilot Studio Agents and the sensitive data or functionality they provide. This approach safeguards enterprise resources by tying up access directly to individually authenticated users, making it significantly harder for unauthorized individuals to impersonate legitimate personnel or bypass security controls.

The use of certificates also enables granular tracking and auditing of user activity, as each action can be mapped to a unique, verified identity. This helps organizations maintain strong oversight, comply with regulatory requirements, and quickly identify or

respond to any potential misuse or security incidents. Ultimately, manual authentication with certificates delivers a higher assurance of user identity, reinforces trust in the system, and demonstrates the organization's commitment to modern, enterprise-grade security practices.

3.3.4 Learn More

Learn more about **Manual Authentication with Certificates:**

- Configure user authentication - Microsoft Copilot Studio | Microsoft Learn
- How to configure Microsoft Entra certificate-based authentication - Microsoft Entra ID | Microsoft Learn
- Microsoft Entra certificate-based authentication technical deep dive - Microsoft Entra ID | Microsoft Learn

3.4 Protect against unauthorized data access

3.4.1 Background

Preventing unauthorized data access is a cornerstone of information security, as it shields sensitive organizational assets from exposure to untrusted individuals or entities. When data is accessed by unauthorized users, whether due to technical vulnerabilities, weak authentication protocols, or insider threats, the organization becomes susceptible to breaches that may compromise privacy, disrupt operations, and erode stakeholder trust. Such incidents can lead to significant financial losses, regulatory penalties, and long-lasting reputational damage. Implementing measures to block anonymous or unauthorized users ensure that confidential information remains accessible only to those with a legitimate, verified need, thereby supporting compliance requirements and reinforcing the organization's defensive posture in an increasingly complex threat landscape.

3.4.2 What to do

To **protect against unauthorized data access**, organizations can enable **Anonymous User Blocking** for **all agents** using **Microsoft Sentinel and Power Platform Admin Center** (PPAC).

3.4.3 More Information

Enabling Anonymous User Blocking through Microsoft Sentinel and Power Platform Admin Center (PPAC) is a proactive security measure that significantly strengthens an organization's defense against unauthorized data access.

- **Restricts Access to Verified Users Only:** By requiring all users to be authenticated before gaining access to agents or sensitive data, this feature ensures that only individuals with legitimate, organization-approved credentials can interact with protected resources. This eliminates the risk posed by anonymous users, who often attempt to exploit security gaps to gain unauthorized entry.
- **Mitigates Multiple Threat Vectors:** Blocking anonymous access helps close common attack routes, such as those arising from weak authentication, technical vulnerabilities, or accidental exposure. Insider threats are also reduced, as every action taken can be traced to a specific, verified identity, discouraging misuse from within the organization.
- **Centralized Security Policy Management:** Microsoft Sentinel and PPAC provide organizations with a unified platform to set, enforce, and monitor access policies across their entire agent ecosystem. This centralization ensures consistency, streamlines administration, and minimizes configuration errors that could lead to accidental data exposure.
- **Real-Time Detection and Response:** With these platforms, organizations can continuously monitor unauthorized access attempts. Any suspicious activity, such as repeated failed authentication or attempts from unknown users, can be promptly flagged and investigated, shortening response times and minimizing potential damage.
- **Supports Regulatory Compliance:** Many data protection regulations require organizations to demonstrate strict control over who can access sensitive information. By enforcing authenticated access only, Anonymous User Blocking

helps meet these compliance obligations, reduces audit findings, and avoids costly penalties.

- **Builds Stakeholder Confidence:** A strong access control strategy reassures customers, partners, and regulators that the organization takes data privacy and security seriously. This trust is essential for maintaining business relationships and upholding the organization's reputation.
- **Enables Comprehensive Audit Trails:** Because access is never anonymous, every interaction with the system is linked to a specific, identified user. This provides a clear, detailed audit trail for investigations, compliance reporting, and incident response, making it easier to reconstruct events and hold individuals accountable.

In summary, enabling Anonymous User Blocking for all agents using Microsoft Sentinel and PPAC does more than just keep unwanted users out—it forms a foundational layer of enterprise security, ensuring that only trusted, verified individuals can interact with valuable organizational data and resources.

