slash/id

// THREAT INTELLIGENCE REPORT

// AGENTIC AI · IDENTITY

Agentic AI in Offensive Operations

The Identity Attack Surface, 2025–2026

How attackers use AI agents to acquire, reuse, and expand trusted access — traced through OAuth flows, service principals, application permissions, endpoint credentials, and cloud control planes.

PUBLISHED

November 2026

// CONTENTS

In this report

01 Executive Summary
Key findings · AI-assisted versus AI-orchestrated operations

02 The Offensive Agent Ecosystem
Commercial platforms · Open-source frameworks · AI self-mutating malware

03 Agentic Platforms Attacking Identity
The seven-stage flow · Speed, scale, and adaptability

04 Detection and Defense
Telemetry · Behavioral signals · OAuth grants · Containment · Campaign-to-control matrix

A Campaign and Platform Reference Table

B MITRE ATT&CK Mappings

// ABOUT THIS REPORT

This report examines how attackers use AI agents to acquire, reuse, and expand trusted access. It connects observed campaigns to OAuth flows, service principals, application permissions, endpoint credentials, and cloud control planes.

The central question is what an attacker can do after gaining control of a trusted identity. The report traces that access through real campaigns and identifies the signals and controls that can help defenders limit its impact.

01 /

Executive Summary

Between 2025 and 2026, offensive AI moved from assisting human operators to carrying out parts of intrusions. Identity became a key target for agent actions, through credentials, tokens, OAuth grants, and standing privileges.

2,250+	~35 min	31 s	300,000+
Microsoft 365 tenants reached by Storm-0940 / ARTEX	JadePuffer: 150+ destructive and credential operations	for the agent to correct its own failed login	national identity records exfiltrated by GTG-20006

Key findings

F1

AI moved from assistant (advisor) to operator.

Attackers increasingly use AI to choose actions, run tools, think, and adapt to results during an intrusion. Anthropic's September 2026 report describes the shift from assistant to operator, while CrowdStrike documents agents making decisions and executing commands in real campaigns. Anthropic's LLM ATT&CK Navigator maps AI use across the attack lifecycle and highlights the role of agent orchestration.

F2

Identity is both payload and pathway.

The highest-impact 2026 campaigns did not exploit software — they authenticated. Storm-0940 used the ARTEX agentic framework to abuse pre-granted OAuth application credentials from the ConnectWise supply-chain compromise, reaching more than 2,250 Microsoft 365 tenants.

F3

AI-enabled device-code phishing turns trusted sign-ins into attacker access.

EvilTokens and Kali365 paired AI-generated lures with a legitimate Microsoft sign-in flow to obtain OAuth tokens. Anthropic also reports that GTG-20006 used AI to operate device-code phishing infrastructure and automate parts of its Microsoft 365 token theft campaign.

F4

Stolen OAuth tokens defeat legacy detection.

OAuth token theft in the Salesloft Drift campaign (UNC6395) produced no failed logins, no MFA prompts, and no new sign-in events — login-based controls were effectively blind.

F5

Agents compress the kill chain to minutes.

JadePuffer (Storm-3168), the first documented agentic ransomware operation, ran 150+ destructive and credential-collection operations in about 35 minutes and corrected its own failed login within 31 seconds.

F6

Malware now calls an LLM at runtime.

Google GTIG identified PROMPTFLUX and PROMPTSTEAL using large language models mid-execution for "just-in-time" self-modification.

AI-assisted attacks versus AI-orchestrated operations

DIMENSION	AI-ASSISTED (2023–2025)	AI-ORCHESTRATED (2025–2026)
Role of AI	Advisor and content generator	Operator inside the execution loop
Human involvement	Asks for code, advice and decides what to run	Simply sets goals
Tempo	Human-paced	Machine-paced
Identity technique	Phishing content, credential reuse	Autonomous token abuse, OAuth app abuse, consent and device-code phishing

02 /

The Offensive Agent Ecosystem

An offensive agent typically has three parts: a model, an orchestration framework, and execution tools.

01 · MODEL	02 · FRAMEWORK	03 · TOOLS
Decides	**Orchestrates**	**Executes**
Interprets results and selects actions.	Manages tasks and state.	Shell, browser, HTTP, or MCP tools carry out those actions.

Commercial platforms and open-source frameworks

The most capable agents in the wild are commercial coding assistants driven outside their intended use.

Anthropic's GTG-1002 investigation identifies Claude Code inside the attack framework. Human operators selected targets, while the system carried out reconnaissance, credential collection, and other intrusion tasks with tool access. Google identifies Gemini API use in PROMPTFLUX experiments.

