Microsoft

Keep Pace

Freddy Dezeure
DCISO for Europe

European digital commitments

1. We will help build a broad AI and cloud ecosystem across Europe.
2. We will uphold Europe's digital resilience even when there is geopolitical volatility.
3. We will continue to protect the privacy of European data.
4. We will always help protect and defend Europe's cybersecurity.
5. We will help strengthen Europe's economic competitiveness, including for open source.

Committed to Europe

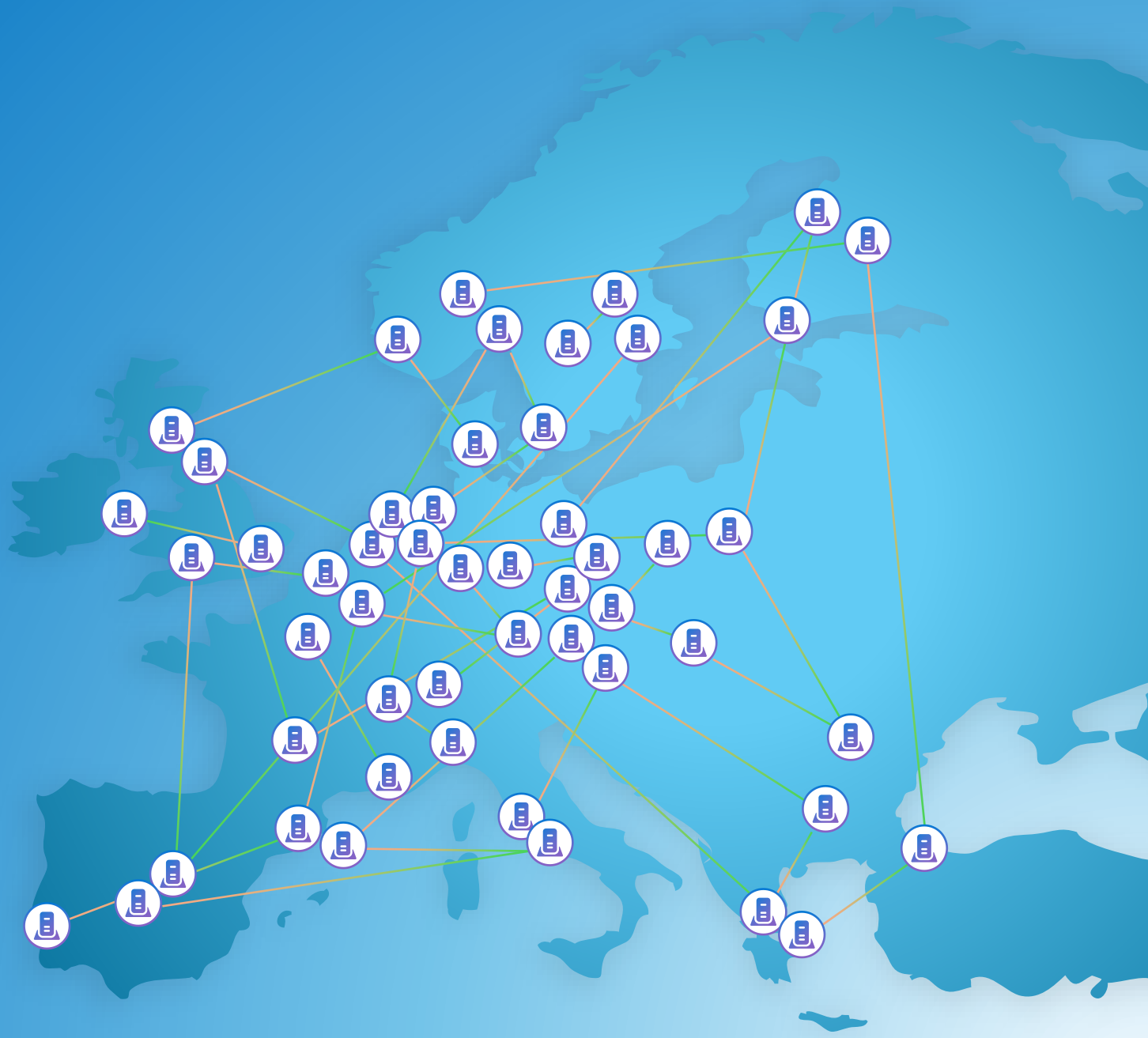
Microsoft Ireland Operations Limited Board

Significant DC footprint

Regionally-specific services

EU data boundary

EU-specific contract terms



COMMITMENT

4

We will always help protect and defend Europe’s cybersecurity.

Deputy CISO for Europe

About my role

A Protect MSFT and its customers

B Comply with European regulations

C Secure by default

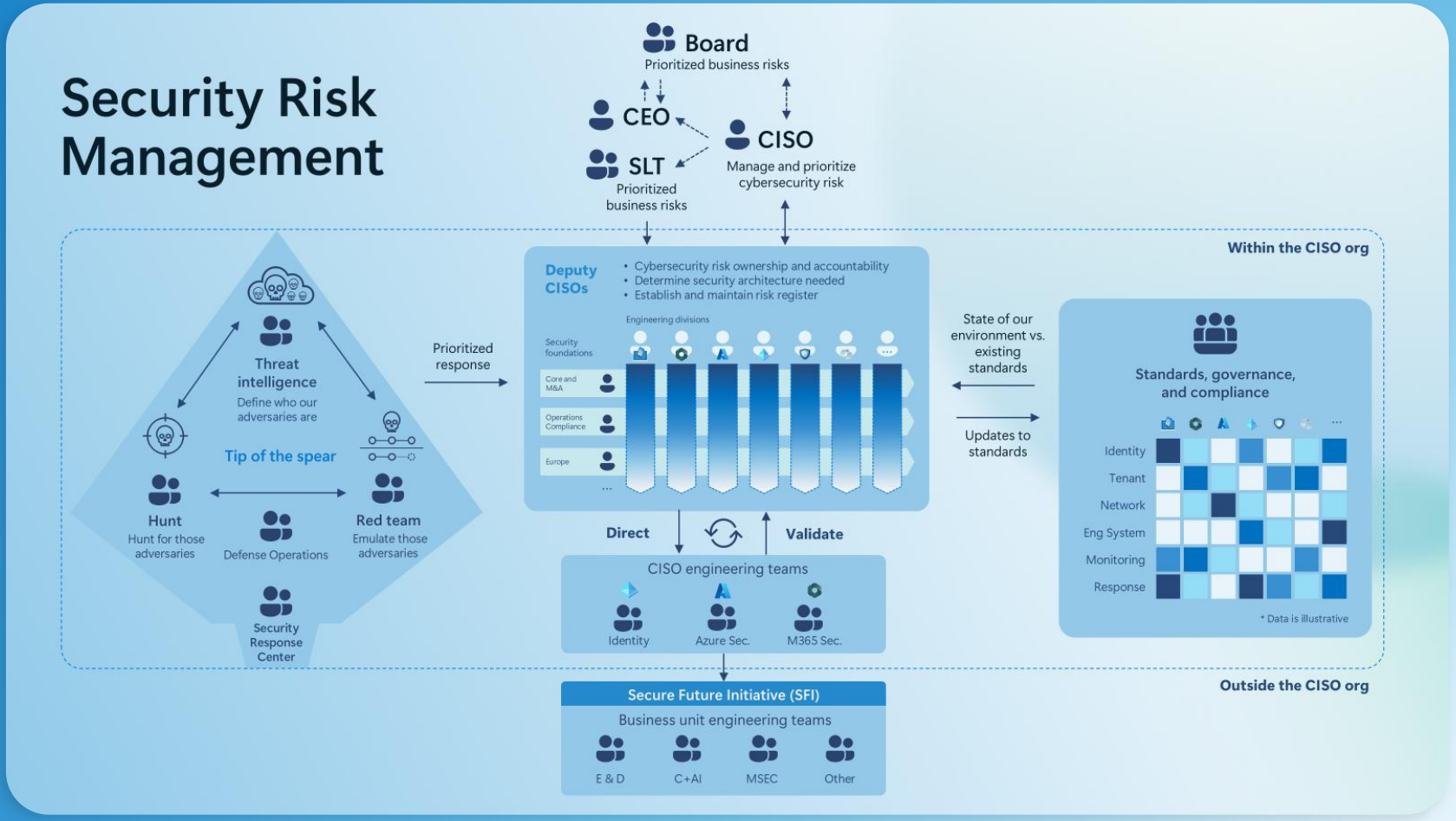
Agenda

1. How Microsoft protects itself
   Active defense - at world scale
2. Frontier models
   Insights for patching
3. AI Agents at work
   Train and contain your agents
4. Incident response
   At pace
5. Defense in depth
   The basics still matter

