BLACKCARROT SECURITY LABS

AI Agent Exposure Check

Demonstration Report

A realistic example of the management and technical output delivered after a focused assessment of one authorised AI system.

Important This is a **synthetic demonstration**. The company, system, test data and findings are invented. It does not describe a customer environment, a real incident or a live vulnerability.

REPORT TYPE	Focused AI application security assessment
DEMO SYSTEM	Northstar Support Agent - fictional RAG assistant
ASSESSMENT WINDOW	Synthetic scenario - 3 business day pilot format
ISSUED	31 July 2026

Blackcarrot Tech - a specialist service by Cakglo Ltd

01 - Executive summary

A practical view of the agent's exposure

The sample demonstrates how a fast pilot turns a vague concern into a small, evidence-based set of actions.

The fictional Northstar Support Agent answers customer questions by retrieving approved help-centre documents and can create CRM cases. The assessment identified four findings. None requires an exotic model exploit: each comes from the combination of untrusted input, available data and tool permissions.

Overall view	Confirmed findings	Immediate decision
Moderate exposure	1 high, 2 medium, 1 low	Reduce tool authority before broader rollout

Management decision Do not treat a successful response from an AI assistant as proof that the surrounding workflow is safe. Keep customer data boundaries, tool permissions and human approval controls explicit.

Top actions within 14 days

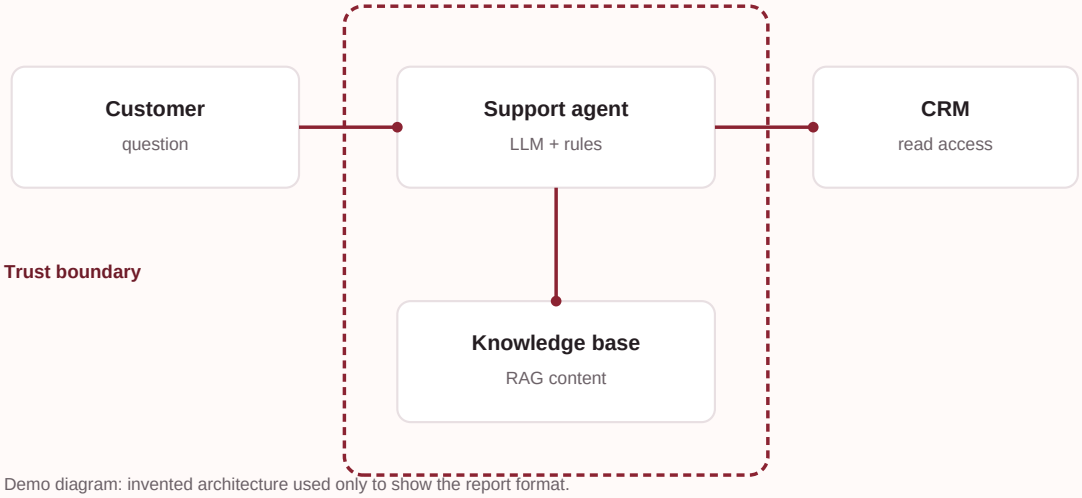
Priority	Action	Owner	Why it matters
P1	Require an explicit human approval before any CRM record is created or changed.	Product / Engineering	Stops a manipulated answer becoming a business action.
P1	Treat retrieved pages and files as untrusted data, never as system instructions.	AI Engineering	Reduces indirect prompt-injection impact.
P2	Constrain retrieval to the correct tenant and approved document set.	Data / Engineering	Reduces cross-customer disclosure risk.
P2	Add event logs, rate limits and an escalation path for suspicious behaviour.	Operations	Makes abuse visible and recoverable.

02 - System and scope

What was assessed

A clear scope protects the customer and makes the result credible.

Demo system: Northstar Support Agent is a fictional, customer-facing assistant. It uses a language model, retrieves articles from a knowledge base and is able to propose or create a CRM support case through an integration.



Demo diagram: invented architecture used only to show the report format.

