INTELLIGENCE BRIEFING
Security Command Center
TLP:CLEAR
2026-07-08 14:46 UTC

Dialogflow CX Trust Boundary Failure: How a Rogue Agent Turned Google's AI Platform Against Its Users

SECURITY ANALYSIS | HIGH | CVSS 7.5

SCC Item ID	SCC-STY-2026-0337
Type	Security Analysis
Severity	HIGH
CVSS Base Score	7.5
Affected Products	Google Dialogflow CX (enterprise chatbot deployments; specific version range not confirmed in available sources)
Published	2026-07-07T16:36:43
Discovery Source	Rss

Executive Summary

A trust boundary failure in Google Dialogflow CX, discovered by Varonis and responsibly disclosed in late 2025, allowed a rogue agent to intercept enterprise chatbot conversations and exfiltrate data without authorization. Google has issued a patch, but the incident exposes a broader risk: AI platforms built on multi-agent architectures inherit the access control weaknesses of their least-secured component. For CISOs, this is a signal that AI pipeline components require the same scrutiny applied to APIs and microservices, trust between agents must be explicit, scoped, and auditable.

Technical Analysis

Varonis researchers identified a trust boundary failure in Google Dialogflow CX that permitted injection of a rogue agent into enterprise chatbot pipelines. According to Varonis's disclosure and the GHSA advisory (GHSA-p5gx-f9rx-95rw), the flaw involved an authentication bypass component combined with insufficient access controls and improper trust delegation between agent components.

The attack surface is the multi-agent handoff model that Dialogflow CX uses to route conversation flows. When an agent delegates to another, passing session context, user data, and conversation state, the platform did not adequately verify that the receiving agent was authorized to handle that data. A malicious actor who could introduce or substitute an agent into this pipeline could, according to Varonis, intercept or manipulate those conversation flows and exfiltrate data from AI-driven interactions.

The CWE profile maps the failure layers clearly: CWE-284 (Improper Access Control) and CWE-863 (Incorrect Authorization) describe the access model failure at the agent handoff; CWE-441 (Unintended Proxy/Intermediary) describes the mechanism by which the rogue agent positioned itself in the data path; CWE-668 (Exposure of Resources to the Wrong Sphere) describes the consequence, conversation data reaching an unauthorized recipient.

Mapped MITRE ATT&CK techniques include T1190 (Exploit Public-Facing Application), T1199 (Trusted Relationship), T1557 (Adversary-in-the-Middle), T1530 (Data from Cloud Storage), T1078 (Valid Accounts), and T1059 (Command and Scripting Interpreter). The T1199 and T1557 mappings are particularly relevant: the attack exploited the trusted relationship between agents and positioned the rogue component as an intermediary, a classic adversary-in-the-middle pattern adapted to AI pipeline architecture.

No in-the-wild exploitation has been confirmed by a second independent source, and no threat actor has been attributed. Google issued a patch following Varonis's responsible disclosure. The source quality score for this item is 0.64 (reflecting a mix of researcher disclosure, official documentation, and corroborating news coverage), with Varonis blog (tier 3), official Google documentation (tier 1), GHSA advisory (tier 1), and Dark Reading report (tier 2) providing corroboration. The precise version range affected has not been confirmed in available source material.

The industry implication extends beyond Dialogflow. Enterprise deployments of AI platforms increasingly rely on multi-agent systems such as LangChain, AutoGen, Google Agentspace, and others, where inter-agent trust is a design assumption, not a verified runtime control. This disclosure is an early datapoint suggesting that agent-to-agent authentication will become a recurring vulnerability class as these architectures become more prevalent in production environments.