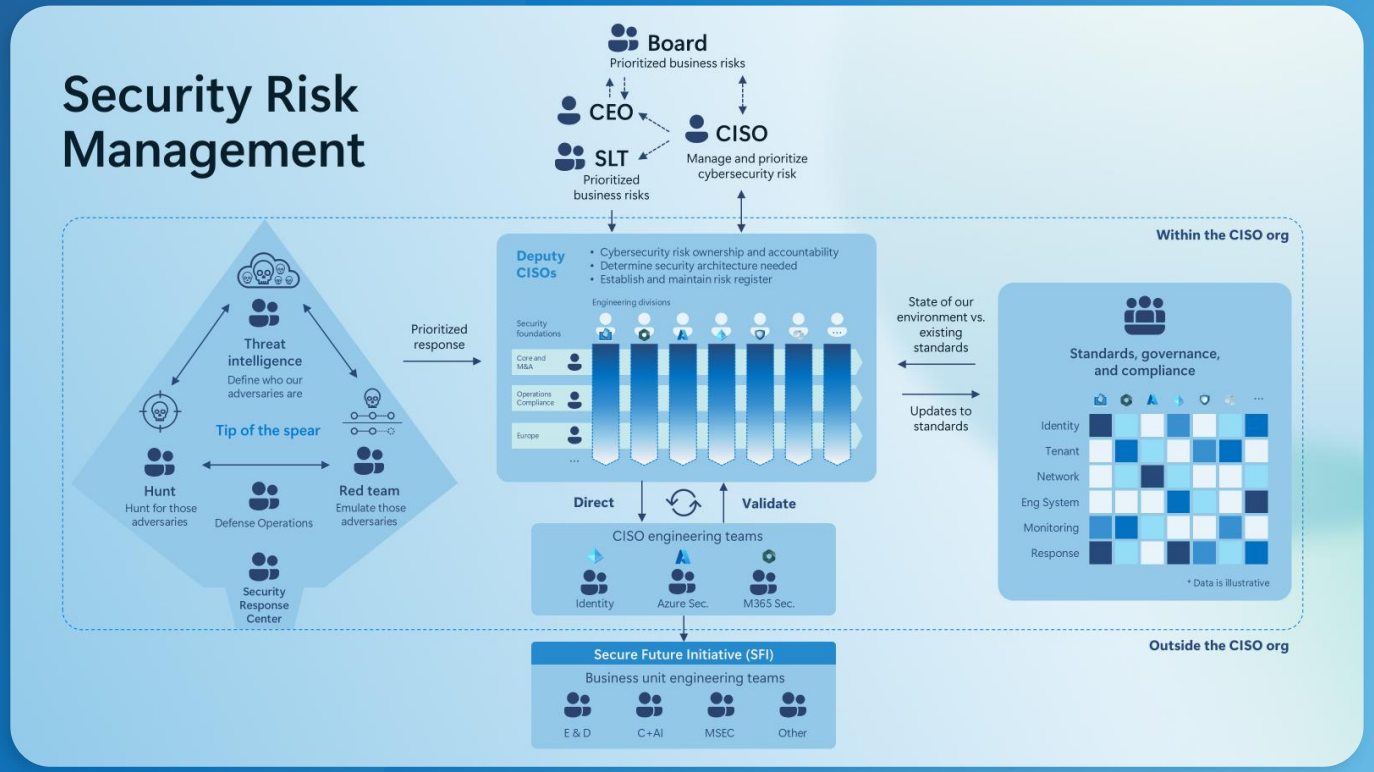
How Microsoft defends itself
Active defense - at world scale

CISO Org - Impactful structure and operation

- Cross-function expertise and decision making
- Quasi-real-time threat-informed defense
- Visibility and impact at scale
- Durable improvements in infrastructure and services