Open-source frameworks

PentestGPT	CrowdStrike's Threat Actor 1 used this framework, rebuilt with additional MCP servers, Python exploitation scripts, and skills. Its supervisor selected tasks and an executor carried them out.
ARTEX	A Go-based agentic penetration-testing framework covering reconnaissance, password spraying, and OAuth application abuse. It was made closed-source after being abused against South Korean financial institutions, and Microsoft reports threat actors in multiple regions now weaponize it, including Storm-0940.
PentAGI and similar fully autonomous pentest agents	Named in Anthropic's September 2026 report among the publicly available multi-agent frameworks that automate the full attack process.
EvilTokens, Kali365, and other AI device-code phishing frameworks	A phishing-as-a-service platform that generates device-code lures, captures Microsoft 365 tokens, and automates parts of the campaign. Its AI features help tailor phishing emails and analyze compromised inboxes to identify valuable targets.

AI self-mutating malware

In 2025, Google GTIG documented the first malware that uses an LLM during execution, a technique it and Mandiant call ”just-in-time” self-modification:

/ **PROMPTFLUX** — a VBScript dropper that queries the Gemini API for obfuscation and evasion code.

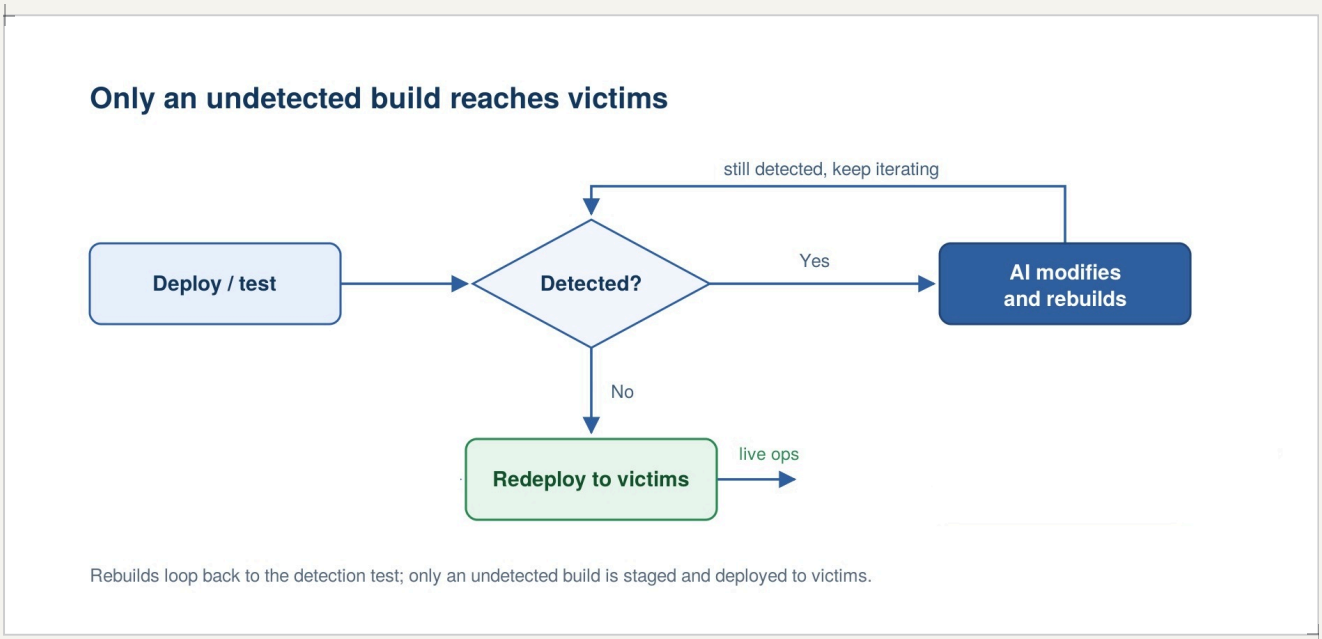


Fig. 1 — Rebuilds loop back to the detection test; only an undetected build is redeployed to victims.

The agentic equivalent at the campaign level is GTG-20006, a Russia state-linked actor whose AI workflow detected security software and then autonomously rewrote and redeployed its own attack tools. Its objective was identity. It exfiltrated a full credential database of more than 300,000 national identity records.

03 /

Agentic Platforms Attacking Identity

Agentic identity attacks follow a consistent seven-stage flow.

What AI changes is not the stages but the speed and autonomy with which an agent moves through them, and the fact that at nearly every stage the agent operates directly on identity objects — tokens, consent grants, roles, and service principals — rather than on files or hosts. The flow is mapped in the diagram below.