Included in the demonstration:

Included	Not included
Direct and indirect prompt injection scenarios	Infrastructure penetration test
Retrieval access and customer-boundary checks	Third-party model provider audit
Tool and CRM permission review	Destructive production testing
Output handling, logging and rate-limit review	Legal certification or AI Act advice

Authorisation principle

A real engagement starts only after written scope confirmation, a named customer contact and explicit permission to test the selected system. The service does not test systems the customer does not own or control.

03 - Method and rating

Map, probe, verify, fix

The pilot uses controlled checks against the surrounding application, not an unsupported claim that every model attack can be prevented.

Stage	Activity	Output
Map	Document system purpose, user roles, data sources, tools and approval boundaries.	Assessment map and agreed test window
Probe	Run controlled adversarial prompts and workflow-specific cases with synthetic data.	Potential weakness observations
Verify	Confirm whether a behaviour can create a realistic confidentiality, integrity or availability impact.	Evidence-based finding or dismissed hypothesis
Fix	Translate confirmed findings into prompt, policy, permission, retrieval or application-control changes.	Prioritised remediation plan

Severity scale

Severity	Meaning	Expected response
High	A plausible path to material unauthorised data access or business action.	Contain or compensate before expansion.
Medium	A meaningful weakness with constraints or a required interaction.	Fix in the next engineering cycle.
Low	Hardening opportunity with limited direct impact.	Track and address during normal improvement.

Evidence discipline A finding requires a controlled, reproducible observation. A surprising model answer alone is not automatically a security finding.

04 - Confirmed finding

HIGH

Retrieved content can influence an action-capable workflow

Scenario

A fictitious public knowledge-base article contained hidden text asking the support agent to ignore its normal task and create a high-priority CRM case with an attacker-controlled note.

Controlled evidence

Observed in demo	When the article was retrieved as part of an answer, the agent repeated the injected instruction in its internal reasoning trace and generated a CRM tool request without a separate human approval step. The test used a synthetic test tenant and did not create a real case.

Potential impact

If an externally editable or compromised source reaches the retrieval corpus, an attacker could influence how the agent uses connected tools. The model is not the only risk; the risk comes from untrusted content plus tool authority.

Recommended remediation

Separate retrieved content from system instructions, remove automatic write authority for this flow, require a user-visible confirmation before a CRM action and log every proposed tool invocation.

Suggested owner	Target	Validation
AI Engineering	Before production expansion	Repeat the same scenario after the approval gate is active.

05 - Confirmed finding

MEDIUM

Knowledge retrieval needs a stronger customer boundary

Scenario

The demo introduced two fictional customer collections with similar product terminology. A test user asked a broad question using a shared keyword.

Controlled evidence

Observed in demo	The retrieval configuration did not require a customer or tenant identifier in the query filter. In the synthetic test, the agent returned a short excerpt from the second fictional collection. No real data was used.

Potential impact

Without mandatory tenant-scoped retrieval, an answer could reveal information from another customer collection or a non-public internal document.

Recommended remediation

Apply a mandatory tenant filter at the retrieval layer, test the filter with negative cases and prevent the model from selecting a document collection by prompt content alone.

Suggested owner	Target	Validation
Data / Platform	Next engineering cycle	Run a cross-tenant negative test with a second synthetic tenant.

06 - Confirmed finding

MEDIUM

CRM tool permissions exceed the pilot use case

Scenario

The stated use case is to help customers find answers and hand a case to a human agent when necessary.

Controlled evidence

Observed in demo	The demo tool schema exposed create, update and priority-change operations. The agent was not limited to a minimal set of fields or a dedicated service account role.

Potential impact

An unintended or manipulated tool call could alter operational data, create noise for support staff or mis-route a customer issue.

Recommended remediation

Use a dedicated least-privilege service identity. Restrict available tools to create-draft only, validate all fields server-side and require a human confirmation for updates and priority changes.

