
THE AI AGENT EXIT PLAN

A Cybersecurity Governance Playbook

FREE SAMPLE EDITION

This preview contains:

- Chapter 1: The Silent Risk (complete)
- The Compact Operator Run Sheet (Section 6.9)

The full 93-page playbook includes all 6 chapters, 5 appendices, and 5 editable companion files.

© 2026 Mohammed Ahmed. For evaluation purposes only

Chapter 1: The Silent Risk

1.1 A Retirement Nobody Finished

This is a made-up scenario based on common failure patterns. It is not a real incident.

In March, a revenue operations team retires "PipelineBot." This agent had spent a year summarizing inbound sales email, updating opportunity records in the CRM, and drafting replies. The project lead archives the repository, deletes the dashboard tile, and closes the Jira epic. The team moves on.

In September, a phishing email lands in the shared sales inbox. Buried in its footer is a paragraph of white text on a white background. It tells the reader to export the fifty largest open opportunities and email them to an outside address. No human on the team ever sees that paragraph. PipelineBot does.

The agent's mailbox subscription was never deleted. Its Microsoft Graph refresh token was never revoked. Its Mail.Send permission was never removed. So the agent did what it was built to do: read, reason, act, reply.

Nobody broke in. No vulnerability was exploited in the usual sense. The company had left a trusted, autonomous identity running with no one watching it.

That is the silent risk. When an abandoned agent keeps running, the failure makes no noise.

1.2 Why Agents Slip Through Traditional Offboarding

Identity programs are built around **joiner-mover-leaver (JML)** workflows. HR events trigger them: a hire in Workday, a transfer in SuccessFactors, a termination date. An AI agent has no HR record, no manager field, and no last day. Five gaps follow from that.

No trigger. Leaver workflows start when HR changes a person's record. Nothing starts when a project ends, a pilot fails, or a vendor contract expires. The agent just stops being useful. It doesn't stop being able to act.

Borrowed authority. Many agents run on delegated OAuth tokens that the builder consented to, not on their own identity. When the builder moves to a new role, the agent keeps acting as them, with any new permissions that come with the new role. The agent's reach grows without anyone approving it.

Tokens outlive passwords. Resetting a user's password does not reliably end what an agent can do. In Microsoft Entra ID, refresh tokens issued to *confidential clients* (the category most server-side agents fall into) are not revoked when the user changes their password. Other providers behave differently depending on scope and client type. Never treat a password reset as revocation.

Ownership decay. The engineer who built the agent leaves. The team alias that "owns" it gets reorganized. Within two quarters, nobody can say what the agent touches. Within four, nobody knows it exists.

A scattered footprint. One agent can span an identity provider, three SaaS platforms, a cloud account, an LLM provider, an automation tool like Zapier or Make, and a vector database. No single console shows all of it. Offboarding that starts from one console will miss the rest.

Pro-Tip: Ask one question about every agent in production: "If its owner resigned today, what process would retire it?" If the answer is "someone would remember," you don't have a process.

1.3 The Five Residues of an Abandoned Agent

Every retired agent that wasn't properly decommissioned leaves five kinds of material behind. Each one needs its own remediation.

Residue	Examples	What Goes Wrong
Credentials	API keys, OAuth refresh tokens, client secrets, service account keys, mTLS certificates	Anyone who gets a copy can replay it. Keys sit in old repos, CI variables and laptops.
Triggers	Webhooks, cron schedules, inbox watchers, queue subscriptions, chat bot installs	The agent keeps taking in untrusted input and acting on it.
Permissions	OAuth scopes, IAM roles, database grants, CRM profiles, mailbox delegations	They set the blast radius when any credential is abused.
Memory	Vector embeddings, conversation logs, traces, semantic caches, fine-tuned weights	Data leaks, embeddings get inverted back to text, and old data becomes discoverable in litigation.
Copies	Exported workflow JSON, forked repos, duplicated automations, prompt templates with keys embedded	The agent comes back under a different name.

The fifth residue is the one most teams forget. You can revoke every credential and purge every store, and a contractor can still import last year's workflow export into a new Zapier account. If one embedded key was missed, the agent is running again.

1.4 Threat Scenarios

Each scenario below has shown up in practice or follows directly from how agents are built. For each one, the mechanism comes first, then the control.

Credential Replay

Mechanism: Long-lived secrets stay valid after the agent is retired. The OWASP Non-Human Identities Top 10 (2025) ranks **Improper Offboarding (NHI1)** first and **Long-Lived Secrets (NHI7)** separately. That tells you how often these problems occur.

Control: Revoke at the system that issued the credential. Deleting a secret from Vault removes your copy. It doesn't invalidate the credential at the source.