3.4.4 Learn More

Learn more about **Anonymous User Blocking:**

- Use governance controls to disable anonymous access | Microsoft Learn
- Data loss prevention example - Require user authentication in agents - Microsoft Copilot Studio | Microsoft Learn
- What is Microsoft Sentinel? | Microsoft Learn
- Roles and permissions in the Microsoft Sentinel platform | Microsoft Learn
- Azure identity and access security best practices | Microsoft Learn
- Manage agent security with enhanced admin controls | Microsoft Learn

3.5 Maintain auditability and accountability

3.5.1 Background

In complex organizational environments, maintaining auditability and accountability is vital for ensuring transparency in the management and the use of sensitive data.

Auditability enables organizations to trace every action taken on critical systems, offering a clear record of who accessed what information, when, and for what purpose. This comprehensive visibility not only deters malicious behavior but also facilitates swift investigation and remediation in the event of an incident. Likewise, accountability ensures that individuals are held responsible for their actions, reinforcing a culture of integrity and compliance. Together, these principles underpin both internal governance and external regulatory obligations, supporting trust among stakeholders and providing vital evidence during audits or legal proceedings.

3.5.2 What to do

To **maintain auditability and accountability**, organizations can **Run Connectors with End-User Credentials** for **Copilot Studio Agents** using **Copilot Studio**.

3.5.3 More Information

By running Connectors with End-User Credentials for Copilot Studio Agents, every action performed through the system is directly attributed to the individual user who initiated it, rather than to a generic service or shared account. This granular level of attribution means that all interactions, whether accessing, modifying, or transmitting data—are recorded with the specific identity of the end user. As a result, organizations can generate precise, user-level audit logs that detail what was done, by whom, and when. This approach not only strengthens traceability and transparency but also supports regulatory compliance by ensuring that evidence of user activity is readily available for audits, investigations, or incident response. Additionally, it reinforces accountability, as individuals are aware that their actions are visible and attributable, which can deter inappropriate behavior and foster a culture of responsibility across the organization.

3.5.4 Learn More

Learn more about **Run Connectors with End-User Credentials:**

- Configure user authentication for actions - Microsoft Copilot Studio | Microsoft Learn
- Use connectors in Copilot Studio - Microsoft Copilot Studio | Microsoft Learn

3.6 Ensure only compliant users/devices access agents

3.6.1 Background

In an era marked by relentless cyber threats and rapidly evolving compliance requirements, it is paramount that only users and devices meeting established security standards are permitted to access critical system agents. Agents often serve as intermediaries to sensitive data, core applications, or privileged services, making them attractive targets for attackers seeking entry points into the enterprise environment. By enforcing strict conditional access and robust endpoint management, organizations can limit exposure to compromised, non-compliant, or untrusted devices—reducing the attack surface and preventing the propagation of malware or unauthorized data flow. This selective gating not only fortifies the organization's defenses against external breaches and insider risks but also ensures adherence to regulatory frameworks that mandate stringent controls over system access. Ultimately, prioritizing compliance at the access layer preserves the integrity of business operations, protects confidential information, and fosters a resilient security posture that can adapt to emerging challenges.

3.6.2 What to do

To **ensure only compliant users/devices access agents**, organizations can enable **Conditional Access and Endpoint Management** for **all agents** using **Microsoft Entra** and **Intune**.

3.6.3 More Information

Microsoft Entra's Conditional Access and Intune's Endpoint Management work together to guarantee that only users and devices meeting security requirements can access critical system agents. Conditional Access checks user identities and device compliance in real time, blocking access from non-compliant or risky sources. Intune ensures all devices are securely configured and up to date. This combined approach reduces the risk of breaches, supports regulatory compliance, and maintains accountability by logging access attempts and outcomes.

