The Governance and Security of Agentic Ecosystems: A Comprehensive Analysis of Microsoft Frameworks for AI Safety

The paradigm shift from generative artificial intelligence to agentic systems represents a fundamental evolution in how computational models interact with the physical and digital worlds. While early iterations of large language models functioned primarily as sophisticated text predictors, current agentic architectures are characterized by their ability to exercise autonomy, utilize external tools, and execute multi-step workflows with minimal human intervention.[1] This transition introduces a novel set of safety and security challenges that necessitate a departure from traditional cyber-defense and model-centric safety benchmarks. Microsoft has responded to this shift by developing an expansive governance ecosystem that integrates technical guardrails, ethical principles, and regulatory compliance tools across its entire stack, from the foundational Azure infrastructure to the pro-code and low-code development environments of AutoGen and Copilot Studio.[4]

Foundational Principles of the Microsoft Responsible AI Standard

The governance of AI agents at Microsoft is rooted in the Responsible AI Standard, a comprehensive internal directive that translates abstract ethical principles into concrete engineering requirements.[4] This standard is built upon six core values: fairness, reliability and safety, privacy and security, transparency, accountability, and inclusiveness.[4] For agentic systems, these principles dictate that models must not only generate safe content but must also interact with external systems in a manner that is predictable, auditable, and aligned with human intentions.[8]

The operationalization of these values involves a structured lifecycle approach categorized into four pillars: Govern, Map, Measure, and Manage.[10] The governance pillar establishes the rules for enacting responsible AI, clearly defining roles and responsibilities for development teams.[8] This includes the mandatory use of AI Impact Assessments, which help teams identify potential harms before a single line of code is written.[11] These assessments utilize specialized guides and templates provided within the Microsoft ecosystem to ensure that the unique risks of agency—such as irreversible tool actions or unintended data exfiltration—are systematically documented.[11]

The mapping and measurement phases involve a rigorous evaluation of the system's capabilities and its failure modes. Microsoft's engineering teams define and operationalize a tooling strategy that includes the Responsible AI Dashboard, which allows for the assessment of model fairness, accuracy, and explainability.[11] For agentic workflows, this measurement extends to evaluating how well an agent adheres to its defined tasks and how resilient it is to adversarial manipulation through techniques like prompt injection.[5] The final management phase focuses on continuous oversight, employing compliance mechanisms and telemetry to monitor the agent's performance in production environments.[5]

Responsible AI Principle	Agentic Application and Requirement	Primary Governance Offering
Fairness	Prevention of biased outcomes in autonomous decision-making and tool selection.	Fairlearn & Responsible AI Dashboard [4]
Reliability & Safety	Ensuring agent actions remain within predefined boundaries and avoid irreversible harm.	Task Adherence API & Groundedness Detection [13]
Privacy & Security	Protecting sensitive data accessed by agents through retrieval-augmented generation.	Microsoft Purview & Entra Agent ID [5]
Transparency	Mandating that agents identify themselves as AI and provide clear reasoning logs.	HAX Toolkit & Transparency Documents [4]
Accountability	Defining human-in-the-loop triggers for high-stakes autonomous actions.	Copilot Studio Data Policies & HITL frameworks [7]
Inclusiveness	Designing agents that support diverse languages and avoid reinforcement of stereotypes.	Inclusive Design Methodology & Multilingual Filters [11]

Strategic Governance Standards and Regulatory Alignment

As agentic systems become increasingly integrated into enterprise workflows, the need for alignment with international standards and emerging regulations has become paramount. Microsoft has achieved ISO/IEC 42001:2023 certification for its Azure AI Foundry models and Microsoft Security Copilot, marking a significant milestone in the establishment of a certifiable AI Management System (AIMS).[10] This standard provides a globally recognized framework for managing AI risks and opportunities throughout the system lifecycle, covering areas such as data management, model transparency, and human oversight.[10]