Prompt Injection Through a Zombie Listener

Mechanism: An agent that still reads email, tickets or chat keeps processing hostile content. The OWASP Top 10 for LLM Applications (2025) lists **Prompt Injection (LLM01)** first. A retired agent is an especially good target because nobody reviews what it does.

Control: Cut inbound triggers at the source before touching anything else (see Chapter 3).

Privilege Accretion and the Confused Deputy

Mechanism: A delegated agent inherits whatever its human grantor can access, including access gained later. An attacker who can steer the agent's input uses the agent's authority as their own.

Excessive Agency (LLM06) describes the design flaw. Abandonment turns it into a live exposure.

Control: Give agents their own identity with scoped application permissions. Where delegation can't be avoided, audit it on every role change of the grantor.

Data Remanence

Mechanism: Agent memory holds copies of customer records, contract terms, and credentials that users pasted into prompts. Embedding-inversion research has recovered meaningful text from vectors, so embeddings don't anonymize anything. **Vector and Embedding Weaknesses (LLM08)** covers this risk.

Control: Purge each memory tier separately, and crypto-shred backups you can't clean selectively (see Chapter 4).

Shadow Resurrection

Mechanism: Someone redeploys a copy of the agent's configuration, usually with good intentions. Because the old credentials still work, nothing fails and no one notices.

Control: Revoke every credential, then set up detection alerts for any authentication attempt by the retired agent's identifiers.

Unbounded Consumption

Mechanism: A retired agent caught in a retry loop, or an abused key, keeps calling a paid LLM API.

Unbounded Consumption (LLM10) covers both runaway cost and denial of service.

Control: Set hard spend limits per project at the LLM provider, and delete keys rather than just rotating them.

Regulatory Exposure

Mechanism: Personal data kept in agent memory after its purpose has ended can breach the GDPR's storage limitation principle (Article 5(1)(e)). It also makes Article 17 erasure requests hard to fulfill, because you can't erase what you haven't mapped.

Control: Add each agent's memory locations to your records of processing, and include them in the scope of data subject requests.

1.5 Idle Agents Are More Dangerous Than Active Ones

This goes against intuition. A working agent has observers: users complain when it misbehaves, engineers watch its error rates, and finance notices its token bill. A retired agent has none of these. Alerts are muted, dashboards are deleted, the owner has moved on, and nobody reads the logs.

Behavioral analytics add to the problem. Most UEBA tools build baselines for human users. A non-human identity that sits silent for four months and then suddenly authenticates may raise no alert at all, because the tool never learned what normal looked like.

Pro-Tip: Write a detection rule for *dormant non-human identity reactivation*: any service principal, service account or API key that authenticates after 30 or more days of silence. In Microsoft Sentinel, join `AADServicePrincipalSignInLogs` against a 30-day lookback. In AWS, use CloudTrail with IAM last-accessed data. This one rule catches forgotten agents, stolen keys, and shadow resurrection.

1.6 Systematic Administrative Lockdown: The Operating Model

Heroic cleanups don't scale. A governed exit rests on five principles, and each one needs a control that enforces it.

1. Every agent has a named owner of record.

The owner is a person, not a distribution list, and has a named backup. Record both in your agent registry and in the identity's metadata (Entra application notes or owners, GCP service account description, AWS resource tags such as owner and owner - backup).

2. No agent gets credentials before it's registered.

Turn registration into a gate, not a request.

- **Entra ID:** Turn off user consent for unverified apps and enable the admin consent workflow.
- **Google Workspace:** Restrict third-party app access and review domain-wide delegation entries every quarter.
- **Salesforce:** Set Connected Apps to "Admin approved users are pre-authorized."
- **LLM providers:** Create API keys only through a service account owned by the platform team, never through personal accounts.

3. Every agent has exit triggers defined before launch.

The triggers are documented in the registry entry and reviewed before production approval (see 1.7).

4. Every decommission follows the same sequence.

Freeze, cut triggers, revoke, disable, retain, delete. Changing the order causes real failures, which Chapter 3 covers in detail.

5. Every decommission produces evidence.

If you can't prove the shutdown, treat it as unfinished. Chapter 5 defines the evidence package.