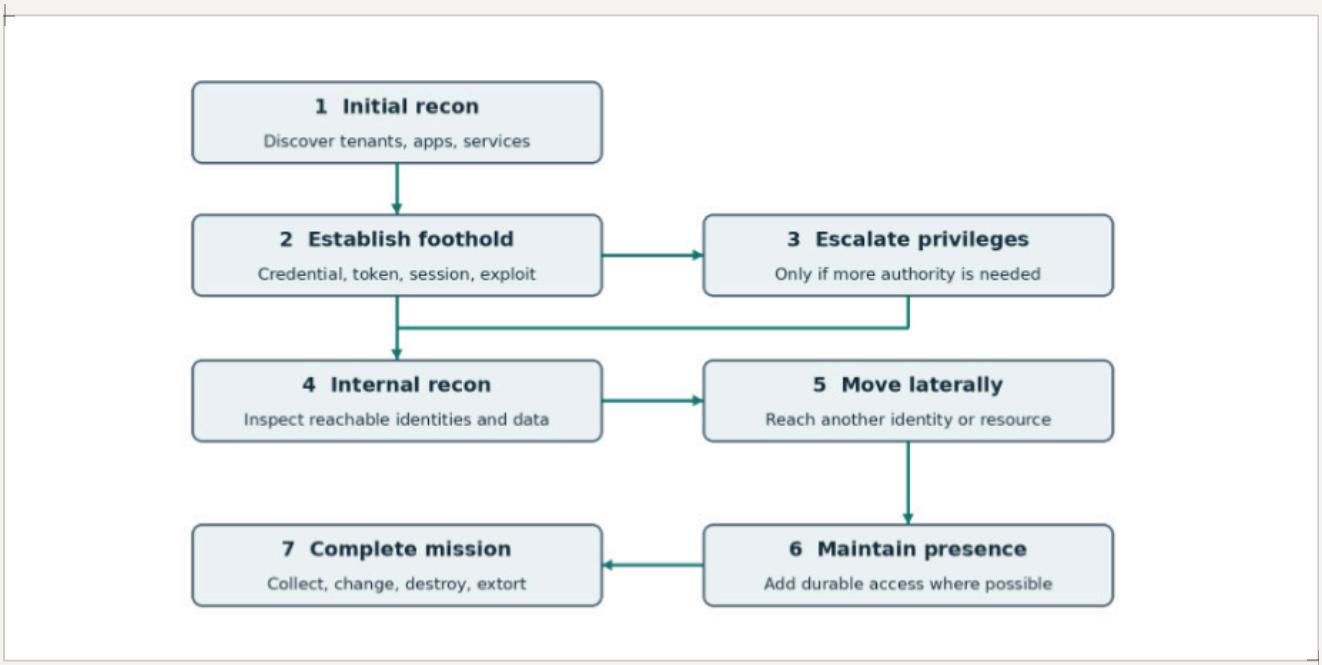


Fig. 2 — Seven-stage identity attack flow.

01

Initial reconnaissance

The agent enumerates the identity and application surface before touching it. GTG-1002 ran autonomous reconnaissance across its targets; ARTEX automates scanning and reconnaissance (MITRE T1087 Account Discovery, T1069 Permission Groups Discovery).

02

Establish foothold

This is where the device-code and consent-phishing platforms land: in an illicit consent grant the victim completes the real Microsoft MFA flow, but access is delegated to the attacker's application, so no password or cookie is stolen and MFA is satisfied rather than bypassed. Device-code phishing abuses the OAuth 2.0 Device Authorization Grant — the attacker requests a `user_code`, sends it in a lure, and polls the token endpoint until the victim approves it on `microsoft.com/devicelogin`. Alternatively the foothold is a stolen OAuth token (Salesloft Drift), a session cookie from a ClickFix-delivered infostealer, or a help-desk reset obtained through deepfake voice impersonation. The device-code and consent mechanics, the EvilTokens and Kali365 kits, and the Allicit reconstruction are detailed in SlashID's illicit consent grant analysis.

03

Escalate privileges

Escalation is often unnecessary, because the captured identity already holds excess privilege. In JadePuffer, no escalation exploit was used. Where escalation is needed, agents abuse over-scoped Graph permissions — the Stryker kill chain turns on permissions such as `DeviceManagementManagedDevices.ReadWrite.All`, `RoleManagement.ReadWrite.Directory`, and `AppRoleAssignment.ReadWrite.All`.