The integration of ISO 42001 with existing frameworks like the NIST AI Risk Management Framework (RMF) allows organizations to build a unified governance posture.[18] While the NIST AI RMF provides a flexible, context-specific playbook for identifying "what" and "why" AI risks are managed, ISO 42001 offers the "how"—a formal structure for implementing and auditing those management systems.[18] Microsoft facilitates this alignment by providing crosswalks and templates that help teams map their responsible AI practices to these international standards, ensuring that deployments are prepared for the stringent requirements of the EU AI Act.[10]

The Microsoft Cloud Adoption Framework for AI agents emphasizes that the most effective standards are those that integrate seamlessly into existing corporate governance rather than creating parallel compliance tracks.[9] This involves mapping AI-specific requirements to established data governance, security, and risk management policies.[19] For example, agents that handle personal data must comply with the General Data Protection Regulation (GDPR) and the EU Data Boundary, commitments that Microsoft has integrated into its standard Data Protection Addendum (DPA).[21] Organizations are encouraged to establish an AI champion network to guide peers through ethical considerations specific to their business domains, transforming abstract principles into day-to-day engineering practices.[19]

Technical Architecture for Agent Safety in Azure AI Foundry

Azure AI Foundry serves as the primary platform for building, deploying, and scaling secure agentic systems. It employs a layered "blueprint" for safety that combines identity management, real-time guardrails, and continuous evaluation.[5] This blueprint is designed to transition agents from experimental pilots to trustworthy core business systems by addressing the most common blockers to adoption: data leakage, prompt injection, and regulatory uncertainty.[5]

A central component of this architecture is the Entra Agent ID. By assigning a unique, trackable identity to every agent, organizations can gain visibility into the "sprawl" of AI agents across their tenant.[5] This allows security teams to treat agents as first-class citizens in their identity governance strategy, applying conditional access policies and monitoring behavior through the same tools used for human employees.[6] This visibility is crucial for establishing accountability and ensuring that every action taken by an autonomous system can be traced back to its origin.[5]

For technical enforcement, Azure AI Foundry integrates the Azure AI Content Safety service, which provides specialized APIs for monitoring agentic interactions.[13] The Task Adherence API is particularly significant for agentic safety, as it detects when an agent's tool use becomes misaligned with the user's intent or the agent's defined role.[13] It monitors for "premature" tool calls—actions taken before sufficient context has been gathered—and "unintended" actions that were not part of the agent's authorized execution plan.[13] This API supports input lengths of up to 100,000 characters, allowing for the analysis of complex, multi-turn conversations.[13]

Complementing this is the Groundedness Detection feature, which addresses the risk of hallucinations.[13] In an agentic context, a hallucination is not merely a factual error but a potential trigger for an incorrect tool action.[2] Groundedness Detection evaluates whether the agent's proposed response is supported by the source materials retrieved during the grounding process, preventing the system from acting on unverified or fabricated information.[13]

Safety Feature in Azure AI Foundry	**Technical Purpose and Risk Mitigation**	**Implementation Context**
Prompt Shields	Blocking direct (UPIA) and indirect (XPIA) prompt injection attacks.	Scanning prompts, files, and tool responses in real-time.[5]
Protected Material Detection	Identifying and blocking copyrighted text or proprietary code.	Preventing output of song lyrics, recipes, or licensed software.[13]
Custom Categories API	Rapid creation of filters for unique organizational harms.	Blocking specific internal project names or sensitive topics.[13]
Network Isolation	Restricting agent communication to private virtual networks.	Using custom VNETs and subnet delegation for data residency.[5]

Task Adherence API	Ensuring tool use remains aligned with instructions.	Monitoring autonomous decision gates for misalignment.[13]

Governance and Runtime Protection in Microsoft Copilot Studio

Microsoft Copilot Studio provides a low-code environment for building agents that are "secure by default".[12] The platform incorporates a range of built-in protections against agent manipulation, but it also offers advanced features for organizations with more stringent oversight requirements.[7] Central to this is the Agent Runtime Protection status, which allows developers to monitor the security health of their agents directly from the management interface.[7]

