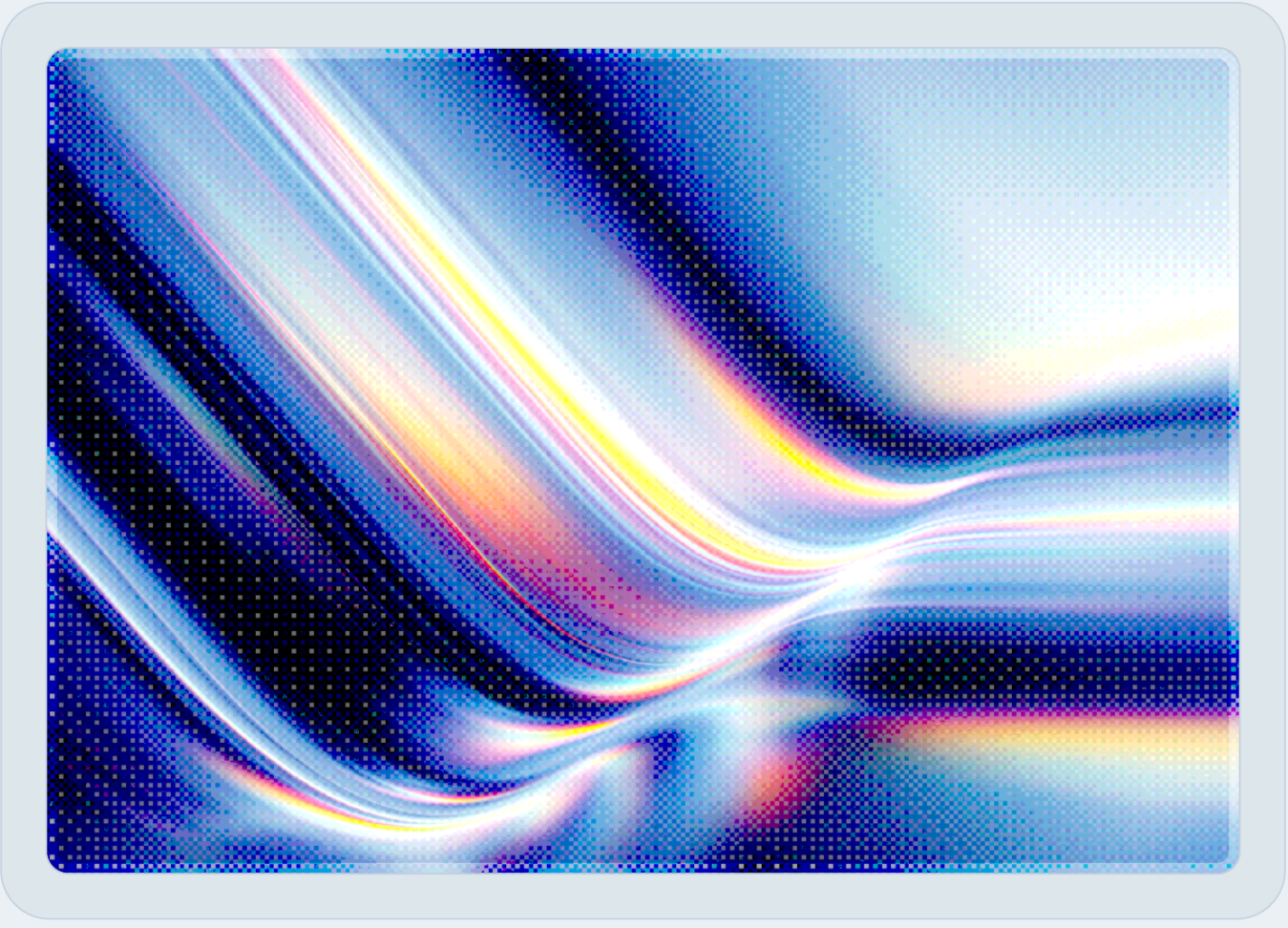
CSAI Foundation | Cloud Security Alliance