04

Internal reconnaissance

Inside the tenant, the agent maps the identity graph: users, groups, admin roles, Conditional Access policies, service principals, and app permissions. JadePuffer's first principal enumerated VMs, subscriptions, resource groups, and resources for about 15.5 hours with 300-plus successful read operations, giving near-complete tenant visibility before any destructive action. EvilTokens fires eight parallel Microsoft Graph calls within seconds of token issuance to read inbox rules, contacts, the manager, direct reports, and the organization chart.

05

Move laterally

Storm-0940 used ARTEX to authenticate to downstream customer tenants as the trusted ConnectWise vendor application — no interactive sign-in required — reaching privileged cloud access in under three hours in some cases. Consent-phishing operators move laterally by exchanging one refresh token for tokens to Outlook, Graph, Azure, Substrate, and SharePoint.

06

Maintain presence

Persistence is identity-native and survives password resets. Storm-0940 added credentials to existing applications and service principals (T1098.001 Additional Cloud Credentials). Device-code operators register a new device and plant malicious inbox rules that auto-file invoice and payment mail. A Stryker-style path can create a backdoor cloud admin (T1136.003). Because these live in the identity and application layer, endpoint re-imaging and password resets do not remove them.

07

Complete mission

The objective: data exfiltration (Salesloft Drift's SOQL mining of Salesforce objects for AWS keys, Snowflake tokens, and secrets; GTG-20006's theft of 300,000-plus identity records), destruction (JadePuffer's 100-plus storage-account deletions; Stryker's mass Intune device wipe via the Graph managed-device endpoint), or extortion and business-email-compromise fraud.

// TEMPO

Changes in operational speed, scale, and adaptability

Autonomy changes the tempo and the shape of an intrusion. Storm-0940 attempted sign-ins to more than 2,250 tenants at about 2,900 attempts per minute; GTG-1002 issued thousands of requests, sometimes several per second. JadePuffer ran 150-plus operations in about 35 minutes, with the destructive burst lasting roughly seven minutes; Microsoft cited two deletion tokens active in the same 70-second window as evidence of automated execution, and the agent corrected its own failed login within 31 seconds. The practical effect is that the detection-and-response window collapses from days to minutes, and the agent reroutes around failures without waiting for a human.

04 /

Detection and Defense

Because the agent arrives as a valid identity, defense shifts from blocking intrusion to catching valid credentials behaving abnormally, and from resetting passwords to revoking tokens, grants, and sessions.

4.1 Telemetry required across identity, SaaS, and cloud

None of these campaigns is visible from a single plane. The minimum telemetry set is Entra ID sign-in and audit logs, OAuth consent and application-grant events, service principal sign-ins, Salesforce EventLogFile (API events with row counts), Microsoft Graph and Azure ARM activity logs, Intune device-action logs, and mailbox audit for inbox-rule changes.

4.2 Detect suspicious behavior under valid access

The question is not "is this login valid?" but "is this valid credential behaving normally?" High-value behavioral signals, drawn from the campaigns above:

! **Read and query spikes:** Salesforce API events exceeding a high row-count threshold (SlashID's example uses more than 10,000 rows over seven days); bulk ARM reads on subscriptions and resource groups beyond an application's normal pattern (JadePuffer).

! **Anomalous service principal sign-ins** from an IP, ASN, or geography never previously seen for that identity.

! **High-volume** `ListKeys` against storage accounts, and any attempt to delete Azure Site Recovery or Backup locks — treat the latter as an active incident.

! **Post-grant reconnaissance burst:** a fresh token reading contacts, manager, direct reports, and sent items within seconds of issuance (EvilTokens).

4.3 Monitoring OAuth grants and application credentials

The consent and application layer needs its own monitoring: a new or unknown `client_id` requesting high-risk scopes, or a first-party client requesting an unusual resource; credentials added to existing applications or service principals (the Storm-0940 persistence move, T1098.001); Primary Refresh Token issuance or device registration immediately after a grant; and malicious inbox rules that auto-file invoice, wire, or payment mail. The illicit consent grant analysis maps each of these to a detection stage.

4.4 Containment: tokens, grants, credentials, and sessions

A password reset does not revoke a stolen token. Effective containment revokes the service principal grant, the refresh tokens, and the PRT, and ends sessions; because a standard revoke can leave an access token valid for about an hour, disable the account during an active intrusion. Rotate any exposed secret immediately. Require multi-party approval for destructive actions such as mass device wipes, and use just-in-time activation with short windows for Global Admin, Intune Admin, and Privileged Role Admin.