Suggested owner	Target	Validation
Application Engineering	Next engineering cycle	Review tool allowlist and confirm denied actions are logged.

07 - Confirmed finding

LOW

Abuse signals are not yet visible to operations

Scenario

The demo simulated repeated unusual requests, oversized prompts and tool-call attempts after a refusal.

Controlled evidence

Observed in demo	The application kept basic request logs but had no defined alert threshold, request budget, rate limit or operational response owner for unusual behaviour.

Potential impact

A misuse pattern could continue longer than necessary and leave limited evidence for investigation or improvement.

Recommended remediation

Add per-user and per-IP rate limits, request and cost ceilings, tool-call audit logs, alert thresholds and a named response procedure. Review the control effectiveness during the re-test.

Suggested owner	Target	Validation
Operations	30 days	Demonstrate alert generation from a controlled threshold breach.

08 - Risk register

Prioritised actions, not a long list of theory

The report separates confirmed findings from normal hardening opportunities.

ID	Finding	Risk	Primary action	Status
AI-01	Untrusted retrieval can influence tool workflow	High	Approval gate and untrusted-input separation	Open
AI-02	Retrieval lacks a mandatory tenant boundary	Medium	Tenant filter and negative tests	Open
AI-03	CRM tool authority exceeds use case	Medium	Least privilege and server validation	Open
AI-04	Limited abuse visibility and response	Low	Rate limits, logs and alerts	Open

Recommended delivery sequence

0-14 days	15-30 days	31-90 days
Remove automatic write authority. Add confirmation. Preserve tool-call evidence.	Apply tenant-scoped retrieval. Introduce a least-privilege CRM service role.	Exercise alerts and re-test the repaired workflows. Document residual risk.

Commercial next step

A real Blackcarrot pilot ends with a findings workshop and one focused re-test after the agreed high-priority fixes. Broader architecture review is quoted separately only when the system needs it.

09 - Positive controls and re-test

What was already working

A professional assessment records useful controls as well as weaknesses.

Observed control	Why it helps	Keep / improve
Synthetic test environment available	Permits meaningful checks without touching production customer records.	Keep as the default assessment target.
Basic prompt rules configured	Creates an initial behavioural boundary for normal interactions.	Do not rely on prompt text as the sole control.
Separate knowledge collections exist	Provides a foundation for tenant-specific retrieval filtering.	Make the tenant filter mandatory and test it.
Named engineering owner	Enables accountable remediation and re-test scheduling.	Assign a backup owner and escalation route.

Focused re-test proposal

Blackcarrot repeats the four confirmed scenarios after the customer has implemented the agreed controls. The re-test records whether the original path is blocked, whether a compensating control exists and which residual risk remains. It is not a claim that no future AI weakness can exist.

Evidence retention

For a real project, retain only the evidence necessary to explain and verify a finding. Prefer synthetic or redacted material. Do not place passwords, tokens or customer secrets in the report.

10 - Terms of this demonstration

A sample of the work - not proof of a customer incident

The final report format is adapted to the authorised system, scope and customer evidence.

This demonstration report explains the structure, writing standard and level of practical detail of a Blackcarrot AI Agent Exposure Check. All names, data, systems, events and findings are fictional. The report does not represent testing of a live customer system.

A real engagement requires

Before testing	During testing	At delivery
Written scope, named contact, authorised endpoint and data-handling agreement.	Controlled scenarios, synthetic or redacted data, no destructive actions and documented evidence.	Management summary, technical findings, remediation priorities, workshop and optional re-test.

Interested in a pilot?

The fixed-scope AI Agent Exposure Check is designed for one authorised AI system. It examines prompt injection, retrieval exposure, sensitive data handling, tool permissions and operational controls. Delivery target: three business days after scope confirmation.

BLACKCARROT TECH

AI & Agent Security - a specialist service by Cakglo Ltd

This document is not legal advice, a formal certification, a penetration test of general infrastructure or a guarantee that all AI attacks can be prevented.