Action Checklist

1. Step 1: Assess exposure, determine if your organization deploys Google Dialogflow CX in any enterprise chatbot, virtual agent, or customer-facing AI pipeline; include third-party integrations that may route through Dialogflow CX under the hood
2. Step 2: Apply the patch, confirm that affected Dialogflow CX deployments have received Google's patch issued following Varonis's disclosure; consult the GHSA advisory (GHSA-p5gx-f9rx-95rw) for remediation scope
3. Step 3: Audit agent trust configurations, review Dialogflow CX security settings per Google's official documentation (https://docs.cloud.google.com/dialogflow/cx/docs/concept/security-settings) to verify that agent handoff permissions are explicitly scoped and not over-permissioned; align with NIST AC-3 (Access Enforcement) and AC-6 (Least Privilege)
4. Step 4: Review access controls on AI pipeline components, verify that each agent component operates under the principle of least privilege (NIST AC-6); implement CIS Controls v8 3.3 (Configure Data Access Control Lists) for any data stores accessible by conversation agents; apply CIS Controls v8 6.3 (Require MFA for Externally-Exposed Applications) to administrative interfaces
5. Step 5: Enable audit logging on agent interactions, confirm that conversation flow events, agent handoffs, and data access events are captured per NIST AU-2 (Event Logging) and AU-12 (Audit Record Generation); retain logs per AU-11 (Audit Record Retention) to support forensic review if exploitation is later identified
6. Step 6: Update your threat model, add agent-to-agent trust boundary failure as a threat pattern in your AI/ML system threat register; map to T1199 (Trusted Relationship) and T1557 (Adversary-in-the-Middle) in

your ATT&CK coverage review

7. Step 7: Extend scrutiny to other agent orchestration platforms, if your organization uses LangChain, AutoGen, or similar multi-agent frameworks, assess whether equivalent inter-agent authentication controls exist; treat this disclosure as a signal for a broader architectural review

8. Step 8: Brief leadership, communicate to CISO and relevant business owners that AI pipeline components carry access control risks equivalent to APIs; frame the risk around data exfiltration from customer-facing AI interactions, not just traditional application vulnerabilities

9. Step 9: Monitor for follow-up disclosures, track GHSA-p5gx-f9rx-95rw and Varonis's blog for additional technical detail; watch for a formal CVE assignment; monitor CISA advisories for any escalation in exploitation status

IR / Forensic Enrichment

Triage Priority	URGENT
Escalation Criteria	Escalate to immediate priority and initiate breach notification assessment if GCP Cloud Audit Logs reveal unauthorized agent handoff events or anomalous Data Access log entries (particularly reads against data stores containing PII, PHI, or financial data) during the exposure window between Varonis's late-2025 disclosure and confirmed patch application, or if the vulnerability is added to the CISA KEV catalog indicating active exploitation in the wild.
Recovery Notes	After confirming Google's platform-layer patch is applied and agent trust configurations have been explicitly scoped, restore normal Dialogflow CX operations while maintaining elevated log monitoring on `google.cloud.dialogflow.cx.v3` Data Access audit logs for a minimum of 30 days post-remediation, specifically watching for anomalous agent handoff sequences or unexpected data store access patterns that could indicate a persistent compromise predating the patch. Verify that all service account IAM bindings modified during containment are reflected in the current IAM policy export and have not been re-elevated through any automated provisioning pipeline. Retain all audit log exports from the exposure window for a minimum of one year, or longer if regulatory obligations (GDPR, HIPAA, PCI-DSS) govern the data processed by the affected chatbot deployments.

Forensic Artifacts	GCP Cloud Audit Logs — Data Access log type for `dialogflow.googleapis.com`: captures agent API calls, session creation, agent handoff events, and entity type access; query for `protoPayload.methodName` values containing `DetectIntent`, `StreamingDetectIntent`, or `TransitionRouteGroup` originating from unexpected service account principals during the exposure window | GCP Cloud Audit Logs — Admin Activity log type for `SetIamPolicy` on Dialogflow CX projects: documents any IAM permission changes on agent service accounts that may indicate attacker persistence or privilege escalation beyond the initial trust boundary failure | Dialogflow CX agent configuration exports (JSON): the `transitionRoutes`, `targetFlow`, and `targetPage` definitions within each agent's exported configuration document the trust relationships that existed at the time of exposure; preserve pre-patch exports as forensic baselines showing which agent handoff paths were implicitly trusted | GCP service account access token audit trail: GCP Cloud Audit Logs for `GenerateAccessToken` and `SignJwt` calls on service accounts bound to Dialogflow CX roles identify whether agent credentials were used to access downstream data stores (e.g., Cloud Storage, Firestore, BigQuery) beyond the expected conversation flow scope | Dialogflow CX conversation history exports: if conversation logging was enabled, raw conversation turn data including webhook request/response payloads and fulfillment calls can be exported via the Dialogflow CX API (`sessions.detectIntent` response logs) and may contain evidence of data exfiltration content or anomalous fulfillment targets introduced by a rogue agent