A revolutionary feature in Copilot Studio is the Advanced Real-Time Protection during agent runtime.[12] This mechanism allows the platform to integrate with external security systems—such as Microsoft Defender for Cloud Apps—to validate an agent's planned actions before they are executed.[12] When an agent intends to call a tool or access a data source, Copilot Studio sends the relevant metadata, prompt history, and tool inputs to the connected security system.[12] The external system has a strict one-second window to approve or block the action based on enterprise-wide security policies.[12] If the system identifies a threat, such as an attempt to execute a high-risk action triggered by a suspected prompt injection, it stops the agent immediately and notifies the user.[7]

Copilot Studio also empowers administrators with granular data policy controls through the Power Platform admin center.[7] Admins can govern the availability of specific agent capabilities, such as the ability to use certain knowledge sources, HTTP requests, or third-party connectors.[7] This ensures that makers can only build agents within the "guardrails" established by the IT and security teams.[7] Furthermore, environment routing can be configured to provide developers with a safe, governed space for experimentation, preventing the accidental deployment of unvetted agents into production environments.[7]

Copilot Studio Control	**Functionality for Enterprise Security**	**Management Interface**
Automatic Security Scans	Continuous validation of agent configurations against safety defaults.	Copilot Studio Authoring Canvas.[7]

Maker Security Warnings	Real-time alerts when default governance settings are modified.	Copilot Studio Agents Page.[7]
Customer-Managed Keys	Enabling encryption using organizational cryptographic assets.	Power Platform Admin Center.[7]
Data Policy Triggers	Restricting autonomous agent triggers to prevent data exfiltration.	Power Platform Admin Center.[7]
Audit Logging	Full visibility into maker activity and agent interactions.	Microsoft Purview & Sentinel.[7]

Security in Pro-Code Frameworks: AutoGen and Magentic-UI

For developers utilizing the AutoGen framework to build sophisticated multi-agent systems, safety is managed through conversation-driven collaboration and technical sandboxing.[3] AutoGen stands out for its flexible agent architecture, which allows developers to define distinct "personas" for each agent, influencing their behavior and decision-making processes through specific instructions and message-passing systems.[3]

A fundamental safety mechanism in AutoGen is the Human-in-the-Loop (HITL) framework, which allows applications to observe and control agent teams.[15] This is primarily implemented through the UserProxyAgent, a built-in component that acts as a proxy for the human user.[15] When the agent team requires feedback or approval—particularly for high-stakes or irreversible actions—the UserProxyAgent can block the execution of the team, transferring control back to the human until feedback is provided.[15] This blocking nature ensures that autonomous agents do not "run away" with a task in the absence of human supervision.[15]

Research prototypes like Magentic-UI, which are built upon AutoGen, extend these HITL concepts through "Action Guards" and "Co-planning".[26] Co-planning allows users to collaborate with agents on a step-by-step plan before execution begins, ensuring that the human and AI are aligned on the intended path.[27] Action Guards specifically target "irreversible" actions, such as closing a browser tab or clicking a button with side effects, by mandating human permission before the agent can proceed.[27] Crucially, Magentic-UI employs sandboxing for tools like browsers and code executors, ensuring that the agent's actions are isolated from the host system and cannot cause persistent or lateral damage.[27]

AutoGen Safety Mechanism	Technical Implementation and Goal	Use Case Example
UserProxyAgent	Blocking execution for human input during a run.	Seeking approval for a large financial transaction.[15]
Max Turns	Pausing the team after a set number of agent responses.	Forcing a check-in after every step of a research task.[15]
Termination Conditions	Defining explicit criteria for an agent to stop and hand off control.	Ending a session when a sensitive topic is reached.[15]
Co-Tasking	Allowing users to take and hand over control dynamically.	A user manually navigating a complex login page for the agent.[26]
Allow-Lists	Restricting agent tool access to specific domains or websites.	Limiting a web-browsing agent to official company sites.[27]