Active defense at scale

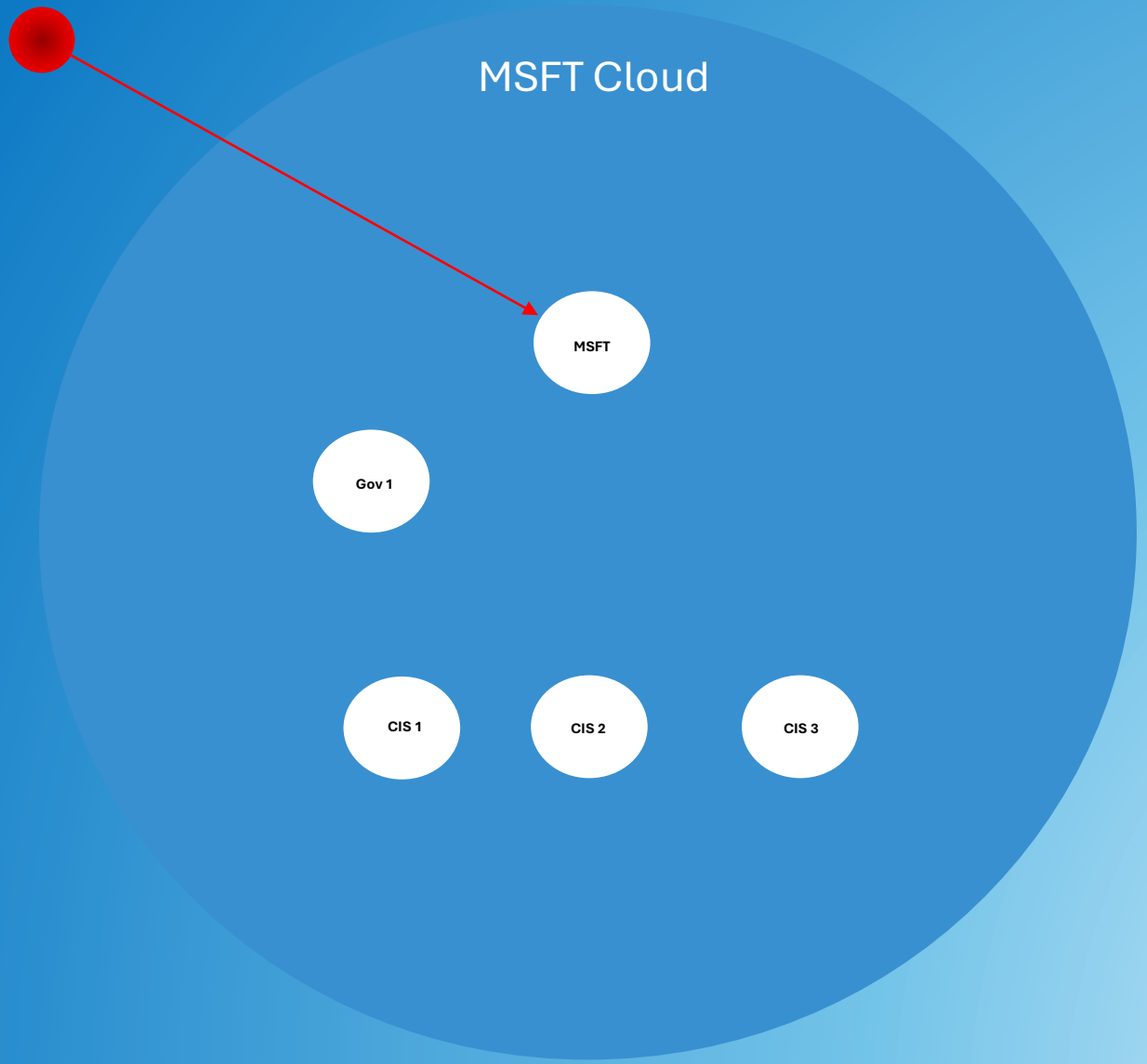


Microsoft tenants

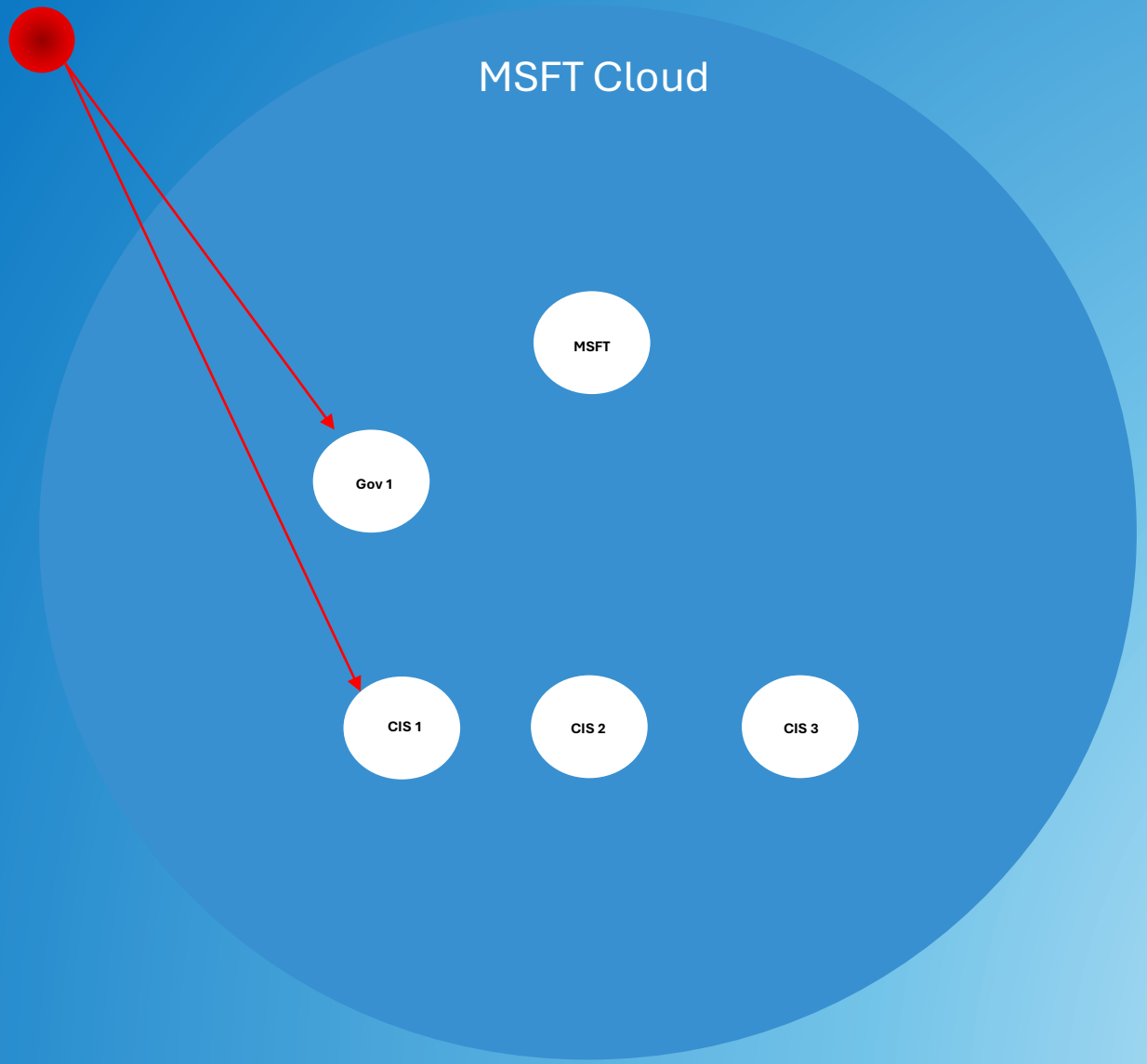
Customer tenants

Microsoft infrastructure and services

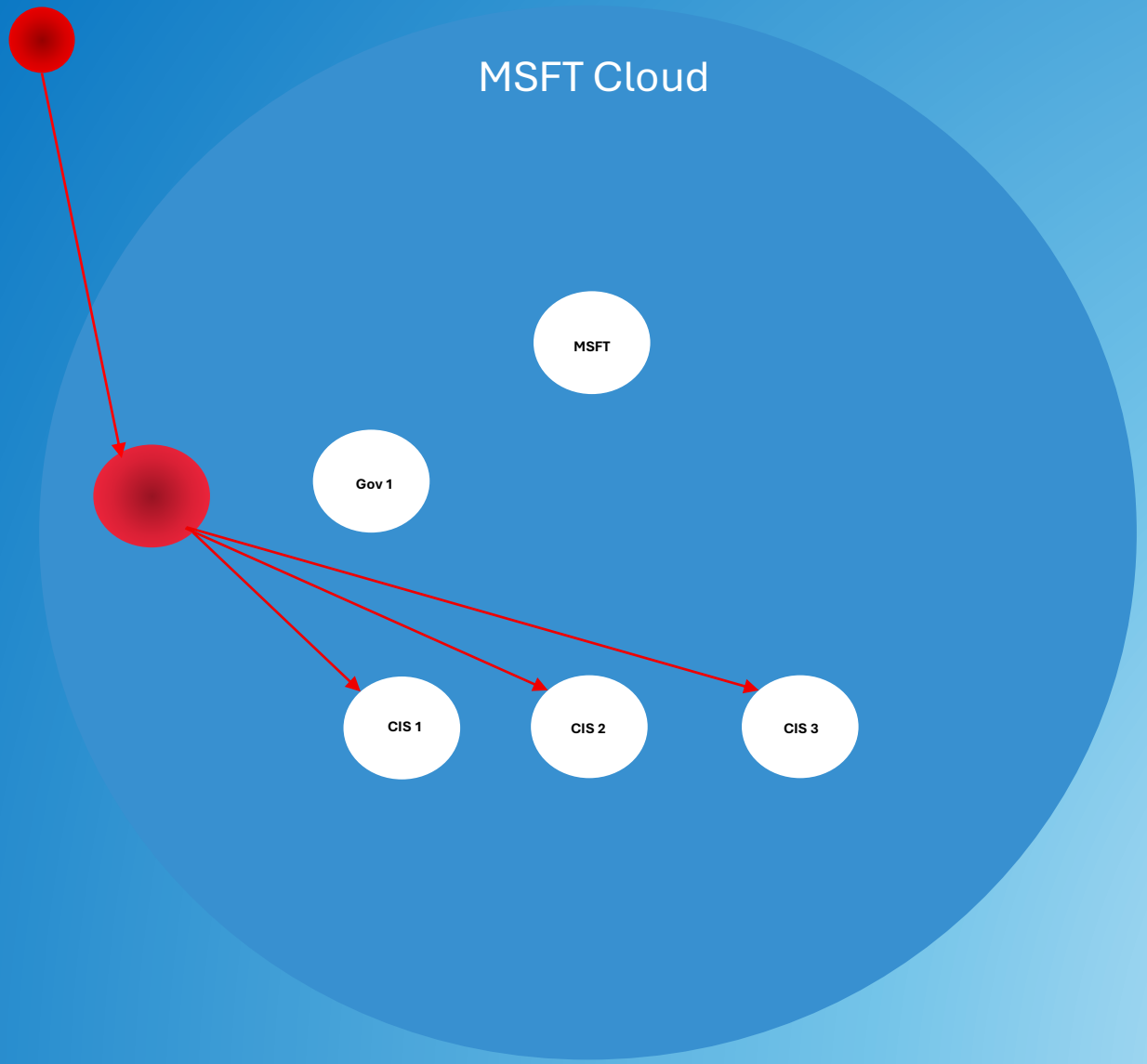
Hunting at scale



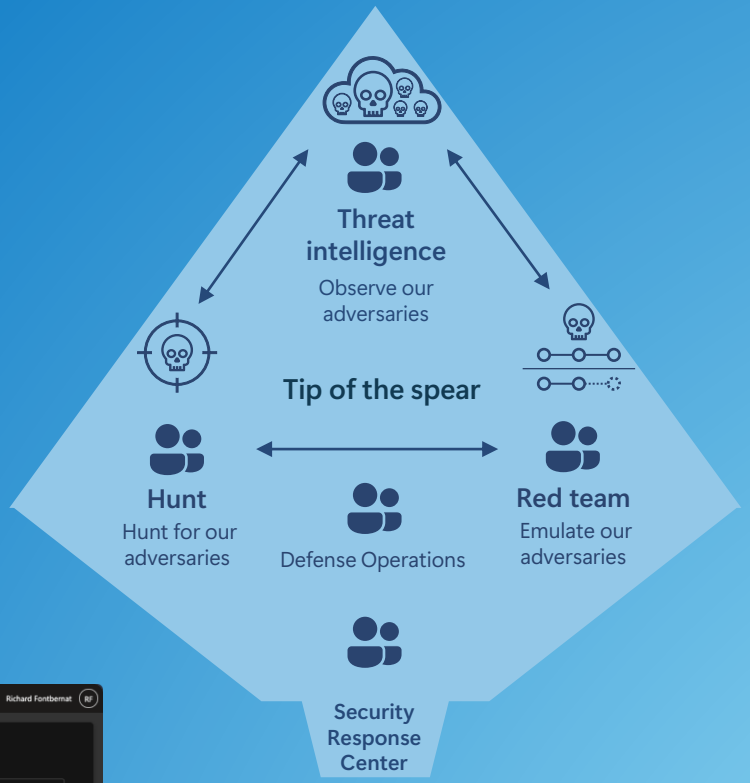
Hunting at scale



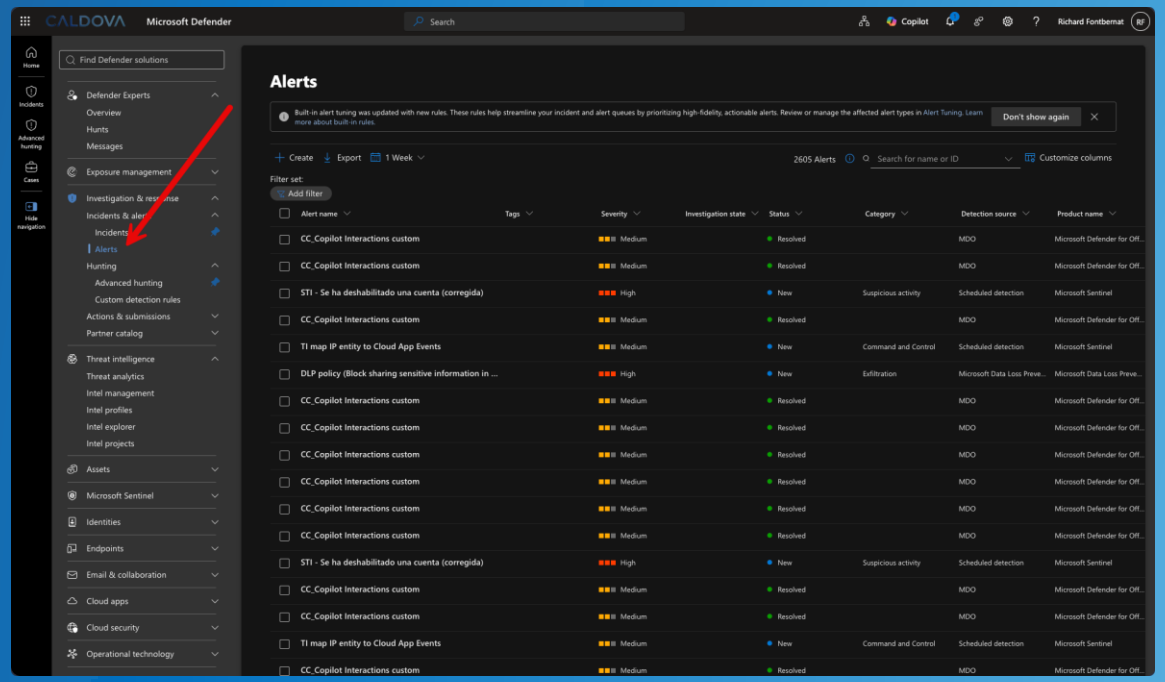
Hunting at scale



NSNs by Microsoft



Customer tenants

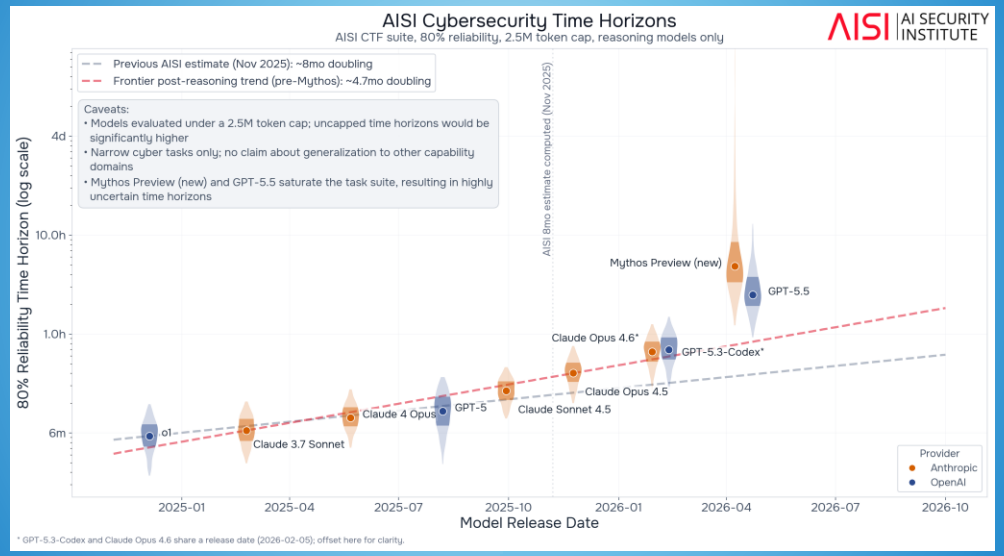


Delivered in Defender XDR

Act immediately if
Title contains “Microsoft Threat Intel:"
ID begins with “ns"

Frontier Models
Insights for patching

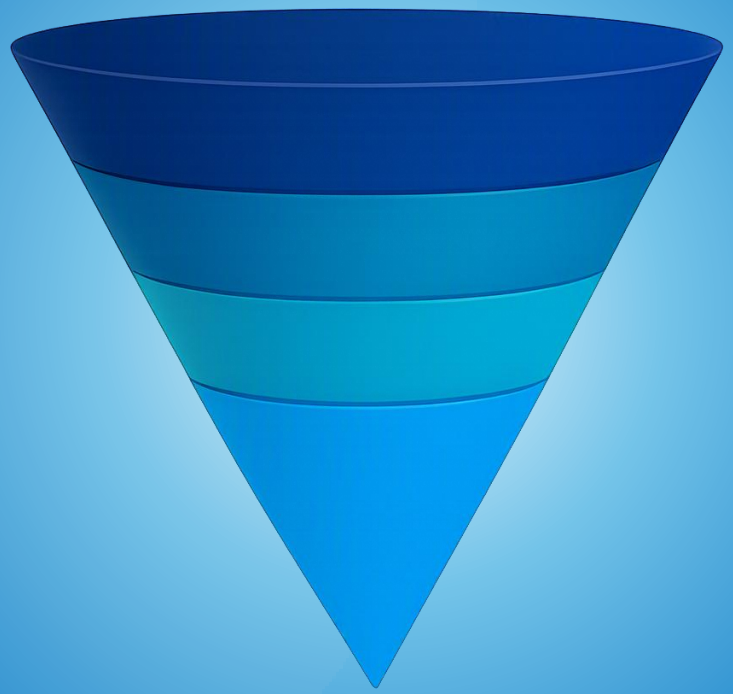
Frontier AI Models



- Finding vulnerabilities in code, faster*
- Finding exploits, faster (even based on binaries)
- Allowing autonomous attacks**
- Continuously improving
- Accessible to our adversaries