Entra Agent ID Administrator Flaw: Service Principal Takeover

When an AI-Agent Privileged Role Reaches the Wider Tenant

2026-04-28

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- **An Entra ID role intended for agent-identity administration had non-agent reach.** Silverfort researcher Noa Ariel disclosed that the Agent ID Administrator role in Microsoft Entra ID could be used to add owners to arbitrary service principals across a tenant, including non-agent applications. With ownership in hand, an attacker could mint new client secrets or certificates and authenticate as the targeted service principal [1][2][3].
- **The boundary failure converted a scoped role into a broad takeover primitive.** Because agent identities in Entra are implemented atop the same service principal infrastructure used by traditional applications, the role's `microsoft.directory/agentIdentities/owners/update` and related permissions leaked into the wider service principal plane. Adding self as owner on the parent Application object was denied as expected; the gap was specific to the service principal surface [1][3].
- **The maximum demonstrated impact reached Global Administrator under research conditions.** Silverfort demonstrated end-to-end privilege escalation by hijacking a service principal that had been assigned the Global Administrator directory role, then signing in with attacker-controlled credentials [1][4]. Silverfort's tenant telemetry — which the company itself notes should be read as a risk-context indicator rather than exploitation evidence — reports that roughly 99% of business tenants in its dataset contain at least one privileged service principal, suggesting the underlying pattern is common in enterprise environments [1][2].
- **Microsoft has patched, but governance work remains.** Microsoft confirmed the issue on March 26, 2026 and rolled the fix to all clouds by April 9, 2026; ownership assignments by Agent ID Administrator over non-agent service principals now return a "Forbidden" error [1][2][5]. The patch closes the immediate authorization gap but does not retire any unauthorized credentials, ownership changes, or role assignments that may already exist in tenants where the role was used before April 9.
- **We read this incident as an early instance of a recurring pattern.** As cloud platforms ship dedicated identity primitives for AI agents, new "agent" roles, scopes, and object types should be treated as full members of the privileged identity inventory rather than as low-

impact administrative conveniences. Independent coverage of the disclosure consistently characterized it as a privilege escalation issue with tenant-wide impact rather than a defect contained to the agent feature surface [2][6][7].

Background

Microsoft introduced Entra Agent ID as a tenant-side identity platform for AI agents, layering an "agent" subtype onto Entra's existing service principal infrastructure rather than building a parallel object model. According to Microsoft's documentation, agent identities are modeled as single-tenant service principals tied to a parent **agent identity blueprint principal**, with the blueprint holding credentials and acquiring tokens on behalf of each child agent identity through Federated Identity Credential relationships. The blueprint authenticates; the child agent identity is what appears as the actor in audit logs and what carries permissions on downstream resources [8].

To administer this new object class, Microsoft shipped the **Agent ID Administrator** built-in role. Microsoft's role catalog gives the role permissions targeted at agent objects, including `microsoft.directory/agentIdentities/owners/update` and `microsoft.directory/agentIdentityBlueprintPrincipals/owners/update`. The role's stated purpose is to handle the lifecycle of AI agent identities — create blueprints, instantiate agent identities, manage their owners — without granting broader application administration rights such as those held by Application Administrator or Cloud Application Administrator [1][9].

The design intent matters because organizations are deploying these objects at scale. Silverfort's tenant telemetry, again reported as risk-context rather than exploitation evidence, indicates that more than half of its surveyed customer tenants already use agent identities, and of those that do, nearly half operate more than 100 agent identities per tenant [1]. CSA's own *Identity and Access Gaps in the Age of Autonomous AI* survey reports that 73% of organizations expect AI agents to become vital within the next year, even as 68% cannot reliably distinguish AI agent activity from human activity in their logs and only 22% report consistent application of access frameworks to agents [10]. The Agent ID Administrator role lands in that gap — a new privileged role granted under the assumption that it is narrowly scoped to a new object class.

Security Analysis

The Boundary Failure

The vulnerability is best understood as a scope-overreach defect, not a code injection or token-forgery flaw. Silverfort observed that when a user holding only the Agent ID Administrator role attempted to add themselves as owner of a non-agent service principal — for example, the service principal of a traditional enterprise application — the directory accepted the operation. Attempts to perform the same operation on the corresponding Application object were denied as expected, which Silverfort attributes to the agent-related ownership permissions being evaluated against the service principal surface without a reliable check that the target service principal was actually an agent identity [1][3].

The architectural reason follows from the shared object model. Agent identities, blueprints, and traditional applications all serialize down to service principals in the directory. Microsoft's documentation describes agent service principals as "modeled as single-tenant service principals with a new 'agent' subtype classification" that uses "existing Microsoft Entra ID service principal infrastructure while adding agent-specific behaviors and constraints" [8]. The Agent ID Administrator role's ownership permissions were applied at the layer that all these objects share, before the subtype classification was enforced. The result was an authorization decision that approximated the directory primitive — *can this principal add owners to a service principal?* — rather than the intended semantic — *can this principal add owners to an agent service principal?*

Attack Path

The exploitation path is three steps and uses standard Microsoft Graph and Azure CLI operations. First, an attacker holding Agent ID Administrator enumerates service principals in the tenant via Microsoft Graph or Azure CLI, looking for high-value targets such as service principals that hold directory roles (for example, Global Administrator or Privileged Role Administrator) or that have consented to high-impact Microsoft Graph application permissions such as `RoleManagement.ReadWrite.Directory`, `Directory.ReadWrite.All`, or `Application.ReadWrite.All` [4][7]. Second, the attacker calls the directory ownership API to add themselves as owner of a chosen non-agent service principal — the exact operation that should have been blocked. Third, with ownership established, the attacker generates a new client secret or certificate against the service principal, then authenticates as that service principal and inherits whatever permissions it holds [1][2][4].