3.6.4 Learn More

Learn more about **Conditional Access and Endpoint Management**:

- Configure Conditional Access in Microsoft Defender for Endpoint - Microsoft Defender for Endpoint | Microsoft Learn
- Enable Robust Security and Governance for Agents in Microsoft 365 Copilot - Microsoft Power Platform Blog
- Microsoft Entra Conditional Access optimization agent - Microsoft Entra ID | Microsoft Learn
- Building Conditional Access policies in Microsoft Entra - Microsoft Entra ID | Microsoft Learn

3.7 Enable cross-tenant and B2B scenarios securely

3.7.1 Background

In today's interconnected business landscape, organizations frequently collaborate across company boundaries, whether with partners, vendors, or remote subsidiaries—necessitating seamless and secure cross-tenant and business-to-business (B2B) interactions. Enabling these scenarios securely is crucial because they extend access to systems and data beyond the traditional perimeter, introducing new vectors for potential compromise. Without rigorous controls, external users or partner organizations might inadvertently gain excessive privileges, access sensitive information, or introduce vulnerabilities into shared environments. Securely managing federated identity credentials allows organizations to validate external identities, enforce granular access policies, and monitor activity with the same diligence applied to internal users. This approach not only mitigates the risk of unintentional data leakage or targeted attacks via third-party channels but also upholds regulatory and contractual obligations pertaining to data protection. Ultimately, secure enablement of cross-tenant and B2B scenarios empowers organizations to collaborate efficiently and confidently, unlocking business value while maintaining robust safeguards around critical assets.

3.7.2 What to do

To **enable cross-tenant and B2B scenarios securely**, organizations can use **Federated Identity Credentials** for **all agents** using **Copilot Studio**.

3.7.3 More Information

By implementing Federated Identity Credentials for all agents leveraging Copilot Studio, organizations create a unified framework for authenticating and authorizing external users, regardless of their home directory or organizational affiliation. Federated credentials allow each agent—whether from a partner, vendor, or subsidiary—to be uniquely identified and validated against their own trusted identity provider. This approach drastically reduces the risk of unauthorized access, since external identities never require direct provisioning within the host tenant and all authentication events are subject to real-time policy enforcement. Furthermore, organizations can apply conditional access rules, monitor agent activity, and rapidly revoke credentials if suspicious behavior is detected, ensuring that only the right individuals access sensitive resources. In this way, Federated Identity Credentials provide a robust, scalable, and auditable security posture for modern cross-tenant and B2B collaboration, aligning operational agility with enterprise-grade protection.

3.7.4 Learn More

Learn more about **Federated Identity Credentials:**

- Overview of federated identity credentials in Microsoft Entra ID - Microsoft Graph beta | Microsoft Learn
- Flexible federated identity credentials (preview) - Microsoft Entra Workload ID | Microsoft Learn
- Set up a Flexible Federated identity credential (preview) - Microsoft Entra Workload ID | Microsoft Learn

3.8 Block Cross Prompt Injection Attacks

3.8.1 Background

A Cross Prompt Injection Attack (XPIA) refers to a security vulnerability where an attacker manipulates prompts or instructions intended for one agent, tool, or AI system, causing them to be inadvertently processed or executed by another. In environments where multiple agents or systems may interact—especially across organizational boundaries or tenants—attackers can exploit these boundaries by injecting malicious commands or data into prompts exchanged between agents. The risk is particularly acute in federated or cross-tenant scenarios, where trust relationships are complex and prompts may traverse different security domains.

Such attacks can lead to unauthorized actions, data leakage, or unintended behavior by agents, undermining the integrity and confidentiality of your workflow. Recognizing and mitigating XPIA is crucial for maintaining a secure agent ecosystem, especially when leveraging advanced features like Federated Identity Credentials and cross-tenant access.

3.8.2 What to do

To **block Cross Prompt Injection Attacks (XPIA)**, organizations can use **real-time threat blocking** for **all agents** using **Microsoft Defender**.

3.8.3 More Information

Microsoft Defender introduces real-time protection for AI agents created with Copilot Studio, addressing emerging security risks like prompt injection attacks that can manipulate AI behavior or expose sensitive data. This integration enables organizations to securely leverage AI agents for task automation and productivity enhancement.

- **Real-time threat blocking:** Defender monitors AI agent tool invocations live and blocks suspicious or high-risk actions such as prompt injections before execution, preventing unauthorized data exposure or misuse.
- **Proactive security during runtime:** The protection acts as a safety net while agents operate, reducing vulnerabilities to malicious inputs and ensuring secure AI agent activity.