* How fast is autonomous AI cyber capability advancing? | AISI Work

** jadepuffer-agentic-ransomware-for-automated-database-extortion



Glasswing insights

- Non-deterministic, multiple runs required
- Limited agreement between models
- Continuous scanning of old and new code
- “Harnesses” matter, a lot
- Large number of findings
- Triage of false positives, severity and exposure
- Large impact on resources (FTEs, tokens)

What can you do?

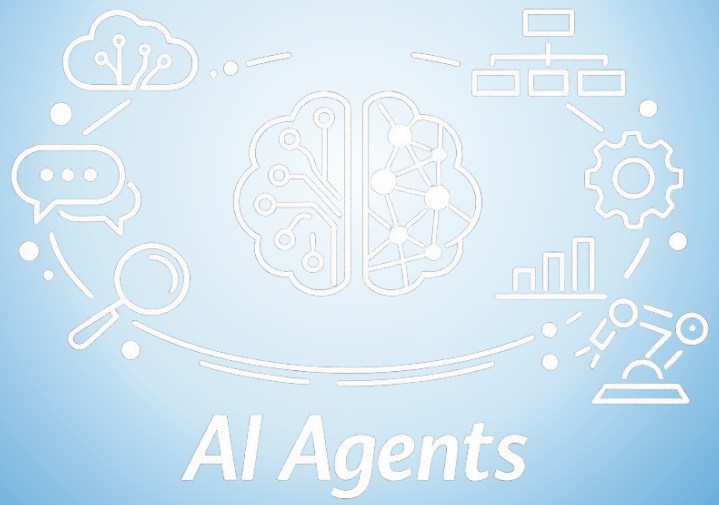
- Scan your own software
  - Inventory of assets + SBOM
  - Don’t wait, use a harness + available models
  - Prepare for a continuous process
- Patch overnight
  - Allocate resources, change your policies, measure, report*
  - Prioritize identity management, edge devices
  - Can we break production to patch our systems?
- Limit exposure
  - Replace legacy software if you can
  - Isolate if you can’t
  - Defend in depth**