Data Governance and Risk Management via Microsoft Purview

Microsoft Purview serves as the unified platform for data security, governance, and compliance, offering specialized tools for managing the risks associated with AI agents and Copilots.[6] The Purview AI Hub provides a centralized management location to secure AI application data and proactively monitor AI usage.[14] This includes "one-click" policies that can be activated to detect risky interactions in AI apps, such as prompts that suggest unethical behavior or attempts to access restricted intellectual property.[14]

Purview’s integration with agents is built upon the principle of inherited protection. When an agent retrieves data to ground its responses, it respects the sensitivity labels and usage rights defined in Microsoft Purview Information Protection.[14] If a document is labeled with a sensitivity tag that restricts viewing rights, the agent will only surface that information to a user who already has the appropriate permissions.[21] Furthermore, Copilot and agent responses will display the highest priority sensitivity label from the data sources used during the chat, providing immediate visual feedback to the user about the confidentiality of the generated content.[7]

Purview Communication Compliance provides an additional layer of protection by detecting regulatory and conduct violations within user prompts and agent responses.[25] This includes monitoring for harassment, threats, and the sharing of sensitive information, such as social security numbers or credit card details.[25] For audit and eDiscovery purposes, Purview can capture all AI interactions in a unified audit log, preserving the history of what was asked, what was answered, and which files were accessed.[6] Organizations can set custom retention policies for these interactions, ensuring that they meet legal and corporate record-keeping requirements.[25]

Purview Governance Offering	**Specific Capability for AI Agents**	**Security Impact**
Insider Risk Management	Detecting large-scale data collection via AI agents.	Mitigating IP theft by employees.[25]
Data Loss Prevention (DLP)	Blocking the exfiltration of sensitive strings in agent output.	Preventing unauthorized data sharing.[5]
Audit Log Integration	Tracking the "who, when, and how" of AI tool usage.	Facilitating forensic investigation and compliance.[6]
One-Click AI Policies	Rapid deployment of ethical behavior monitoring.	Streamlining the security setup for new agents.[14]
Sensitivity Label Inheritance	Carrying over document protections to AI responses.	Maintaining data confidentiality at scale.[14]

Red Teaming and Adversarial Resilience: Insights from the AIRT

The Microsoft AI Red Team (AIRT) has conducted adversarial probing of more than 100 generative AI products, providing critical insights into the evolving threat landscape for agentic systems.[4] Red teaming in this context involves emulating real-world attacks against end-to-end systems, moving beyond simple model benchmarks to identify vulnerabilities in how agents interact with tools, memory, and external environments.[2]

A primary finding from the AIRT's research is that agentic AI introduces a novel set of attack vectors, such as AI Agent Context Poisoning (or memory poisoning).[28] In these attacks, malicious instructions are injected into an agent's context through untrusted inputs—such as a pre-filled prompt link or a malicious email—that instruct the agent to "remember" the attacker's content as a trusted source.[30] This can lead to persistent manipulation of the agent's behavior across multiple sessions, allowing the attacker to establish themselves as an "authoritative" source for future citations or tool actions.[30] The AIRT has identified specific indicators of compromise for these attacks, including URL query parameters containing keywords like "memory," "trusted," and "authoritative".[30]

Another critical lesson is that simple attacks are often as effective as complex ones.[2] Hand-crafted prompts and "fuzzing" techniques frequently uncover vulnerabilities in end-to-end systems that automated benchmarks miss.[2] However, to achieve necessary scale, Microsoft employs automation through the open-source PyRIT (Python Risk Identification Tool) framework.[2] PyRIT allows red teams to orchestrate thousands of adversarial probes against a target system, evaluating responses for risks such as sensitive data leakage from internal knowledge bases.[2]