In Silverfort's published demonstration, the targeted service principal had been assigned the Global Administrator directory role; once the attacker authenticated as that principal with new credentials, they reached effectively unrestricted control of the tenant [1][4][6]. Silverfort summarized the dynamic with the observation that, prior to the fix, the role "allowed assigning ownership over service principals beyond agent-related identities, effectively enabling similar capabilities to roles such as Application Administrator, but without being scoped specifically to agent use cases" [1][6].

Disclosure Timeline and Patch State

The disclosure timeline is precise and well-sourced across the primary and secondary reporting. Silverfort identified the behavior on February 24, 2026 and reported it to the Microsoft Security Response Center on March 1, 2026. Microsoft confirmed the behavior on March 26, 2026, advanced a fix to pre-release on April 4, 2026, and completed the rollout to all cloud environments on April 9, 2026. Public disclosure followed on April 23, 2026 [1][2]. Following the patch, attempts by Agent ID Administrator to assign ownership over non-agent service principals are blocked with a "Forbidden" response from the directory [1][5][7].

The patch is server-side and tenant-wide, so no customer action is required to receive it. However, defenders should not treat the fix as full remediation of past exposure. If the role was assigned to any account during the vulnerable window and that account was compromised or misused, ownership entries, secrets, and certificates added during the window persist in the directory. Those artifacts continue to grant authentication paths that Microsoft's patch does not unwind.

Scope Considerations

Silverfort's prevalence figures — that approximately 99% of tenants in its sample contain at least one privileged service principal, and that more than half of surveyed tenants use agent identities — should be read as risk-context indicators rather than exploitation telemetry [1]. There is no public evidence at this writing that the issue was exploited in the wild before patching. The relevance of the figures is that the *prerequisite* asset for impactful exploitation — a privileged non-agent service principal worth hijacking — is common in enterprise Entra tenants of any meaningful size, per Silverfort's telemetry. Where the Agent ID Administrator role was assigned during the vulnerable window, the upper bound on potential impact in a given tenant is set by the most privileged service principal in that tenant, not by the population of agent identities.

Recommendations

Immediate Actions

Defenders should begin with directory hygiene targeted at the vulnerable window between Entra Agent ID general availability and the April 9, 2026 patch. The first task is to enumerate every account that currently holds, or held during that window, the Agent ID Administrator role and treat those accounts as having had elevated capability comparable to Application Administrator over the entire service principal plane. Where accounts have departed, been compromised, or were assigned the role outside of an explicit agent-administration use case, prioritize them for review.

The second task is to audit ownership and credential changes on all service principals between late February and mid-April 2026. Microsoft Entra audit logs record `Add owner to service principal`, `Add service principal credentials`, and related events with the actor identity, the target service principal, and a correlation identifier; Silverfort highlights these as the operative detection signals, and Hackread's coverage corroborates the same audit-log monitoring guidance [1][4]. Particular attention should be paid to ownership additions made by accounts whose only privileged role at the time was Agent ID Administrator, and to credential additions on service principals that hold directory roles or high-impact Microsoft Graph application permissions.

The third task is targeted credential rotation. For any non-agent service principal where an unexpected ownership entry or credential addition occurred during the window, rotate all credentials, remove unrecognized owners, and review consented application permissions. Confirm that no newly added Federated Identity Credentials or sign-in policies persist on the affected principals.

Short-Term Mitigations

Beyond incident-driven cleanup, organizations should formalize the Agent ID Administrator role as a tier-zero privileged role for assignment-control purposes. The role should be governed through Microsoft Entra Privileged Identity Management with just-in-time activation, justification, approval, and time-bound elevation, mirroring the controls already applied to Application Administrator and Cloud Application Administrator. Standing assignments should be the exception, not the default, and emergency-access (break-glass) accounts should not hold the role.

Service principal ownership itself deserves the same scrutiny as group ownership and role assignment. We recommend establishing a baseline of expected owners for each privileged service principal, and alerting on any deviation regardless of which role made the change. Where feasible, restrict ownership of

high-impact service principals to a small number of designated identity-team accounts, and avoid assigning ownership to individual user accounts that also hold productivity or development roles.

Detection content should be updated to monitor for the specific pattern Silverfort describes: an account with Agent ID Administrator that performs ownership or credential operations against a service principal whose `servicePrincipalType` or agent subtype indicates it is not an agent identity. Sign-in events for service principals that newly acquire credentials, especially those tied to directory roles, should be correlated with the originating ownership change.