* BOD 26-04: Prioritizing Security Updates Based on Risk | CISA

** https://sip.security.microsoft.com/securenow

AI Agents at work
Train and contain your agents

Microsoft
DCISO
for Europe

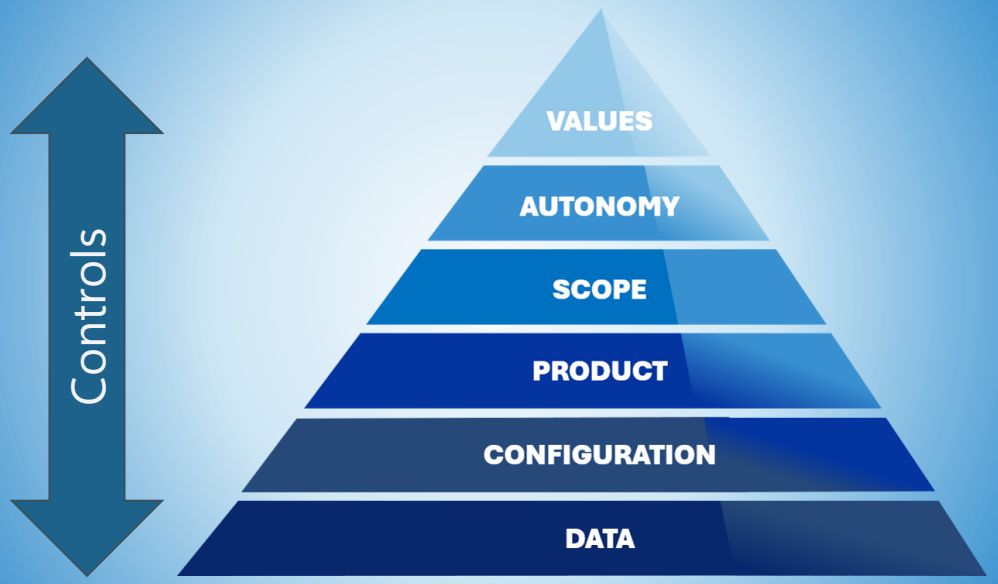


AI agents

- Democratization of development – fast adoption
- Gaps in our understanding of Agent behaviour
  - Black boxes
  - Non deterministic
  - Autonomous
  - Agent-to-agent communication/cooperation
  - Pace of execution
- Lack of “social systems”

Our control frameworks

- Product: product security
  - OWASP AI
  - MITRE ATLAS
- Configuration: enterprise security
  - NIST CSF, NIST 800-53, ISO 27001
  - MITRE ATT&CK
  - CIS Baselines
- Update frequency not at pace
- Coverage not sufficient



Proposed reference model

- Values
  - Should AI agents respect the corporate values?
- Autonomy
  - Can AI agents modify, delete data?
  - Can they carry out financial transactions and to which threshold?
  - Are they allowed to take initiative?
  - Can they launch physical action?
  - Is a human-in-the-loop required for critical actions?
  - Do we need a “kill switch”?
  - Can we roll back their actions?
- Scope
  - To what data, services, systems the agents have access?
  - Do AI agents have access to the internet and to other agents?
  - What are their primary role and success factors?
- Data layer is the foundation
  - Provenance, lineage, classification, integrity
- How do we train and contain our AI agents?

Control Objectives expanded to AI

Principles

Outcomes

Asset Management

- An up-to-date centralized inventory of all identities (human, service and AI).
- An up-to-date centralized inventory of all models by approved entity, including their applicable use cases.
- An up-to-date centralized inventory of agents including roles & responsibilities and owners.
- An up-to-date centralized inventory of primary data assets and source of record.
- An up-to-date centralized inventory of HW and SW assets, tenants, services, and applications.
- An up-to-date centralized inventory of cryptographic keys their expiration date and where they are used.
- An up-to-date centralized inventory of exceptions.
- AI Bill of Materials (AI-BOM) is maintained.

Identity & Access Management

- Interactive logins use phishing-resistant MFA or approved token protocols + AI model or AI Agent equivalent.
- Non-interactive logins use modern authentication methods + AI model or AI Agent equivalent.
- Implement and manage break glass accounts in a secure manner.
- Access is protected with controls that account for contextual risk.
- Single-Sign-On is enforced.
- Identity/passwords are not saved in the browser - enforce use of password managers.
- Access tokens are temporal and live for the minimum amount of time required.
- Access is managed during the Joiners/Leavers/Movers process (humans and AI Agents).
- Kill switch to terminate agents
- Identities and entitlements are pruned to remove unnecessary and dormant access.
- Protect against insider risk (human and AI Agents).

Least Privilege

- Admins/users/guests/agents are managed differentiated and provided least privileges.
- Minimum privileges are provided to applications/ scripts / services.
- Entitlements granting/removing/reviews are managed appropriately.
- Role based access to actions, systems, services, activities (for humans and AI Agents).
- Enforce human-in-the-loop enforced for critical Agent decisions.

Data Protection

- Data classification, provenance, lineage.
- Role based access to data (read, extract, modify) for humans and Agents.
- Data loss prevention.
- Prevent AI data poisoning.
- Backup and recovery (data, models, Agents).
- Validate training data quality, representativeness, and absence of harmful bias.
- Prevent sensitive data inference and memorization.

Encryption & Key Management

- Encryption at rest using secure levels of ciphers.
- Encryption in transit (SSL, TLS) using secure levels of ciphers.
- Key rotation.
- Minimize key access (HSM/customer managed keys/key vaults).

Network Security

- Close access to internet by default / open by decision.
- Control and monitor inbound/outbound network access for agents.
- Secure communications with CSP services requiring external access (RDP, SSH…).
- Secure communications with MCP servers.
- Segmentation, zoning.

Control Objectives expanded to AI

Principles

Outcomes

Vulnerability Management

- Continuous vulnerability management of products/cloud services/agents.
- Identify and mitigate risk of exposed attack vectors, open internet surfaces.
- Protect the integrity of the AI models, agents and their training.

Application Management

- Restrictions on installation of applications/services/agents.
- Scan and remove applications / services / agents during lifetime.
- Restrictions on use of applications/services/agents (Only execute authorized code).
- Services segmentation.
- Control the authorized and appropriate use of MCP tools.
- Mitigate AI input/output risks and monitor for anomalous behavior.
- Verify provenance and trustworthiness of AI models/libraries before deployment.
- Secure AI training and build environments by applying least privilege and network isolation to ML pipelines and integrating security checkpoints and automated security testing into AI development workflows.
- Conduct AI-focused threat modeling for AI systems, identifying AI-specific attack vectors (prompt injection, data poisoning, model theft, agent abuse).
- Detect and manage AI behavior drift, reground model if drifted.

Malware Defenses

- Model and Agent malware detection.
- Cloud infrastructure malware detection.
- Network (Telemetry) - protect user infrastructure using CSP telemetry.

Logging & Alerting

- Receive security alerts from the CSP.
- Automated configuration monitoring — detect deviations from secure defaults or policies (drift avoidance).
- Gather sufficient information from the infrastructure to detect / respond to anomalous behaviors.
- Gather information from alerts / anomalies and respond via automation.
- Establish AI-specific incident response plans covering model failure, harmful outputs, and bias incidents.

Awareness & Training

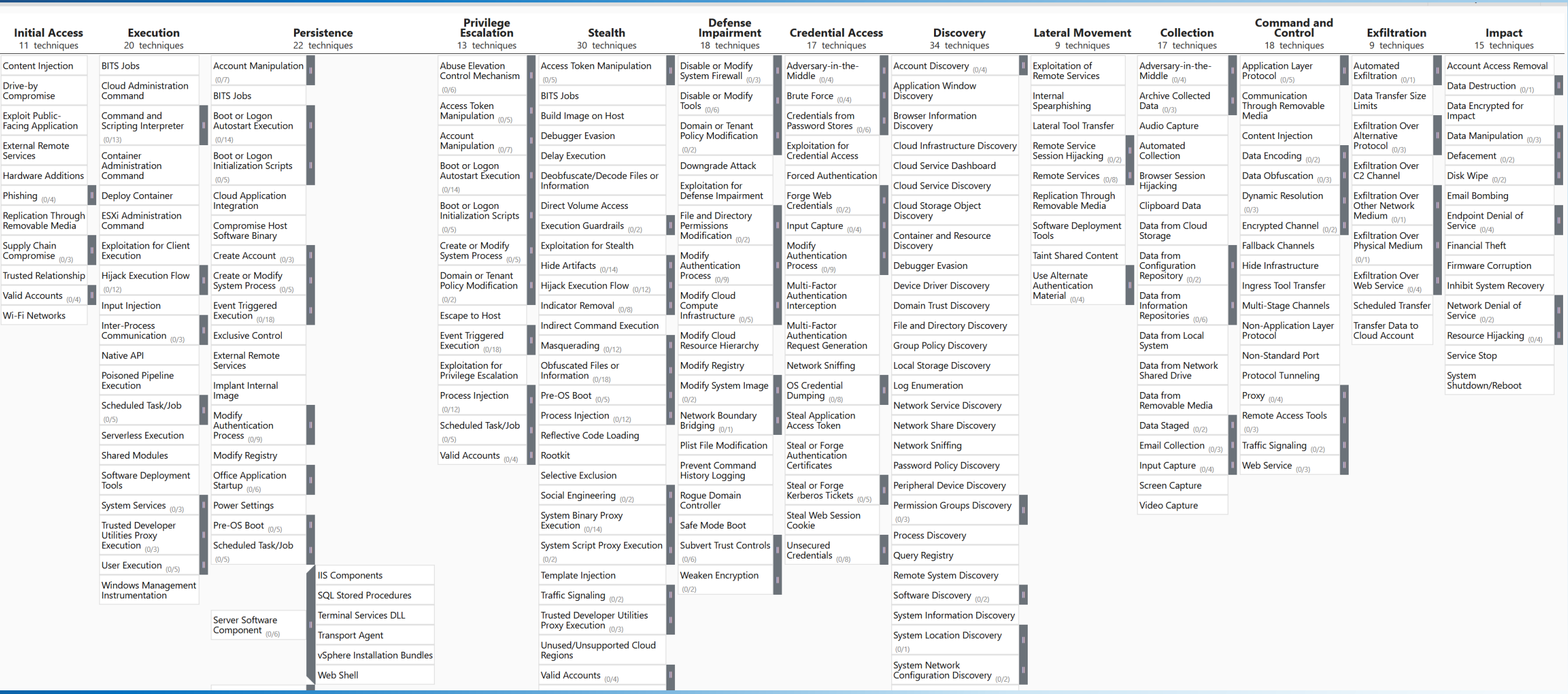
- Training humans in the use of AI and AI Agents, also in their role as "Human in the Loop".
- Train humans and AI Agents to recognize social engineering attacks.
- Train humans and AI Agents in corporate values and responsible AI.
- Train agents in their role and autonomy.
- Train incident responders on AI-specific incidents.

