DEMO REPORT: INTENTIONALLY UNSAFE WORKFLOW

Sample AI Workflow Security Review

Prepared by	CSINT Research
Generation date	2026-07-31
Source workflow	`workflows/unsafe-support-agent.json`
Workflow name	Demo - Unsafe Support Agent
Scanner version	n8n-ai-agent-security-lab 1.2.0
Review type	Static path analysis with an isolated runtime contract comparison

Scope

This sample reviews one exported n8n workflow: an intentionally unsafe customer support agent that receives public webhook input, passes it to an AI agent node, sends an email and calls a URL taken from workflow data. The fixture is published for teaching and testing. It ships deactivated and was never connected to a production system.

Result summary

Static score: 10/100 (F)

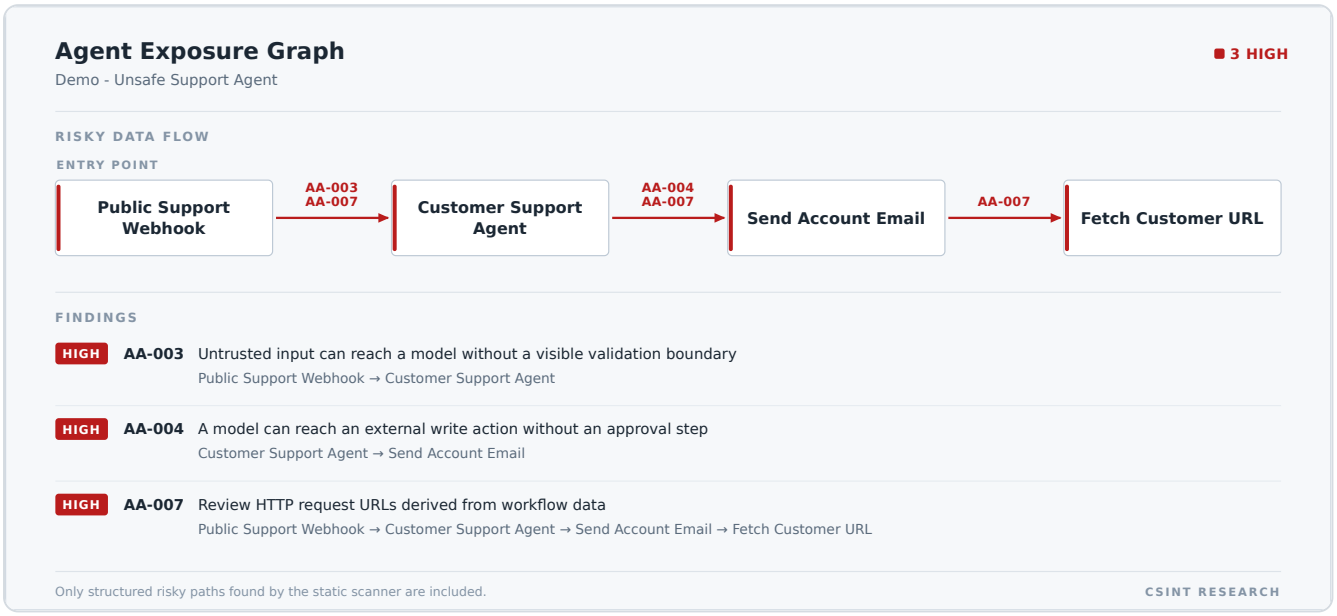
The static review found 9 items: 0 critical, 4 high, 4 medium and 1 low. Four findings sit on a single connected path from the public entry point to an outbound HTTP request. The five most important findings are documented on page 3, and the hardened comparison on page 4 shows the same workflow after remediation.

Limitation

> A static score is a review aid, not proof of exploitation and not proof of security. Static analysis cannot see runtime authorization, credential scopes, upstream controls or model behavior. This document demonstrates the review format on a fixture that is unsafe on purpose; a real report covers a redacted customer workflow.

Exposure graph

The scanner turns structured risky paths into a single figure. Boxes are workflow nodes. Arrows are data flow proven from the export. Finding IDs sit on each risky edge, and the ledger below the flow maps every ID and severity back to the exact path.



How to read finding IDs and paths

- **AA-003** on the first edge: untrusted webhook input reaches the model with no validation boundary in between.
- **AA-004** on the second edge: model output reaches an email action with no approval step.
- **AA-007** spans the whole flow: the final HTTP request URL is derived from data that entered at the public webhook.
- The same IDs appear in the Markdown report and in SARIF output, so a finding can be tracked from the graph to code scanning.