Per-Action IR Details

Step 1: Assess exposure — determine if your organization deploys Google Dialogflow CX in any enterprise chatbot, virtual agent, or customer-facing AI pipeline; include third-party integrations that may route through Dialogflow CX under the hood

NIST Phase: Preparation

Reference: NIST 800-61r3 §2 — Preparation: asset inventory and exposure assessment prior to confirmed incident

Controls: CIS 1.1 (Establish and Maintain Detailed Enterprise Asset Inventory), CIS 2.1 (Establish and Maintain a Software Inventory)

Compensating: Query cloud billing and IAM consoles directly: run `gcloud projects list` and `gcloud services list --enabled --filter='name:dialogflow'` across all GCP projects to surface Dialogflow CX API enablement. For third-party integrations, audit webhook endpoints and OAuth service accounts in the GCP IAM console for any service account with `dialogflow.` role bindings.

Evidence: This is a pre-incident assessment step and does not alter live state; no volatile capture required. Document GCP project IDs, Dialogflow CX agent names, and associated service account identifiers as the baseline inventory before any remediation actions are taken.

Step 2: Apply the patch — confirm that affected Dialogflow CX deployments have received Google's patch issued following Varonis's disclosure; consult the GHSA advisory (GHSA-p5gx-f9rx-95rw) for remediation scope

NIST Phase: Eradication

Reference: NIST 800-61r3 §3.4 — Eradication: eliminate the vulnerability from affected systems after containment is confirmed

Controls: CIS 7.3 (Perform Automated Operating System Patch Management), CIS 7.4 (Perform Automated Application Patch Management), CIS 7.2 (Establish and Maintain a Remediation Process)

Compensating: Dialogflow CX is a managed Google Cloud service — patching is applied by Google at the platform layer and is not operator-triggered via a traditional patch workflow. Confirm remediation status by reviewing the GHSA-p5gx-f9rx-95rw advisory directly at github.com/advisories/GHSA-p5gx-f9rx-95rw and cross-referencing the Google Cloud release notes for Dialogflow CX at cloud.google.com/dialogflow/cx/docs/release-notes for entries dated after the Varonis disclosure in late 2025.

Evidence: Before treating the system as remediated, capture current Dialogflow CX agent configuration exports via `gcloud dialogflow cx agents export` and preserve GCP Cloud Audit Logs (Data Access log type: `google.cloud.dialogflow.cx.v3.Agents`) covering the period between initial Varonis disclosure and confirmed patch application. These logs document whether any unauthorized agent handoff events occurred in the exposure window and must be preserved before any configuration changes are made.

Step 3: Audit agent trust configurations — review Dialogflow CX security settings per Google's official documentation (docs.cloud.google.com/dialogflow/cx/docs/concept/security-settings) to verify that agent handoff permissions are explicitly scoped and not over-permissioned; align with NIST AC-3 (Access Enforcement) and AC-6 (Least Privilege)

NIST Phase: Containment

Reference: NIST 800-61r3 §3.3 — Containment: restrict the scope of a potential compromise by limiting what the threat can access or traverse

Controls: NIST AC-3 (Access Enforcement), NIST AC-6 (Least Privilege)

Compensating: Export agent configuration JSON for each Dialogflow CX agent using the GCP Console or `gcloud dialogflow cx agents export --agent= --output-file=agent_config.json`. Manually review the `transitionRoutes` and `targetFlow`/`targetPage` handoff definitions for any wildcard or implicit trust grants between agents. Flag any agent-to-agent handoff that does not require explicit caller identity verification as a misconfiguration.

Evidence: Before modifying any agent trust configuration, export and preserve the current agent configuration artifacts (agent_config.json exports) as forensic snapshots. Also capture GCP Cloud Audit Logs for `google.cloud.dialogflow.cx.v3.SessionEntityTypes` and session-level API calls to document existing handoff patterns in the pre-remediation state. Altering permissions destroys the evidentiary record of what over-permissioned paths existed.