Lesson from AI Red Teaming	**Strategic Implication for Security Teams**	**Recommended Action**
Understand system capabilities first.	Vulnerabilities are highly dependent on the application context.	Tailor red teaming scenarios to specific agent tools.[2]
Simple attacks often suffice.	Human creativity can bypass complex model filters.	Combine automated probing with manual red teaming.[2]
Red teaming is not benchmarking.	Static metrics fail to capture novel, evolving harms.	Define new harm categories based on real-world usage.[2]

Leverage automation for scale.	Manual testing cannot cover the entire input space.	Use PyRIT and the AI Red Teaming Agent for high-volume tests.[2]
The human element is crucial.	Nuanced harms require cultural and emotional intelligence.	Involve subject matter experts in safety evaluations.[2]
AI amplifies existing security risks.	Traditional flaws like improper error handling still matter.	Maintain standard security hygiene alongside AI safety.[2]
Securing AI is an ongoing process.	Safety requires a "break-fix" cycle of continuous updates.	Schedule regular audits and update filters frequently.[2]
RAI risks are pervasive but complex.	Issues like bias are subjective and hard to measure.	Guard against both intentional and accidental harms.[2]

Legal Safeguards and the Customer Copyright Commitment

Intellectual property concerns represent a significant hurdle for many organizations exploring the use of AI agents for content creation or code generation. Microsoft addresses these concerns through the Customer Copyright Commitment, a pledge to assume legal responsibility for potential copyright claims arising from the use of its commercial AI services.[32]

This commitment extends Microsoft's existing intellectual property indemnity to cover claims related to the output generated by Copilot services and the Azure OpenAI Service.[32] If a third party sues a commercial customer for copyright infringement for using these tools, Microsoft will defend the customer and pay for any adverse judgments or settlements.[32] This protection is intended to reassure customers that the copyright liability of using Microsoft's AI products remains with the provider, not the user.[32]

To qualify for this commitment, customers must adhere to a "shared responsibility" model.[32] They are required to use the built-in content filters and other safety systems—such as Protected Material Detection—to prevent the generation of infringing materials.[32] Furthermore, customers must not intentionally provide inputs to the agent that they do not have the legal rights to use.[32] This legal shield is effectively a partnership where Microsoft provides the technical mitigations and the customer provides the ethical oversight of their own inputs.[32]

Feature of Copyright Commitment	**Coverage and Eligibility Details**	**Key Exclusions**
Broad IP Indemnity	Covers copyright, patent, trademark, and trade secrets.	Free products and consumer-facing AI services.[33]
Adverse Judgment Payment	Microsoft pays settlements or court-ordered damages.	Claims where the user bypassed safety filters.[32]
Global Application	Effective starting October 1, 2023, for paid commercial versions.	Modifications made to the output by the customer.[33]
Tool Integration	Requires use of Protected Material Detection guardrails.	Inputs that the user knew were infringing.[32]

Operationalizing Agent Safety: Case Studies and Real-World Implementation

The practical application of these governance tools is evident across various industries where organizations are deploying Microsoft-powered agents to solve complex business problems. These real-world cases demonstrate that safety is not a secondary consideration but a foundational element that enables business impact.

Financial Services and Customer Engagement

Raiffeisen Bank International (RBI) utilized the Azure OpenAI Service to launch its own internal version of ChatGPT, focusing on improving productivity while adhering to strict banking regulations.[11] By deploying within the Azure service boundary, RBI ensured that its prompts and data remained isolated from public models, a core requirement for financial institutions.[11] Similarly, the AI answer engine Perplexity.AI leveraged Azure AI Studio to double its throughput while cutting costs, using the platform's built-in safety filters to manage the risks of ungrounded or toxic outputs.[11]