Scope & Autonomy

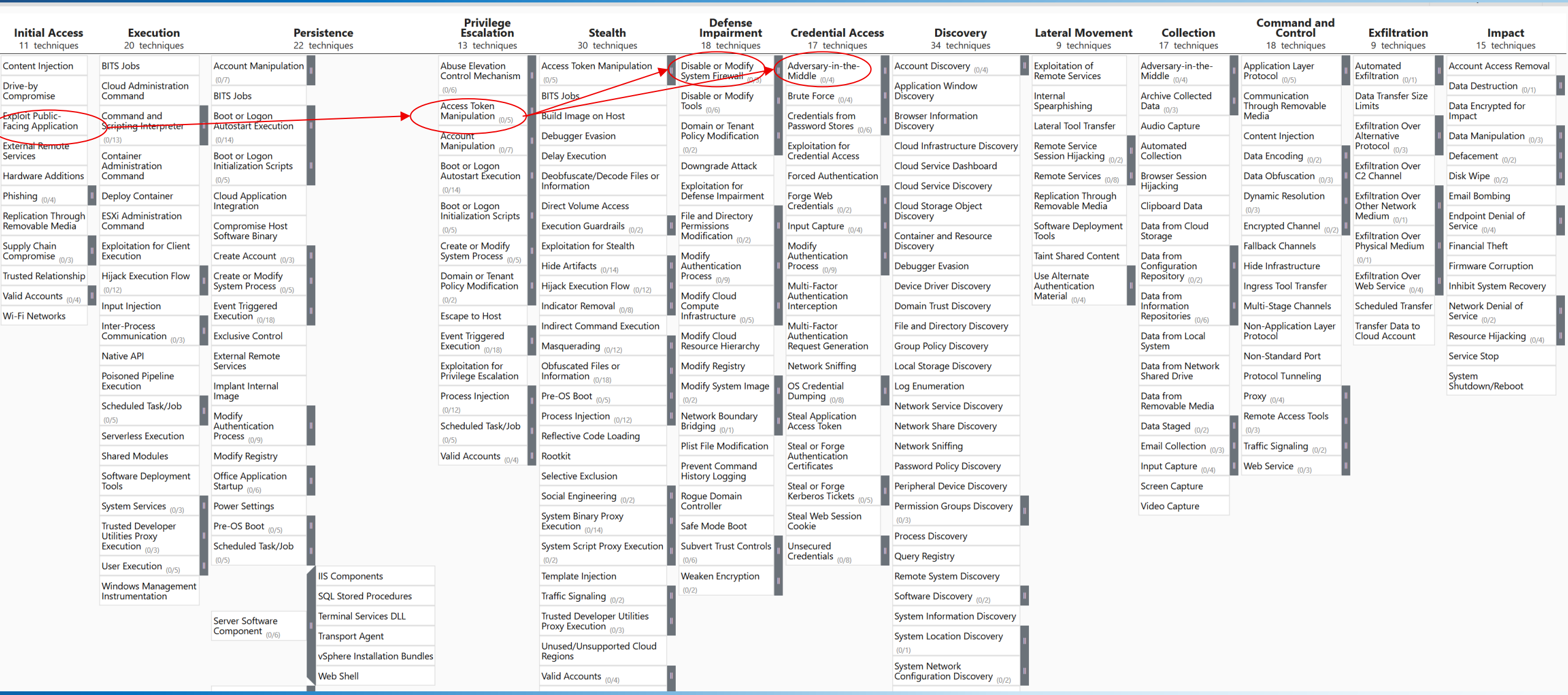
- Implement similar guardrails on scope and autonomy for Agents as for humans.
- Implement audit trails and logging for Agent's activity.
- Containment of Agent behavior/activity in case of misbehavior.
- Perform AI red teaming.