Step 4: Review access controls on AI pipeline components — verify that each agent component operates under the principle of least privilege (NIST AC-6); implement CIS 3.3 (Configure Data Access Control Lists) for any data stores accessible by conversation agents; apply CIS 6.3 (Require MFA for Externally-Exposed Applications) to administrative interfaces

NIST Phase: Containment

Reference: NIST 800-61r3 §3.3 — Containment: reduce the blast radius of a trust boundary failure by enforcing access segmentation across AI pipeline components

Controls: NIST AC-6 (Least Privilege), CIS 3.3 (Configure Data Access Control Lists), CIS 6.3 (Require MFA for Externally-Exposed Applications), CIS 6.5 (Require MFA for Administrative Access)

Compensating: Use `gcloud projects get-iam-policy ` to enumerate all IAM bindings for service accounts associated with Dialogflow CX agents. Identify any service account holding `roles/dialogflow.admin`, `roles/datastore.user`, or broader data-access roles beyond `roles/dialogflow.client`. For each over-permissioned binding, replace with a custom IAM role scoped to the minimum required Dialogflow API methods. Enforce MFA on the GCP Console administrative interface via Google Workspace 2-Step Verification policy.

Evidence: Before revoking any IAM bindings, capture the full IAM policy export (`gcloud projects get-iam-policy --format=json`) and GCP Cloud Audit Logs for `SetIamPolicy` events on affected projects. These preserve the pre-remediation permission state. Also capture any active GCP service account key files in use, as revoking credentials without logging their prior scope eliminates evidence of what data stores were reachable by a rogue agent under the trust boundary failure.

Step 5: Enable audit logging on agent interactions — confirm that conversation flow events, agent handoffs, and data access events are captured per NIST AU-2 (Event Logging) and AU-12 (Audit Record Generation); retain logs per AU-11 (Audit Record Retention) to support forensic review if exploitation is later identified

NIST Phase: Detection Analysis

Reference: NIST 800-61r3 §3.2 — Detection and Analysis: establish log coverage to detect evidence of trust boundary exploitation and support retrospective forensic review

Controls: NIST AU-2 (Event Logging), NIST AU-12 (Audit Record Generation), NIST AU-11 (Audit Record Retention), CIS 8.2 (Collect Audit Logs)

Compensating: Enable GCP Cloud Audit Logs for Dialogflow CX at the project level: in the GCP Console, navigate to IAM & Admin > Audit Logs, select the Dialogflow API, and enable DATA_READ and DATA_WRITE log types. Export logs to a Cloud Storage bucket or BigQuery dataset with a retention policy of no less than 90 days (adjust to your regulatory requirement). For retrospective analysis without a SIEM, use `gcloud logging read 'resource.type="audited_resource" AND protoPayload.serviceName="dialogflow.googleapis.com"' --format=json` to query agent interaction events directly.

Evidence: This step enables logging and does not alter live session state, so no volatile capture is required before execution. However, document whether Data Access logging was previously disabled — the absence of logs for the exposure window is itself a forensic finding that must be recorded in the incident timeline, as it indicates potential blind spots during which rogue agent handoffs could have occurred undetected.

Step 6: Update your threat model — add agent-to-agent trust boundary failure as a threat pattern in your AI/ML system threat register; map to T1199 (Trusted Relationship) and T1557 (Adversary-in-the-Middle) in your ATT&CK coverage review

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: incorporate new threat patterns discovered during an incident into organizational threat models and detection coverage

Compensating: Document the Dialogflow CX agent-to-agent trust boundary failure as a new threat pattern in your existing threat register (a spreadsheet or JIRA ticket is acceptable). For each multi-agent AI system in scope, create a data flow diagram annotating trust boundaries and handoff authentication mechanisms. Use the MITRE ATLAS matrix (atlas.mitre.org) as a supplementary reference for AI-specific attack patterns relevant to agent hijacking scenarios, alongside ATT&CK technique mapping.

Evidence: No live state is altered by this step; no volatile capture is required. Attach the agent configuration exports and IAM policy snapshots captured during earlier steps as supporting evidence in the threat model update, so the documented threat pattern is grounded in the actual configuration that was found vulnerable rather than a hypothetical.