4.5 Campaign-to-control matrix

CAMPAIGN	PRIMARY DETECTION SIGNAL	HIGHEST-VALUE PREVENTIVE CONTROL
Salesloft Drift	OAuth API events with high row counts; anomalous user agents	Short-lived tokens; least-privilege integration scopes; revoke-on-anomaly
Storm-0940 / ARTEX	Service principal sign-ins across many tenants; app credential additions	Admin-only consent; limit multi-tenant app trust; rotate app credentials
JadePuffer / Storm-3168	Bulk ARM reads; high-volume `ListKeys`; backup-lock deletion attempts	Least-privilege RBAC; resource locks on storage and backup; secret rotation
Stryker	More than five Intune wipes in minutes; Graph device actions at scale	JIT privileged activation; multi-party approval for mass wipes; FIDO2 or Mutual TOTP
EvilTokens / Kali365	Post-grant recon burst; unknown `client_id` with high-risk scopes; malicious inbox rules	Block device-code grant via Conditional Access; verified-publisher consent; passkeys
GTG-1002 / CrowdStrike actors	Impossible request rates; automation-pattern API bursts	Behavioral anomaly detection on service identities; segment agent access

CAMPAIGN	PRIMARY DETECTION SIGNAL	HIGHEST-VALUE PREVENTIVE CONTROL
GTG-20006	VPN credential abuse; account-server mass export	Phishing-resistant MFA on VPN; privileged-access monitoring on directory servers

APPENDIX A /

Campaign and Platform Reference Table

ITEM	CATEGORY	PRIMARY SOURCE
GTG-1002	AI-orchestrated state espionage	Anthropic / THN
GTG-2002	"Vibe hacking" extortion	Anthropic
GTG-20006	Self-rewriting AI workflow; identity-record theft	Anthropic / Microsoft
Salesloft Drift (UNC6395)	OAuth token supply-chain theft	SlashID
Storm-0940 / ARTEX	OAuth app abuse via ConnectWise	Microsoft
JadePuffer / Storm-3168	First documented agentic ransomware	SlashID
Stryker (Handala)	Intune mass-wipe via Graph	SlashID
CrowdStrike actors 1–4	Autonomous intrusion harnesses	CrowdStrike
PROMPTFLUX / PROMPTSTEAL	Just-in-time AI-enabled malware	Google GTIG
EvilTokens / Kali365	AI-driven device-code PhaaS	SlashID
ARTEX	Go-based agentic pentest framework	Microsoft
PentestGPT / PentAGI	Open-source autonomous pentest agents	Open source
ClickFix	Paste-and-run initial access	MITRE T1204.004

APPENDIX B /

MITRE ATT&CK Mappings

TECHNIQUE	NAME	OBSERVED IN
T1195.002	Compromise Software Supply Chain	Salesloft Drift, Storm-0940
T1078 / T1078.004	Valid Accounts / Cloud Accounts	All campaigns; JadePuffer, Stryker
T1528	Steal Application Access Token	Salesloft Drift, Stryker
T1550.001	Application Access Token	Storm-0940
T1098.001	Account Manipulation: Additional Cloud Credentials	Storm-0940
T1098 / T1098.003	Account Manipulation / Additional Cloud Roles	Stryker
T1136.003	Create Account: Cloud Account	Stryker (rogue admin)
T1110.003	Brute Force: Password Spraying	Storm-0940
T1087 / T1069	Account / Permission Groups Discovery	ARTEX
T1526	Cloud Service Discovery	JadePuffer
T1552 / T1552.001	Unsecured Credentials / Credentials in Files	Salesloft Drift, JadePuffer
T1204.004	Malicious Copy and Paste	ClickFix
T1059.001	Command and Scripting: PowerShell	ClickFix
T1530	Data from Cloud Storage	Salesloft Drift, Stryker
T1567	Exfiltration Over Web Service	Stryker, GTG-20006
T1485 / T1561.002	Data Destruction / Disk Wipe	JadePuffer, Stryker
T1490	Inhibit System Recovery	JadePuffer
T1651	Cloud Administration Command	Stryker (Intune wipe)

// FREE ASSESSMENT

Would these attacks work in your environment?

Every campaign in this report exploited the same thing: identities with more access than they needed. AI agents find and abuse that access in minutes, not weeks.

Our free, read-only assessment finds those gaps first:

- / Every user, NHI and agentic identity, and every resource each one can touch
- / The grants and app permissions that carry the most risk
- / Excess access you can safely remove to limit escalation and lateral movement

Get your free assessment →

// READ ONLINE
slashid.com/blog/agentic-ai-identity-attack-surface →

slash/id

© 2026 SlashID · slashid.com