Incident response
At pace

MITRE ATT&CK


Initial Access
11 techniques
Content Injection
Drive-by Compromise
Exploit Public-Facing Application
External Remote Services
Hardware Additions
Phishing (0/4)
Replication Through Removable Media
Supply Chain Compromise (0/3)
Trusted Relationship
Valid Accounts (0/4)
Wi-Fi Networks
Execution
20 techniques
BITS Jobs
Cloud Administration Command
Command and Scripting Interpreter (0/13)
Container Administration Command
Deploy Container
ESXi Administration Command
Exploitation for Client Execution
Hijack Execution Flow (0/12)
Input Injection
Inter-Process Communication (0/3)
Native API
Poisoned Pipeline Execution
Scheduled Task/Job (0/5)
Serverless Execution
Shared Modules
Software Deployment Tools
System Services (0/3)
Trusted Developer Utilities Proxy Execution (0/3)
User Execution (0/5)
Windows Management Instrumentation
Persistence
22 techniques
Account Manipulation (0/7)
BITS Jobs
Boot or Logon Autostart Execution (0/14)
Boot or Logon Initialization Scripts (0/5)
Cloud Application Integration
Compromise Host Software Binary
Create Account (0/3)
Create or Modify System Process (0/5)
Event Triggered Execution (0/18)
Exclusive Control
External Remote Services
Implant Internal Image
Modify Authentication Process (0/9)
Modify Registry
Office Application Startup (0/6)
Power Settings
Pre-OS Boot (0/5)
Scheduled Task/Job (0/5)
Server Software Component (0/6)
IIS Components
SQL Stored Procedures
Terminal Services DLL
Transport Agent
vSphere Installation Bundles
Web Shell
Privilege Escalation
13 techniques
Abuse Elevation Control Mechanism (0/6)
Access Token Manipulation (0/5)
Account Manipulation (0/7)
Boot or Logon Autostart Execution (0/14)
Boot or Logon Initialization Scripts (0/5)
Create or Modify System Process (0/5)
Domain or Tenant Policy Modification (0/2)
Escape to Host
Event Triggered Execution (0/18)
Exploitation for Privilege Escalation
Process Injection (0/12)
Scheduled Task/Job (0/5)
Valid Accounts (0/4)
Stealth
30 techniques
Access Token Manipulation (0/5)
BITS Jobs
Build Image on Host
Debugger Evasion
Delay Execution
Deobfuscate/Decode Files or Information
Direct Volume Access
Execution Guardrails (0/2)
Exploitation for Stealth
Hide Artifacts (0/14)
Hijack Execution Flow (0/12)
Indicator Removal (0/8)
Indirect Command Execution
Masquerading (0/12)
Obfuscated Files or Information (0/18)
Pre-OS Boot (0/5)
Process Injection (0/12)
Reflective Code Loading
Rootkit
Selective Exclusion
Social Engineering (0/2)
System Binary Proxy Execution (0/14)
System Script Proxy Execution (0/2)
Template Injection
Traffic Signaling (0/2)
Trusted Developer Utilities Proxy Execution (0/3)
Unused/Unsupported Cloud Regions
Valid Accounts (0/4)
Defense Impairment
18 techniques
Disable or Modify System Firewall (0/3)
Disable or Modify Tools (0/6)
Domain or Tenant Policy Modification (0/2)
Downgrade Attack
Exploitation for Defense Impairment
File and Directory Permissions Modification (0/2)
Modify Authentication Process (0/9)
Modify Cloud Compute Infrastructure (0/5)
Modify Cloud Resource Hierarchy
Modify Registry
Modify System Image (0/2)
Network Boundary Bridging (0/1)
Plist File Modification
Prevent Command History Logging
Rogue Domain Controller
Safe Mode Boot
Subvert Trust Controls (0/6)
Weaken Encryption (0/2)
Credential Access
17 techniques
Adversary-in-the-Middle (0/4)
Brute Force (0/4)
Credentials from Password Stores (0/6)
Exploitation for Credential Access
Forced Authentication
Forge Web Credentials (0/2)
Input Capture (0/4)
Modify Authentication Process (0/9)
Multi-Factor Authentication Interception
Multi-Factor Authentication Request Generation
Network Sniffing
OS Credential Dumping (0/8)
Steal Application Access Token
Steal or Forge Authentication Certificates
Steal or Forge Kerberos Tickets (0/5)
Steal Web Session Cookie
Unsecured Credentials (0/8)
Discovery
34 techniques
Account Discovery (0/4)
Application Window Discovery
Browser Information Discovery
Cloud Infrastructure Discovery
Cloud Service Dashboard
Cloud Service Discovery
Cloud Storage Object Discovery
Container and Resource Discovery
Debugger Evasion
Device Driver Discovery
Domain Trust Discovery
File and Directory Discovery
Group Policy Discovery
Local Storage Discovery
Log Enumeration
Network Service Discovery
Network Share Discovery
Network Sniffing
Password Policy Discovery
Peripheral Device Discovery
Permission Groups Discovery (0/3)
Process Discovery
Query Registry
Remote System Discovery
Software Discovery (0/2)
System Information Discovery
System Location Discovery (0/1)
System Network Configuration Discovery (0/2)
Lateral Movement
9 techniques
Exploitation of Remote Services
Internal Spearphishing
Lateral Tool Transfer
Remote Service Session Hijacking (0/2)
Remote Services (0/8)
Replication Through Removable Media
Software Deployment Tools
Taint Shared Content
Use Alternate Authentication Material (0/4)
Collection
17 techniques
Adversary-in-the-Middle (0/4)
Archive Collected Data (0/3)
Audio Capture
Automated Collection
Browser Session Hijacking
Clipboard Data
Data from Cloud Storage
Data from Configuration Repository (0/2)
Data from Information Repositories (0/6)
Data from Local System
Data from Network Shared Drive
Data from Removable Media
Data Staged (0/2)
Email Collection (0/3)
Input Capture (0/4)
Screen Capture
Video Capture
Command and Control
18 techniques
Application Layer Protocol (0/5)
Communication Through Removable Media
Content Injection
Data Encoding (0/2)
Data Obfuscation (0/3)
Dynamic Resolution (0/3)
Encrypted Channel (0/2)
Fallback Channels
Hide Infrastructure
Ingress Tool Transfer
Multi-Stage Channels
Non-Application Layer Protocol
Non-Standard Port
Protocol Tunneling
Proxy (0/4)
Remote Access Tools (0/3)
Traffic Signaling (0/2)
Web Service (0/3)
Exfiltration
9 techniques
Automated Exfiltration (0/1)
Data Transfer Size Limits
Exfiltration Over Alternative Protocol (0/3)
Exfiltration Over C2 Channel
Exfiltration Over Other Network Medium (0/1)
Exfiltration Over Physical Medium (0/1)
Exfiltration Over Web Service (0/4)
Scheduled Transfer
Transfer Data to Cloud Account
Impact
15 techniques
Account Access Removal
Data Destruction (0/1)
Data Encrypted for Impact
Data Manipulation (0/3)
Defacement (0/2)
Disk Wipe (0/2)
Email Bombing
Endpoint Denial of Service (0/4)
Financial Theft
Firmware Corruption
Inhibit System Recovery
Network Denial of Service (0/2)
Resource Hijacking (0/4)
Service Stop
System Shutdown/Reboot

MITRE ATT&CK – attack flow



Incident response process

Source: SANS

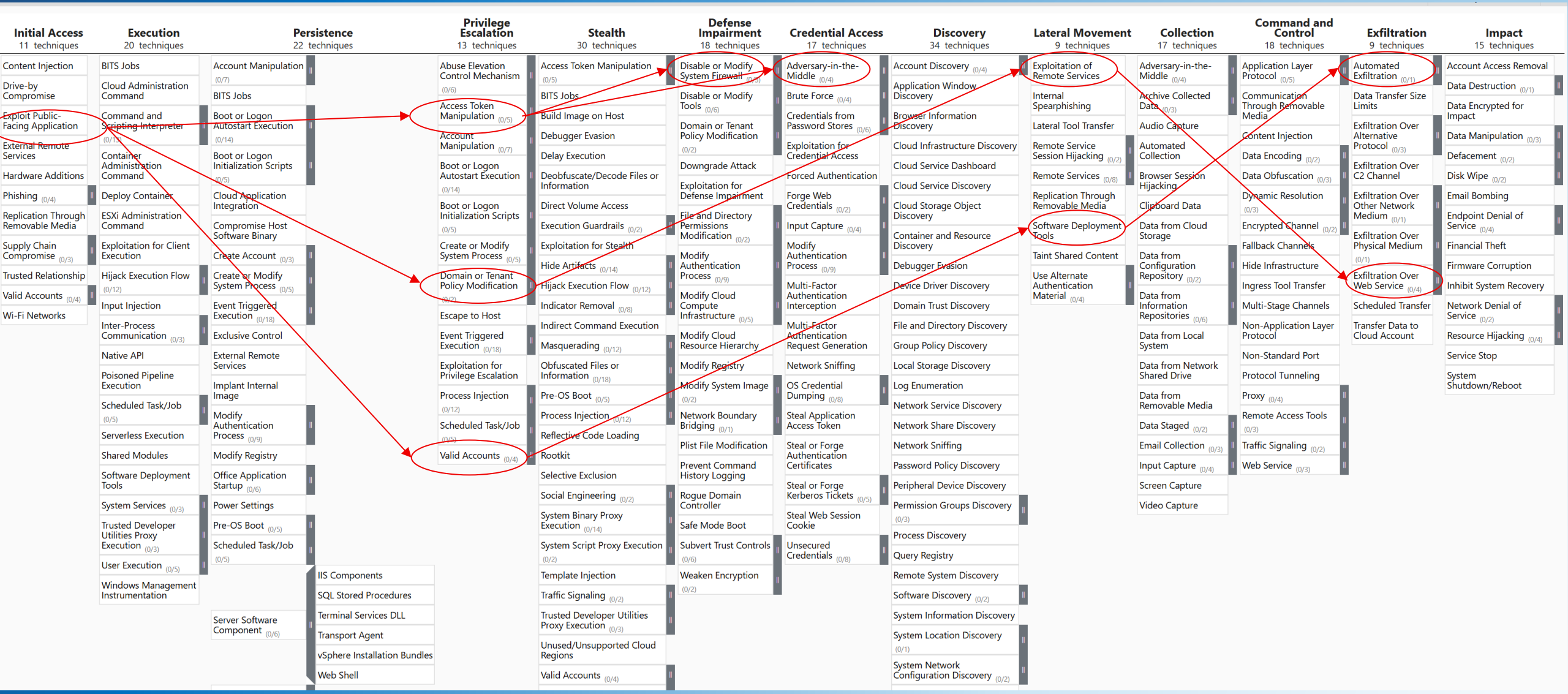
Identify

- When did the event happen?
- Has the source (point of entry) of the event been discovered?
- What is the scope of the compromise?
- Is it spreading?
- Does it affect operations?
- …