Strategic Considerations

The deeper lesson is that platform-introduced "AI-agent" roles, scopes, and object types are now part of the privileged identity inventory and should be governed accordingly. Identity teams should require a written design review whenever a cloud provider introduces a new agent-class role, including questions about which underlying primitives the role can act on, how the platform distinguishes agent objects from traditional ones at the authorization layer, and what audit signals exist to detect abuse. Treat the absence of clear answers as a signal to delay broad assignment of the role until those answers exist.

Organizations should also reconsider the privilege footprint of service principals more broadly. Any service principal granted a directory role such as Global Administrator or any of the high-impact Microsoft Graph application permissions becomes a takeover target for any future role or feature that can write to service principal ownership or credentials. Reducing the count of such principals — by retiring legacy applications, consolidating consented permissions, and preferring delegated permissions where possible — shrinks the attack surface for not just this issue but the next one in the same class.

Finally, integrate AI-agent identity management into existing non-human identity (NHI) governance rather than building a parallel program. CSA's *Identity and Access Gaps in the Age of Autonomous AI* survey reports that organizations leaning on governance mechanisms as a stopgap for missing identity-level controls — disabling identities or revoking tokens after the fact — were the most common containment posture, at 49% [10]. The Agent ID Administrator incident illustrates why this is insufficient: by the time governance reacts, the takeover has already produced new credentials with their own lifetimes.

CSA Resource Alignment

The Agent ID Administrator boundary failure maps cleanly onto several existing CSA frameworks and prior publications. Within the **MAESTRO** agentic AI threat modeling framework, the issue sits at the intersection of the Agent Identity layer (where the new agent service principal subtype lives) and the underlying Foundation/Infrastructure layer (where the shared service principal primitive operates). The threat is precisely the cross-layer authorization mismatch MAESTRO is designed to surface, in which a control assumed to be scoped to one layer's object type is enforced at a deeper, more permissive layer [11].

The **AI Controls Matrix (AICM)** – CSA's vendor-neutral controls framework, which builds on the Cloud Controls Matrix and contains 243 control objectives across 18 security domains – provides the control-objective vocabulary for the recommendations above, particularly its identity and access management domain and its agent-identity controls covering least privilege, lifecycle management, and credential governance for non-human identities [12]. Organizations using AICM to assess AI deployments should add Agent ID Administrator-class roles to the privileged identity inventory referenced by those controls.

CSA's **Using Zero Trust to Counter Identity Spoofing & Abuse** white paper frames the broader pattern at issue: a legitimate identity (the Agent ID Administrator account) abusing a legitimately granted permission to act outside of its intended scope, which the paper categorizes as identity abuse rather than identity spoofing. The detective and policy controls the paper recommends – continual verification, attribute-based authorization, comprehensive visibility into ownership and credential events on privileged objects – are the operative controls for the recommendations in this note [13]. CSA's *Identity and Access Gaps in the Age of Autonomous AI* survey provides the population-level evidence that the controls above are not yet consistently in place for agent identities in enterprise environments [10].

References

[1] Ariel, Noa (Silverfort). "Agent ID Administrator scope overreach: Service Principal takeover in Entra ID." Silverfort Blog, April 23, 2026.

[2] The Hacker News. "Microsoft Patches Entra ID Role Flaw That Enabled Service Principal Takeover." The Hacker News, April 2026.

[3] Cybersecurity News. "Hackers Can Abuse Entra Agent ID Administrator Role to Hijack Service Principals." Cybersecurity News, April 2026.

[4] Hackread. "Microsoft Entra Agent ID Flaw Enabled Tenant Takeover via Privilege Escalation." Hackread, April 26, 2026.

[5] Security Affairs. "Microsoft fixes Entra ID flaw enabling privilege escalation." Security Affairs, April 2026.

[6] Sharma, Shweta. "Microsoft patched an 'agent-only' role that was not." CSO Online, April 27, 2026.

[7] Cryptika Cybersecurity. "Hackers Can Abuse Entra Agent ID Administrator Role to Hijack Service Principals." Cryptika, April 2026.

[8] Microsoft Learn. "Agent identities, service principals, and applications - Microsoft Entra Agent ID." Microsoft, April 8, 2026.

[9] Microsoft Learn. "Authorization in Microsoft Entra Agent ID." Microsoft, April 3, 2026.

[10] Cloud Security Alliance. "Identity and Access Gaps in the Age of Autonomous AI." CSA, March 2026.

[11] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[12] Cloud Security Alliance. "AI Controls Matrix." CSA, 2025.

[13] Cloud Security Alliance. "Using Zero Trust to Counter Identity Spoofing & Abuse." CSA, 2025.