Step 7: Extend scrutiny to other agent orchestration platforms — if your organization uses LangChain, AutoGen, or similar multi-agent frameworks, assess whether equivalent inter-agent authentication controls exist; treat this disclosure as a signal for a broader architectural review

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: use findings from one incident to drive proactive hardening of architecturally similar systems before they are exploited

Controls: CIS 7.1 (Establish and Maintain a Vulnerability Management Process), CIS 2.2 (Ensure Authorized Software is Currently Supported)

Compensating: Enumerate multi-agent framework dependencies by scanning application dependency manifests: `pip show langchain autogen` on Python hosts, or `grep -r 'langchain\|autogen\|semantic-kernel' requirements*.txt pyproject.toml` across application repositories. For each identified framework, review whether agent-to-agent calls are authenticated via explicit identity tokens or rely on implicit in-process trust — document findings in a one-page architectural risk memo for CISO review.

Evidence: This is a proactive review step that does not alter live state; no volatile capture is required. Preserve current dependency version snapshots and architectural diagrams as the baseline state before any remediation actions are taken on adjacent platforms, so changes can be audited against the pre-review configuration.

Step 8: Brief leadership — communicate to CISO and relevant business owners that AI pipeline components carry access control risks equivalent to APIs; frame the risk around data exfiltration from customer-facing AI interactions, not just traditional application vulnerabilities

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: communicate incident findings and systemic risk implications to organizational leadership to drive policy and investment decisions

Compensating: Prepare a one-page executive summary using the GHSA-p5gx-f9rx-95rw advisory and Varonis disclosure as source material. Quantify blast radius in concrete terms: number of Dialogflow CX agents deployed, volume of customer conversations processed per month, and data classifications (PII, PHI, financial) accessible by those agents. Attach the IAM policy export from Step 4 to demonstrate actual over-permissioned paths rather than hypothetical risk.

Evidence: No live state is altered; no volatile capture is required. Include the log gap documentation from Step 5 (whether Data Access logging was previously disabled) in the leadership brief — the absence of detection capability during the exposure window is a material risk factor that informs breach notification assessment and insurance reporting obligations.

Step 9: Monitor for follow-up disclosures — track GHSA-p5gx-f9rx-95rw and Varonis's blog for additional technical detail; watch for a formal CVE assignment; monitor CISA advisories for any escalation in exploitation status

NIST Phase: Post Incident

Reference: NIST 800-61r3 §4 — Post-Incident Activity: maintain situational awareness for evolving threat intelligence on a known vulnerability to enable timely response to exploitation escalation

Controls: CIS 7.1 (Establish and Maintain a Vulnerability Management Process)

Compensating: Configure a free GitHub Advisory Database watch on GHSA-p5gx-f9rx-95rw via GitHub's 'Watch' notification feature. Set a Google Alert for 'Dialogflow CX CVE' and 'GHSA-p5gx-f9rx-95rw' to catch CVE assignment or public PoC publication. Subscribe to CISA's Known Exploited Vulnerabilities (KEV) catalog RSS feed at cisa.gov/known-exploited-vulnerabilities-catalog to detect any escalation to active exploitation status. Review weekly; assign a named owner to the monitoring task.

Evidence: No live state is altered; no volatile capture is required. If a formal CVE is assigned or CISA adds the vulnerability to the KEV catalog, immediately re-enter the detection_analysis phase: re-query GCP Cloud Audit Logs for `google.cloud.dialogflow.cx.v3` events covering the full exposure window using the query from Step 5 to determine whether any exploitation activity preceded the organizational response.

Detection Guidance

No specific IOC values (C2 domains, payload hashes, IP addresses) appear in the available source material. The Varonis blog (varonis.com/blog/rogue-agent-dialogflow-attack) is the primary technical disclosure and may contain indicators not reproduced in the item data, consult that source directly for any published indicators.

For behavioral detection, focus on the adversary-in-the-middle and trusted relationship abuse patterns mapped to this vulnerability:

• Dialogflow CX audit logs: Review Google Cloud audit logs for unexpected agent invocations, agent handoffs to unrecognized agent IDs, or session transfers that deviate from approved conversation flow paths. Per NIST AU-6 (Audit Record Review, Analysis, and Reporting), establish a baseline of normal agent invocation patterns and alert on deviations.

• Data access anomalies: Monitor for unusual data retrieval events from cloud storage or backend integrations during conversation sessions (T1530, Data from Cloud Storage). Flag sessions where an agent accesses resources outside its expected data scope.