Contain

- Limit spread and impact

MITRE ATT&CK – in fast forward



Incident response process – in fast forward

Source: SANS

Identify

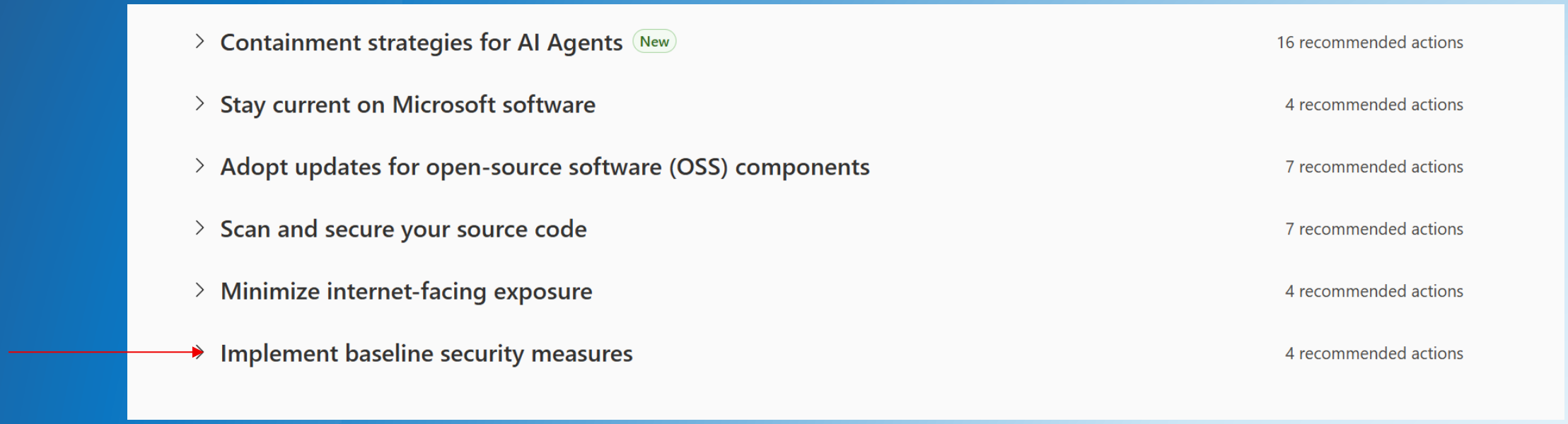
- When did the event happen?
- Has the source (point of entry) of the event been discovered?
- What is the scope of the compromise?
- Is it spreading?
- Does it affect operations?
- **Limit spread and impact**
- …

Incident response

- Identify your most (and least) critical assets
- Obtain agreement to burn down least critical
- Identify trigger points in attack flow and infra
- Contain without perfect knowledge
- Prepare for Agents becoming your adversaries
- Use AI in defense: red teaming, detection, response

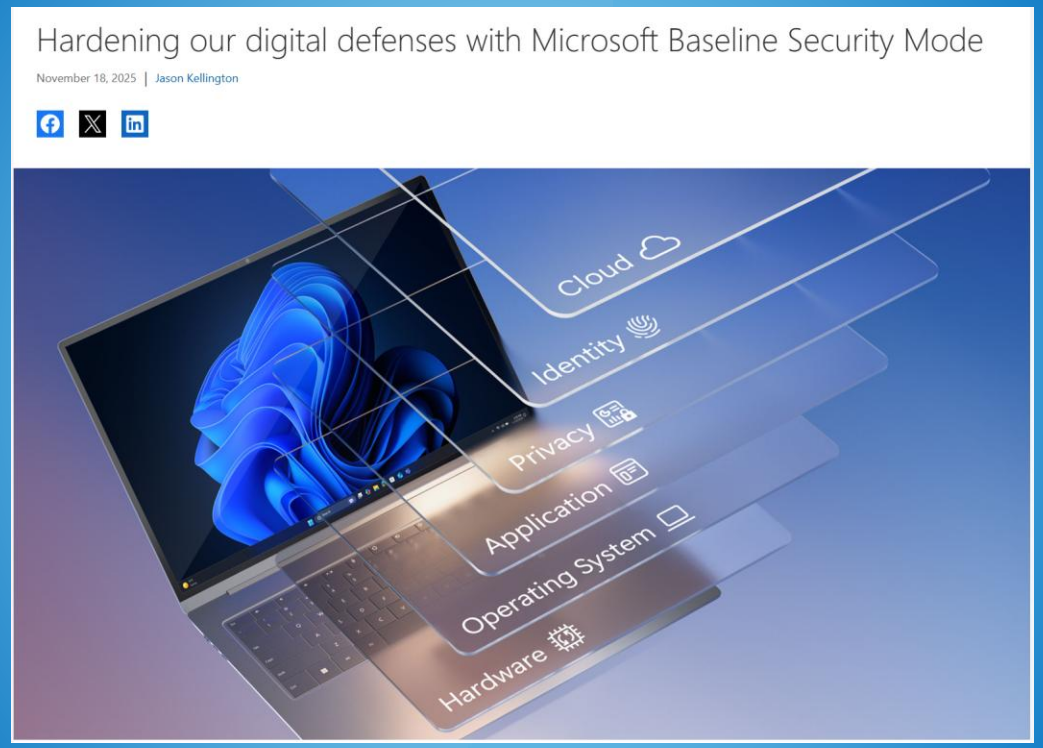
Defense in depth
The basics still matter

Defense in an AI- accelerated threat landscape



https://sip.security.microsoft.com/secure-now

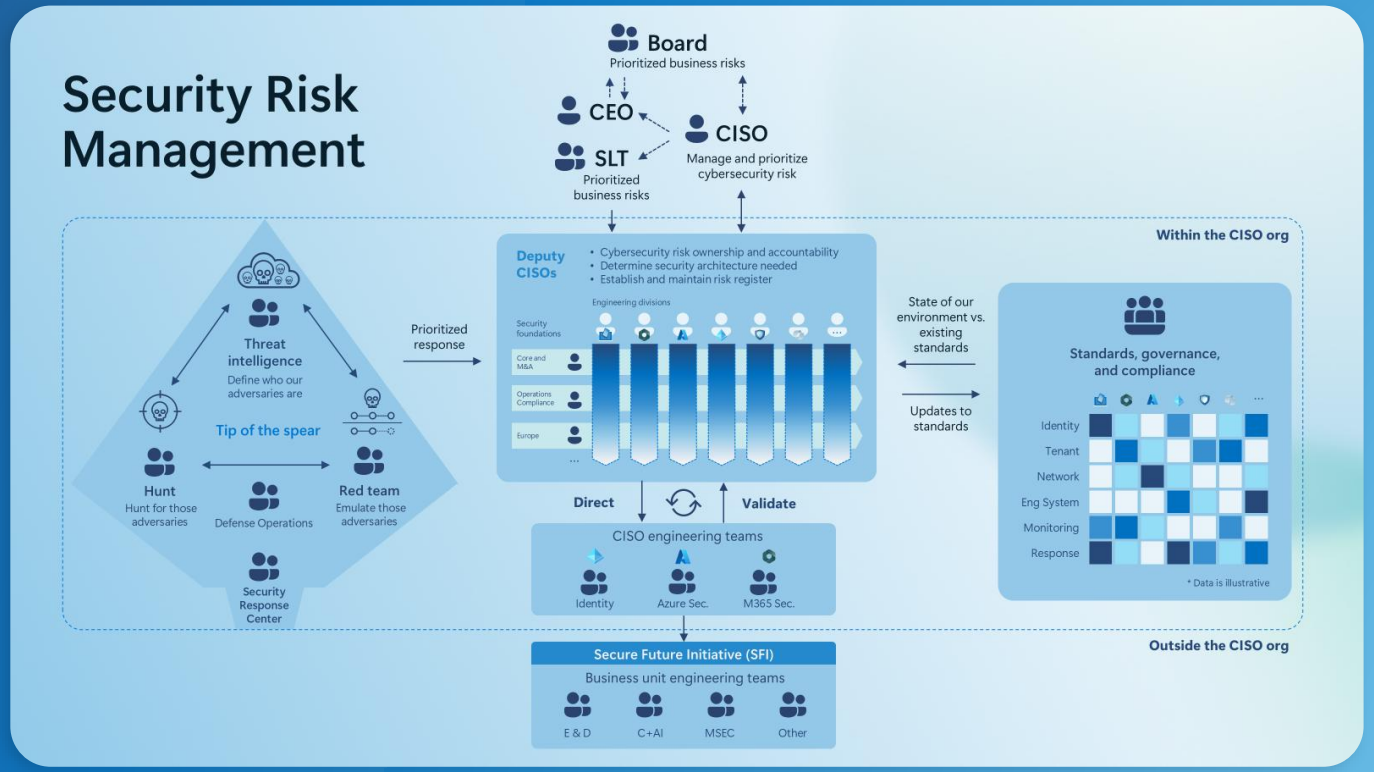
Microsoft Baseline Security Mode



- How Microsoft protects itself
- Stage rollout with built-in impact analysis, exclusions
- One entry point in the Microsoft 365 admin center
- V1: 18 secure settings
- V2: 25 additional settings across Microsoft 365, Entra, Intune, Purview, Dynamics 365, and Azure in 10/2026
- All new tenants will come with BSM controls

*** Baseline security mode settings | Microsoft Learn**

Secure by Design and Default



Microsoft tenants

Customer tenants

Microsoft infrastructure and services

Wrapping up

- The basics still matter, a lot
- Expand your controls to agents
- Deploy controls at pace - by design/default
- Use AI and AI Agents to defend
- Update your risk assessments and playbooks
- Prepare to defend with incomplete knowledge

Microsoft

Thank you!

Freddy Dezeure
DCISO for Europe