Software Development and Creative Industries

Unity, a leader in the gaming industry, built a developer assistant with Azure OpenAI to streamline troubleshooting for its global community.[11] By integrating safety into the developer assistant, Unity ensured that the AI provided reliable technical guidance without violating the intellectual property of third-party developers.[11] The use of Protected Material Detection is particularly relevant in these scenarios to prevent the accidental leakage of proprietary code structures.[13]

Social Impact and Healthcare

In the social impact sector, CARE relies on Azure AI to prepare for global emergencies, using AI insights to analyze data from diverse sources.[11] The reliability of these insights is paramount, as incorrect predictions could have life-altering consequences.[11] In India, IWill Therapy and IWill CARE use Azure-based Hindi-speaking bots to expand access to mental health services.[11] This deployment emphasizes the importance of multilingual safety models, as the system must accurately detect self-harm or violent intent across multiple languages and cultural contexts.[11]

Sustainability and Operations

PwC has integrated Microsoft Sustainability Manager with other AI-driven solutions to help clients drive sustainable growth.[11] These agentic systems analyze vast amounts of supply chain data to identify opportunities for carbon reduction.[11] The governance of these agents involves ensuring that the data used for sustainability reporting is accurate and auditable, a task supported by the transparency and logging features of the Microsoft Cloud.[8]

Organization	**AI Agent Use Case**	**Key Safety/Governance Tool Used**
Raiffeisen Bank	Internal ChatGPT for productivity.	Azure OpenAI Service Boundary & Data Isolation.[11]
IWill Therapy	Multilingual mental health support bot.	Azure AI Content Safety (Multilingual).[11]
CARE	Emergency readiness and global data analysis.	Groundedness Detection & Reliable AI.[11]

Unity	Developer troubleshooting assistant.	Protected Material Detection for Code.[11]
PwC	Sustainability and supply chain analysis.	Microsoft Purview for Data Auditability.[11]

Identifying and Mitigating Emerging Risks in Agentic Systems

The transition from isolated LLM chat interfaces to integrated agents that "live" within the enterprise data environment has created new risks related to oversharing and insider threats.[6] Microsoft telemetry indicates that "oversharing" is one of the primary concerns for IT leaders deploying AI.[6] This occurs when an agent—acting with the permissions of a user—retrieves sensitive data that the user may have had access to but should not necessarily be feeding into an AI reasoning loop.[6]

To mitigate this, the Copilot Control System provides reports and dashboards that help IT teams identify content that has been shared too broadly.[6] For example, if a "Confidential" file is accidentally shared with the "All Employees" group, an agent could potentially surface that information in response to a seemingly benign query.[6] By applying Purview policy recommendations, organizations can fix these permissions before the agent can access the sensitive files.[6]

Furthermore, the risk of "Recommendation Poisoning" highlights a shift in attacker tactics.[30] In a recommendation poisoning scenario, an attacker might compromise a website or a public document that an agent is likely to "browse" for grounding.[30] By embedding malicious instructions in that source, the attacker can manipulate the agent's future recommendations or "poison" its memory.[30] This necessitates a strategy of "zero trust for AI," where every input—even from supposedly trusted internal or grounded sources—is treated as potentially malicious and scanned by real-time classifiers like Prompt Shields.[5]

Emerging Agentic Risk	**Mechanism of Exploitation**	**Mitigation Strategy**
Recommendation Poisoning	Malicious data grounding sources.	Continuous scanning of tool responses via Prompt Shields.[5]

Identity Sprawl	Unmanaged "shadow" agents in the tenant.	Mandating Entra Agent ID for all deployments.[5]
Tool Over-Privilege	Agents having R/W access to sensitive APIs.	Least-privilege credential management and action guards.[7]
Context Injection	Malicious instructions hidden in emails or files.	Cross-Prompt Injection (XPIA) Classifiers.[5]
Latent Bias	Agents reinforcing stereotypes through tool choice.	Regular fairness audits via RAI Dashboard.[11]