- **Unified incident investigation:** Blocked actions generate alerts integrated into Microsoft Defender XDR, allowing security teams to analyze AI-related threats with full context and streamlined workflows.
- **Enhanced security and productivity synergy:** The collaboration between Defender and Copilot Studio offers unified threat visibility, intelligent automation, and real-time recommendations, empowering organizations to confidently deploy AI agents.

3.8.4 Learn More

Learn more about **Federated Identity Credentials:**

- https://aka.ms/MCS-XPIA

4 Conclusion

In summary, the Agent Security Playbook provides a comprehensive roadmap for managing and securing agents within the Microsoft 365 ecosystem. By exploring foundational concepts, actionable steps, and advanced features across security, management, and governance pillars, this playbook empowers organizations to address the evolving landscape of digital threats with confidence.

Implementing the strategies and controls outlined here will help safeguard sensitive information, ensure compliance with regulatory standards, and foster a culture of accountability and continuous improvement. Whether your organization is just beginning its journey with agent security or seeking to refine mature practices, this playbook offers guidance adaptable to your unique needs.

By prioritizing robust security controls, leveraging management best practices, and embracing auditability and cross-tenant collaboration, organizations can build resilient, trusted environments that support productivity and innovation. Use this playbook as both a reference and a guide and remain proactive in adapting your security posture as technologies and threats evolve.

Together, these efforts will not only protect your digital assets but also empower your teams to work securely and effectively, today and into the future.

5 Appendix

While the primary focus of this playbook centers on Security Controls, it is important to acknowledge that Management Controls and Agent Reporting, though not strictly classified as security features, play a pivotal role in the overall health and resilience of your Microsoft 365 environment. The concerns addressed by these controls—such as operational consistency, compliance, and visibility—not only complement security initiatives but also help mitigate risks that can arise from unmanaged agents or insufficient oversight. The following appendix provides an overview of management controls and agent reporting considerations, recognizing their relevance and practical application alongside security controls in supporting a robust organizational posture.

5.1 Management Controls

Management controls are organizational policies and procedures that guide the deployment and oversight of agents in Microsoft 365. They help ensure agents are properly managed throughout their lifecycle, enforce compliance, and maintain operational consistency.

5.1.1 Prevent shadow IT and unmanaged agent sprawl

5.1.1.1 Background

Shadow IT refers to the use of information technology systems, devices, software, or services without explicit organizational approval. It often arises when employees use unauthorized tools to meet their needs, potentially introducing security risks and reducing visibility for IT teams. Unmanaged agent sprawl occurs when security or monitoring agents—software components installed on endpoints—are deployed widely but without centralized oversight. This can lead to redundant, outdated, or conflicting agents running across the environment, complicating management and increasing the risk of vulnerabilities.

5.1.1.2 What to do

To **prevent shadow IT and unmanaged agent sprawl**, organizations can enable **Agent Visibility and Lifecycle Management** for **all agents** using **Microsoft 365 Admin Center**.

5.1.1.3 More Information

Enabling Agent Visibility and Lifecycle Management for all agents through the Microsoft 365 Admin Center provides organizations with centralized oversight of every deployed agent. This comprehensive view allows IT teams to detect unauthorized or redundant agents, ensure all agents are properly updated, and manage their entire lifecycle. As a result, organizations can minimize the risk of shadow IT—where tools are used without approval—and prevent unmanaged agent sprawl, reducing complexity and improving security across their environment.

5.1.1.4 Learn More

Learn more about **Agent Visibility and Lifecycle Management:**

- Application lifecycle management (ALM) with Microsoft Power Platform - Power Platform | Microsoft Learn

5.1.2 Mitigate risk from rogue or misconfigured agents

5.1.2.1 Background

Rogue or misconfigured agents can introduce risks, such as unauthorized access to sensitive data, system instability, performance degradation, and the creation of security vulnerabilities. These agents may bypass organizational controls, making it harder to detect malicious activity or prevent data leaks, and could conflict with other security tools, weakening the overall security posture.

5.1.2.2 What to do

To **mitigate risks from rogue or misconfigured agents**, organizations can use **Agent Blocking and Unblocking** for **Shared Agents** using **Microsoft 365 Admin Center**.