• Authentication events: Watch for account or service account activity consistent with T1078 (Valid Accounts), specifically, service accounts used by Dialogflow CX agents accessing APIs or data stores outside their normal operational pattern. Align with NIST AU-2 (Event Logging) for authentication event capture.

• Conversation flow integrity: If your deployment supports it, audit conversation transcripts for evidence of unexpected data exposure, unusual response content, or responses that appear to originate from an agent not in the approved flow. This aligns with NIST AU-14 (Session Audit).

• Proxy/intermediary indicators: Unusual latency in agent responses, unexpected intermediate hops in API call chains, or responses that include data not supplied by the user in the current session may indicate an intermediary agent in the pipeline (CWE-441 / T1557).

MITRE D3FEND countermeasures applicable to this attack pattern include D3-UAP (User Account Permissions) to restrict agent service account scope, D3-MFA (Multi-factor Authentication) on administrative and API access per CIS Controls v8 6.5, D3-LAM (Local Account Monitoring) to detect anomalous service account behavior, and D3-CRO (Credential Rotation) for any service account credentials that may have been exposed through the vulnerable pipeline.

Framework Mappings

MITRE-ATTACK

- **T1530** — Data from Cloud Storage
- **T1078** — Valid Accounts
- **T1059** — Command and Scripting Interpreter
- **T1190** — Exploit Public-Facing Application
- **T1199** — Trusted Relationship
- **T1557** — Adversary-in-the-Middle

NIST-800-53R5

- **AC-2** — Account Management
- **AC-6** — Least Privilege
- **IA-2** — Identification and Authentication (Organizational Users)
- **IA-5** — Authenticator Management
- **CM-7** — Least Functionality
- **SI-3** — Malicious Code Protection
- **SI-4** — System Monitoring
- **SI-7** — Software, Firmware, and Information Integrity
- **CA-8** — Penetration Testing
- **RA-5** — Vulnerability Monitoring and Scanning
- **SC-7** — Boundary Protection
- **SI-2** — Flaw Remediation
- **AC-3** — Access Enforcement

OWASP-TOP10-2021

- **A01:2021** — Broken Access Control

CIS-V8

- **6.1** — Establish an Access Granting Process

- **6.2** — Establish an Access Revoking Process
- **6.8** — Define and Maintain Role-Based Access Control
- **6.3** — Require MFA for Externally-Exposed Applications
- **7.3** — Perform Automated Operating System Patch Management
- **7.4** — Perform Automated Application Patch Management

SOC2-TSC

- **CC6.1** — The entity implements logical access security software, infrastructure, and architectures over protected information assets

HIPAA-SECURITY

- **164.312(a)(1)** — Access Control
- **164.312(d)** — Person or Entity Authentication

ISO-27001-2022

- **A.8.8** — Management of technical vulnerabilities

NIST-CSF-2

- **DE.AE-08** — Incidents are declared when adverse events meet the defined incident criteria

MITRE ATT&CK Mapping

Technique ID	Technique Name	Tactic
T1530	Data from Cloud Storage	Collection
T1078	Valid Accounts	Defense-Evasion
T1059	Command and Scripting Interpreter	Execution
T1190	Exploit Public-Facing Application	Initial-Access
T1199	Trusted Relationship	Initial-Access
T1557	Adversary-in-the-Middle	Credential-Access

Sources

Source	URL	Tier
Security News	https://www.darkreading.com/application-security/dialogflow-cx-rogu...	T2
Security settings | Dialogflow CX	https://docs.cloud.google.com/dialogflow/cx/docs/concept/security-s...	T1
Rogue Agent: How a Single Code Block Could Hijack Your ...	https://www.varonis.com/blog/rogue-agent-dialogflow-attack	T3

Source	URL	Tier
An authentication bypass vulnerability in Google Cloud...	https://github.com/advisories/GHSA-p5gx-f9rx-95rw	T1
Rogue Agent Flaw in Google Dialogflow CX Enabled Cross ...	https://www.mallory.ai/stories/019f3cb3-d98b-73cf-84f3-f6d7cd545b00	T3

Generated 2026-07-08 14:46 UTC by TJS Security Command Center