Operationalizing Governance: The Agent Success Kit and Adoption Framework

To assist organizations in navigating this complex landscape, Microsoft provides an "Agent Success Kit".[1] This resource is designed to help IT administrators prepare their tenants for AI agents and enable users to create and use them responsibly.[1] The kit includes technical readiness guides, user enablement materials, and scenarios for common business use cases.[1]

The adoption process follows a structured path:

1. **Assign Identity:** Every agent must be known and tracked using Entra Agent IDs to prevent unmanaged sprawl.[5]
2. **Establish Built-in Controls:** Organizations should implement Prompt Shields, groundedness checks, and harm filters as a standard baseline for all agents.[5]
3. **Continuous Evaluation:** Red teaming and automated safety evaluations should be performed both before deployment and throughout the agent's production life.[2]
4. **Protect Sensitive Data:** Applying Purview labels and DLP ensures that data protection policies are honored by agentic outputs.[5]
5. **Monitor with Enterprise Tools:** Telemetry should be streamed into existing security operations tools like Microsoft Defender XDR and Sentinel for unified investigation and response.[5]
6. **Regulatory Alignment:** Organizations must map their technical metrics and governance decisions to international standards like the NIST AI RMF and ISO 42001 to ensure legal compliance.[5]

By following this blueprint, organizations can extract the significant productivity benefits of agentic AI—such as reducing phishing triage time from 30 minutes to 3 minutes—while maintaining a robust defense against the evolving threat landscape.[35] The governance of AI agents is ultimately about building a system of trust, where technology, policy, and human oversight work in concert to ensure that artificial agency remains a force for organizational innovation and social good.

Works cited

1. Microsoft Copilot Studio, accessed February 11, 2026, https://adoption.microsoft.com/en-us/ai-agents/copilot-studio/
2. Enhancing AI safety: Insights and lessons from red teaming | The ..., accessed February 11, 2026, https://www.microsoft.com/en-us/microsoft-cloud/blog/2025/01/14/enhancing-ai-safety-insights-and-lessons-from-red-teaming/
3. A Developer's Guide to the AutoGen AI Agent Framework - The New Stack, accessed February 11, 2026, https://thenewstack.io/a-developers-guide-to-the-autogen-ai-agent-framework/
4. Responsible AI: Ethical policies and practices | Microsoft AI, accessed February 11, 2026, https://www.microsoft.com/en-us/ai/responsible-ai
5. Agent Factory: Creating a blueprint for safe and secure AI agents ..., accessed February 11, 2026, https://azure.microsoft.com/en-us/blog/agent-factory-creating-a-blueprint-for-safe-and-secure-ai-agents/
6. Copilot Control System for Enterprise-Ready AI - Microsoft, accessed February 11, 2026, https://www.microsoft.com/en-us/microsoft-365-copilot/copilot-control-system
7. Security and governance - Microsoft Copilot Studio | Microsoft Learn, accessed February 11, 2026, https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance
8. Responsible AI Principles and Approach | Microsoft AI, accessed February 11, 2026, https://www.microsoft.com/en-us/ai/principles-and-approach
9. Governance and security for AI agents across the organization - Cloud Adoption Framework | Microsoft Learn, accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/ai-agents/governance-security-across-organization
10. Microsoft Azure AI Foundry Models and Microsoft Security Copilot achieve ISO/IEC 42001:2023 certification, accessed February 11, 2026, https://azure.microsoft.com/en-us/blog/microsoft-azure-ai-foundry-models-and-microsoft-security-copilot-achieve-iso-iec-420012023-certification/
11. Responsible AI Tools and Practices - Microsoft, accessed February 11, 2026, https://www.microsoft.com/en-us/ai/tools-practices
12. Strengthen agent security with real-time protection in Microsoft Copilot Studio, accessed February 11, 2026, https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/strengthen-agent-security-with-near-real-time-protection-in-microsoft-copilot-studio/