5.1.2.3 More Information

By leveraging Agent Blocking and Unblocking in the Microsoft 365 Admin Center, organizations can quickly restrict or restore access for Shared Agents that are identified as rogue or misconfigured. This control helps prevent unauthorized activities, limits exposure to sensitive data, and reduces the risk of operational disruption by stopping problematic agents before they can cause harm or conflict with other systems.

5.1.2.4 Learn More

Learn more about **Agent Blocking and Unblocking:**

- Manage agents for Microsoft 365 Copilot in the Microsoft 365 admin center - Microsoft 365 admin | Microsoft Learn

5.1.3 Prevent accidental deployment in production

5.1.3.1 Background

Accidental deployment in production can expose systems to untested or unstable code, leading to unexpected outages, data loss, or security breaches. Such incidents disrupt critical business operations, damage user trust, and often require urgent remediation efforts to restore normal functionality.

5.1.3.2 What to do

To **prevent accidental deployment in production**, organizations can use **Environment Routing** for **Copilot Studio Agents** using **Power Platform Admin Center** (PPAC).

5.1.3.3 More Information

By configuring Environment Routing in the Power Platform Admin Center, organizations can direct Copilot Studio Agents to specific environments, such as development, testing, or production. This setup ensures that new changes are first deployed in non-production environments, allowing thorough validation before reaching live systems. As a result, Environment Routing acts as a safeguard, minimizing the risk of untested or unstable code being accidentally released into production and helping maintain system stability.

5.1.3.4 Learn More

Learn more about **Environment Routing:**

- Environment routing - Power Platform | Microsoft Learn

5.1.4 Educate makers and prevent risky configurations

5.1.4.1 Background

When makers are not properly educated about system security and best practices, they may inadvertently set up configurations that expose vulnerabilities or bypass critical safeguards. This lack of awareness increases the likelihood of introducing weak points, making systems more susceptible to errors and threats.

5.1.4.2 What to do

To **educate makers and prevent risky configurations**, organizations can configure **Maker Security Warnings** for **Copilot Studio Agents** using **Copilot Studio**.

5.1.4.3 More Information

By setting up Maker Security Warnings within Copilot Studio for Copilot Studio Agents, organizations provide timely guidance and alerts to makers about system security risks and best practices. These warnings help makers recognize potential vulnerabilities as they build or configure agents; increase their security awareness; and discourage actions that could compromise the system. As a result, makers become better equipped to avoid risky configurations and strengthen overall security.

5.1.4.4 Learn More

Learn more about **Maker Security Warnings:**

- Enable maker welcome content - Power Platform | Microsoft Learn

5.1.5 Prevent oversharing and data leakage

5.1.5.1 Background

Oversharing and data leakage can expose sensitive information to unauthorized individuals, leading to privacy violations, security breaches, and potential financial or reputational harm for organizations. When too much data is accessible or shared beyond intended audiences, it increases the risk that confidential or critical details may be misused or fall into the wrong hands.

5.1.5.2 What to do

To **prevent oversharing and data leakage**, organizations can enable **Sharing Limits** for **all agents** using **Power Platform Admin Center** (PPAC).

5.1.5.3 More Information

By enabling Sharing Limits for all agents through the Power Platform Admin Center (PPAC), organizations can control and restrict how much data agents can be shared and with whom. This reduces the chance of sensitive information being inadvertently exposed outside intended audiences, helping to safeguard confidential data and minimize the risk of privacy breaches or unauthorized access.

5.1.5.4 Learn More

Learn more about **Sharing Limits:**

- Control how agents are shared - Microsoft Copilot Studio | Microsoft Learn
- Block and limit sharing for Copilot Studio | Microsoft Learn

5.1.6 Ensure compliance with internal review processes

5.1.6.1 Background

Failing to ensure compliance with internal review processes can result in unvetted content or data being published or shared, increasing the likelihood of errors, policy violations, or the inadvertent release of sensitive information. This oversight undermines organizational safeguards designed to protect data integrity and minimize risks associated with unauthorized disclosures.

5.1.6.2 What to do

To **ensure compliance with internal review processes**, organizations can configure **Agent Publishing Approval Workflow** for **Copilot Studio Agents** using **Copilot Studio**.