> Only structured risky paths found by the static scanner are drawn. The graph is not a full workflow diagram, and a node that does not appear here still needs manual review.

Five most important findings

HIGH **AA-003 Untrusted input can reach a model without a visible validation boundary**

Path: `Public Support Webhook → Customer Support Agent`

Remediation: add a deterministic validation step before the model. Enforce size, type and allowlist rules, separate instructions from data, and test direct and indirect prompt injection cases.

HIGH **AA-004 A model can reach an external write action without an approval step**

Path: `Customer Support Agent → Send Account Email`

Remediation: require explicit human approval for messages, writes and account changes. Use least-privilege credentials for the final action.

HIGH **AA-006 No rate or cost boundary was detected before model usage**

Evidence: `External inputs: Public Support Webhook`

Remediation: add per-user and per-origin limits, a maximum input size, a model-call budget, timeouts and a bounded retry policy.

HIGH **AA-007 Review HTTP request URLs derived from workflow data**

Path: `Public Support Webhook → Customer Support Agent → Send Account Email → Fetch Customer URL`

Remediation: parse URLs with a real URL parser, allowlist schemes and destinations, block private and metadata IP ranges, and disable redirects unless required.

MEDIUM **AA-002 Public webhook authentication is not enforced by the trigger**

Evidence: `Webhook nodes without native authentication: Public Support Webhook`

Remediation: use header, JWT or basic authentication at the webhook boundary. If custom validation is retained, reject before processing and rate-limit failures.

> The remaining findings (AA-005 output validation, AA-008 timeouts and retries, AA-009 audit record, AA-010 failure route) are listed in the full Markdown report.

Hardened comparison

Workflow	Static result	Findings
Demo - Unsafe Support Agent	10/100 (F)	9 findings, 4 high severity
Demo - Hardened Support Agent	93/100 (A)	1 medium finding kept for manual review

What changed

- The webhook now requires authentication before any processing.
- A validation node enforces input rules and a rate limit before the model.
- Model output is checked against a schema before it is used.
- A human approval step sits between the model and the email action.
- The outbound fetch only accepts allowlisted destinations.
- External actions are written to an audit log, and a failure route exists.

What remains for manual review

One medium finding stays open on purpose. `AA-007`: the hardened workflow still derives an outbound URL from workflow data. The workflow validates the scheme and host first, but the scanner keeps every dynamic URL path visible so a person confirms the allowlist is correct. The score was not forced to 100.

> This mirrors how a real review ends: automated checks close most paths, and the items that need human judgement are named instead of hidden.

Runtime contract result

The regression gate sends synthetic requests to an isolated staging copy of the hardened workflow and checks the behavior that must not change. Result: **8/8 checks passed.**

Check	Expected staging behavior	Result
Missing authentication	Reject	PASS
Invalid staging signature	Reject	PASS
Missing or unexpected fields	Reject	PASS
Prompt injection marker	Keep behind approval	PASS
Invalid approval token	Reject	PASS
Valid request	Queue without external action	PASS
Repeated request ID	Return the same operation	PASS
Unsupported method	Reject	PASS

The staging fixture contains no email, HTTP request, database or AI nodes. A valid approval only increments an isolated test counter, so the run ends with 0 simulated external actions.

What this test proves, and what it does not

- It proves the staging copy rejects unauthenticated, malformed and unapproved requests, and that a replayed request does not run twice.
- It does not prove model safety, credential scope, production authorization or the absence of vulnerabilities.
- It only tests the supplied workflow export against an explicitly allowlisted staging endpoint. Production systems are never touched.

Review boundaries

- This is a defensive static and configuration review with an isolated staging comparison. It is not a penetration test.
- It is not a compliance certification and does not map findings to any regulatory framework.
- It is not a security guarantee. A clean report means the reviewed paths showed no findings, nothing more.
- Production systems are not attacked, changed or connected to at any point.
- Findings reference public guidance: OWASP Top 10 for LLM Applications 2025 and the NIST AI RMF Generative AI Profile.

Recommendation

Fix the highest findings first: the injection path into the model, the missing approval step, the missing cost boundary and the dynamic outbound URL. Then rescan the export and rerun the runtime contract so the fixes are confirmed by the same checks that found the problems. Retest again after any later workflow change; the scanner is free and runs locally.

Pilot scope: one redacted workflow, manual review, remediation notes and one retest.