13. What is Azure AI Content Safety? - Azure AI services | Microsoft Learn, accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview
14. Microsoft Purview and Copilot: The perfect union to guarantee the security of your company, accessed February 11, 2026, https://www.plainconcepts.com/microsoft-purview-copilot-2/
15. Human-in-the-Loop — AutoGen - Microsoft Open Source, accessed February 11, 2026, https://microsoft.github.io/autogen/stable//user-guide/agentchat-user-guide/tutorial/human-in-the-loop.html
16. Content Safety in Foundry Control Plane | Microsoft Azure, accessed February 11, 2026, https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety
17. ISO 42001: Auditing and Implementing Framework | CSA - Cloud Security Alliance, accessed February 11, 2026, https://cloudsecurityalliance.org/blog/2025/05/08/iso-42001-lessons-learned-from-auditing-and-implementing-the-framework
18. Integrating the NIST AI RMF and ISO 42001: A Practical Guide - FairNow, accessed February 11, 2026, https://fairnow.ai/map-nist-ai-rmf-iso-42001/
19. Establishing Responsible AI Policies for AI Agents across ..., accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/ai-agents/responsible-ai-across-organization
20. AI Policy Template (June 2024) - AI Governance Library, accessed February 11, 2026, https://www.aigl.blog/ai-policy-template-june-2024/
21. Data, Privacy, and Security for Microsoft 365 Copilot, accessed February 11, 2026, https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-privacy
22. Enterprise data protection in Microsoft 365 Copilot and Microsoft 365 Copilot Chat, accessed February 11, 2026, https://learn.microsoft.com/en-us/copilot/microsoft-365/enterprise-data-protection
23. Microsoft Security Copilot, accessed February 11, 2026, https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot
24. Data, privacy, and security for Azure AI Agent Service - Foundry Tools | Microsoft Learn, accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/ai-foundry/responsible-ai/agents/data-privacy-security?view=foundry-classic
25. Use Microsoft Purview to manage data security & compliance for Microsoft Copilot Studio, accessed February 11, 2026, https://learn.microsoft.com/en-us/purview/ai-copilot-studio
26. Magentic-UI: Towards Human-in-the-loop Agentic Systems - Microsoft, accessed February 11, 2026, https://www.microsoft.com/en-us/research/wp-content/uploads/2025/07/magentic-ui-report.pdf

27. Magentic-UI, an experimental human-centered web agent - Microsoft Research, accessed February 11, 2026, https://www.microsoft.com/en-us/research/blog/magentic-ui-an-experimental-human-centered-web-agent/
28. Lessons From Red Teaming 100 Generative AI Products - arXiv, accessed February 11, 2026, https://arxiv.org/pdf/2501.07238
29. Planning red teaming for large language models (LLMs) and their applications - Azure OpenAI in Microsoft Foundry Models, accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/red-teaming?view=foundry-classic
30. Manipulating AI memory for profit: The rise of AI Recommendation Poisoning, accessed February 11, 2026, https://www.microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/
31. AI Red Teaming Agent (preview) - Microsoft Foundry, accessed February 11, 2026, https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/ai-red-teaming-agent?view=foundry-classic
32. Microsoft announces new Copilot Copyright Commitment for customers, accessed February 11, 2026, https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/
33. Introducing the Microsoft Copilot Copyright Commitment, accessed February 11, 2026, https://techcommunity.microsoft.com/discussions/businessapplicationsforpartners/introducing-the-microsoft-copilot-copyright-commitment/3922303
34. Incident Analysis for AI Agents - AAAI Publications, accessed February 11, 2026, https://ojs.aaai.org/index.php/AIES/article/download/36596/38734/40671
35. Agentic AI Security Guide | Security Insider - Microsoft, accessed February 11, 2026, https://www.microsoft.com/en-us/security/security-insider/emerging-trends/agentic-ai-for-smarter-security