5.1.6.3 More Information

By implementing the Agent Publishing Approval Workflow in Copilot Studio, organizations can require that all agent content goes through designated internal review steps before publication. This structured approval process helps confirm that content meets organizational standards, reduces the risk of policy violations, and prevents the release of unapproved or sensitive information.

5.1.6.4 Learn More

Learn more about **Agent Publishing Approval Workflow:**

- Key concepts - Publish and deploy your agent - Microsoft Copilot Studio | Microsoft Learn

5.1.7 Reduce risks of insecure deployments

5.1.7.1 Background

Insecure deployments expose systems and data to potential threats such as unauthorized access, data breaches, and exploitation of vulnerabilities. Without adequate security measures, attackers can compromise applications, disrupt services, or steal sensitive information, leading to possible financial losses, reputational damage, and regulatory penalties.

5.1.7.2 What to do

To **reduce risks of insecure deployments**, organizations can configure **Automatic Security Scans** for **Copilot Studio Agents** using **Copilot Studio**.

5.1.7.3 More Information

By configuring Automatic Security Scans for Copilot Studio Agents, organizations can proactively identify and address vulnerabilities before deployments go live. These scans

help detect security gaps early, ensuring that only thoroughly vetted, secure code is released, thereby minimizing the likelihood of breaches and reducing overall risk to systems and data.

5.1.7.4 Learn More

Learn more about **Automatic Security Scans:**

- Automatic security scan in Copilot Studio - Microsoft Copilot Studio | Microsoft Learn

5.2 Agent Reporting

Agent reporting involves systematically tracking and analyzing agent activities and performance, enhancing transparency and accountability. This process helps organizations identify issues, optimize resources, and supports effective governance and decision-making.

5.2.1 Enable forensic analysis and compliance audits

5.2.1.1 Background

Not enabling forensic analysis and compliance audits poses a significant risk because it limits an organization's ability to investigate security incidents, track user actions, and demonstrate regulatory compliance. Without these capabilities, uncovering the root cause of issues or proving adherence to standards becomes challenging, potentially leading to undetected breaches or non-compliance penalties.

5.2.1.2 What to do

To **enable forensic analysis and compliance audits**, organizations can enable **Audit Logs for Agent Activity** for **all agents** using **Microsoft Purview** and **Microsoft Sentinel**.

5.2.1.3 More Information

By enabling Audit Logs for Agent Activity with Microsoft Purview and Microsoft Sentinel, organizations gain detailed records of user actions and system events. These logs

provide a traceable history that supports thorough forensic investigations into security incidents and helps demonstrate compliance with regulatory requirements by documenting adherence to policies and controls.

5.2.1.4 Learn More

Learn more about **Audit Logs for Agent Activity:**

- View audit logs for admins, makers, and users of Copilot Studio - Microsoft Copilot Studio | Microsoft Learn
- Audit logs for Copilot and AI applications | Microsoft Learn

5.2.2 Support governance and lifecycle tracking

5.2.2.1 Background

Failing to support governance and lifecycle tracking can result in a lack of oversight and control over assets, data, or processes. This increases the risk of unauthorized changes, data loss, and difficulties in meeting regulatory or organizational requirements, ultimately undermining accountability and operational transparency.

5.2.2.2 What to do

To **support governance and lifecycle tracking**, organizations can use **Agent Inventory** for **all agents** using **Microsoft Sentinel** and **Power Platform Admin Center** (PPAC).

5.2.2.3 More Information

By utilizing Agent Inventory within Microsoft Sentinel and the Power Platform Admin Center, organizations gain a centralized view and control of all agents. This enables effective tracking of agent status, changes, and ownership throughout their lifecycle. As a result, organizations can maintain oversight, ensure compliance with policies and regulations, and foster accountability and transparency in their operations.

5.2.2.4 Learn More

Learn more about **Agent Inventory:**

- Manage agents for Microsoft 365 Copilot in the Microsoft 365 admin center - Microsoft 365 admin | Microsoft Learn
- View agent inventory (preview) - Power Platform | Microsoft Learn
- Search your inventory of custom agents from Copilot Studio | Microsoft Learn