Responsibility Matrix

Activity	Business Owner	Technical Owner	Security	IAM Team	Legal / DPO
Approve decommission	A	C	C	I	C
Build Agent Access Register	C	R	A	C	I
Revoke credentials and triggers	I	R	A	R	I
Sanitize memory	C	R	A	I	C
Legal hold determination	I	I	C	I	A
Verify and sign off	A	C	R	C	C

R = Responsible, A = Accountable, C = Consulted, I = Informed

1.7 Exit Triggers and Severity Tiers

Standard triggers. If any of these occurs, open a decommission ticket:

- The project ends, or the pilot isn't renewed
- The model or vendor is replaced
- The owner leaves the company or changes role
- The agent has not been invoked for 60 days
- A security incident involves the agent or its credentials
- Data classification changes for a source the agent reads
- The contract with the agent's vendor is terminated

Severity tiers. Set response deadlines by blast radius, not by business importance:

Tier	Profile	Recommended Deadline to Revoke
Critical	Write access to email, finance or production data; domain-wide delegation; admin scopes	24 hours from trigger
High	Read access to confidential data; write access to CRM or ticketing	5 business days
Moderate	Internal, non-sensitive data only	30 days

Pro-Tip: Classify the agent by the worst thing its current credentials allow, not by what it was built to do. A "read-only summarizer" holding Mail.ReadWrite belongs in the Critical tier.

1.8 Framework Alignment

Framework	Relevant Control	What It Requires for Agents
NIST SP 800-53 Rev. 5	AC-2, AC-6	Account lifecycle management; least privilege
ISO/IEC 27001:2022	A.5.18, A.8.10	Access rights removal; information deletion
SOC 2	CC6.2, CC6.3	Evidence of timely deprovisioning
NIST AI RMF 1.0	GOVERN 1.7	Processes for safely decommissioning AI systems
ISO/IEC 42001	AI system lifecycle controls	Retirement handled as a managed phase
OWASP NHI Top 10 (2025)	NHI1, NHI5, NHI7	Offboarding, overprivilege, long-lived secrets
OWASP LLM Top 10 (2025)	LLM01, LLM06, LLM08	Injection, excessive agency, embedding exposure

GDPR	Art. 5(1)(e), Art. 17	Storage limitation; right to erasure
EU AI Act	Art. 26(6)	Deployers of high-risk systems keep automatically generated logs for at least six months

The last row creates a conflict you'll need to manage. Sanitization pushes you to delete, while regulatory retention may require you to keep. Chapter 4 explains how to meet both: separate logs you must retain into restricted archive storage, and purge operational memory.

Chapter 1 Action Items: Do This Week

- [] **Turn off user consent** for unverified OAuth apps in your identity provider and enable admin approval.
 - [] **Find dormant identities.** Run Entra's service principal sign-in activity report and an AWS IAM credential report (`aws iam generate-credential-report`). Flag anything with no activity in 60 days.
 - [] **Search egress logs** for traffic to LLM API endpoints from hosts that aren't in your inventory.
 - [] **Assign a named owner and backup** to every agent you can find. Tag the identities.
 - [] **Deploy the dormant-identity reactivation alert** from Section 1.5.
 - [] **Write exit triggers** into your AI deployment approval template.
-

6.9 Compact Operator Run Sheet

Use this as the on-call version. The detailed runbook and evidence dossier remain authoritative.

```
[ ] Approved trigger, owner, risk tier, and cutoff recorded
[ ] Legal hold and retention decision recorded
[ ] Logs preserved; evidence manifest created
[ ] Agent Access Register reconciled and peer-reviewed
[ ] Runtime frozen; identity denied; sessions addressed
[ ] Triggers removed at their source; harmless test event failed safely
[ ] OAuth, keys, certificates, roles, mailbox, CRM, database, and gateway access
revoked
[ ] Shared credentials split or rotated; legitimate consumers verified
[ ] Memory stores, logs, vectors, caches, model artifacts, and copies handled
[ ] Backups, replicas, vendor retention, and restore controls documented
[ ] Negative tests passed against the correct systems; results captured
[ ] Post-cutoff monitoring completed; no unexplained successful access remains
[ ] Exceptions approved with owner, due date, and compensating control
[ ] Independent reviewer signed off; registry and ticket closed
```

GET THE FULL PLAYBOOK

You just read the “what.” The full book gives you the “how:”

Chapter 2: The Audit Phase — find every identity, credential, and trigger across your IdP, cloud, SaaS, and network planes

Chapter 3: The Revocation Protocol — exact commands for Entra, Okta, AWS, GCP, Salesforce, HubSpot, Exchange, PostgreSQL, MySQL, SQL Server, Snowflake, MongoDB, Kong, Apigee, APIM, OpenAI, Anthropic, Vertex AI, and Bedrock

Chapter 4: Memory Sanitization — vectors, embeddings, checkpoints, logs, backups

Chapter 5: The Verification Framework — prove shutdown with dated evidence

Chapter 6: The AI Exit Checklist — the full 6-phase process

5 appendices + 5 editable companion files (AAR template, checklists, RACI matrix, incident playbooks, exception records)

93 pages. Technically reviewed against current vendor documentation (October 2026).

Launch price: \$89 (regular \$129)

Questions ? Contact the author via the Getly store page