MEMO · SECURITY ARCHITECTURE TEAM · TO THE PROJECT COMMITTEE

The AI helpdesk agent is not the risk. *What it is allowed to do is.*

We are about to put an AI agent in front of IT support. It chats with employees, searches the knowledge base, opens tickets, starts password resets, and talks to our ITSM vendor through a connector. The main risk is not that the AI "gets tricked", it is what a manipulated agent can do once it acts. An agent that can only answer questions causes little damage; ours can touch tickets, passwords and employee data, so a chat window becomes a real attack surface.

Prepared for · The Project Committee **Scope** · AI helpdesk agent, its tools & the ITSM connector **Ask** · Sign off the controls as release blockers

01 THREAT MODEL

Who would attack it

Four ways a chat window in front of IT support becomes an attack surface. None of them require breaking the model, only using what the agent is allowed to do.

01 **Prompt-injection attacker**
Hides instructions in a message, a ticket or a document the agent reads, and makes the agent act on them.

02 **Malicious employee**
Social-engineers the agent instead of a human technician to get access they shouldn't have.

03 **Compromised vendor connector**
A trusted integration that starts feeding the agent poisoned data, or harvesting what we send it.

04 **The over-permissioned agent itself**
Not malicious, but with more permissions than any task needs, so every small manipulation becomes a big incident.

02 IMPACT

What can go wrong

The damage doesn't come from the AI being wrong. It comes from a manipulated agent using the tools and access we handed it.

1 **Data theft through the agent's own tools**
Hidden instructions make the agent look up employee accounts and leak the results. The exit door is a tool we gave it.

2 **Privilege escalation through tickets**
An employee gets the agent to create tickets with approval fields pre-filled, and downstream automation trusts the ticket. Nobody hacked IAM; the ticket system did the work.

3 **Resetting the wrong password**
"I'm locked out of jsmith's account, I'm covering for him." If the agent takes the target name from the message and never checks who is asking, that's an account takeover. This exact bug hit Meta's AI agent in production, letting anyone reset any Instagram account password.

4 **PII leaking from chat history**

Transcripts fill up with pasted passwords and personal data. Over-retained, over-shared or badly access-controlled history means one breach exposes all of it.

5 **Leaked credentials & exposed components**

The chatbot and its components (MCP servers, application servers, databases, storage) are internet-reachable and can be probed, exploited, or accessed with reused leaked credentials.

03 CONTROLS

How we stop it

THE PRINCIPLE

The agent never gets to authorize anything. It understands requests; the architecture decides.

Policy layer between agent and every tool

Every action passes authentication, an ownership check (does the requester own the account they're touching?) and rate limits. Anything ambiguous is denied.

Fresh verification for sensitive actions

A password reset only completes with MFA or a verified channel, so even a fully manipulated agent can't finish the job. Admin accounts are excluded from the agent entirely.

Tickets are requests, not approvals

The agent can write descriptions; approval, priority and requester identity are set by the platform, never by the AI.

Chat history is minimized

Redacted, isolated per user, encrypted, and deleted after 90 days.

Everything is logged and monitored

Every tool call is recorded; we alert on unusual patterns (bulk lookups, resets on accounts the user never logged into). A kill switch cuts the agent's credentials instantly.

No internet exposure, no long-lived credentials

The agent is kept off the internet, with managed identity instead of static secrets. If exposure is unavoidable: mandatory authentication and WAF protection.

04 RESIDUAL RISK

The risk picture

Every high-and-critical risk drops to a manageable residual once the controls above are in place. This is the risk you are being asked to accept.

RISK	BEFORE CONTROLS	AFTER CONTROLS	OWNER
Data theft via tool calls	CRITICAL	MEDIUM	CIO
Privilege escalation via tickets	HIGH	LOW	CIO
Wrong-password reset / takeover	CRITICAL	LOW	CIO
PII leak from chat history	HIGH	LOW	DPO
Compromised connector	MEDIUM	LOW	CIO
Leaked credentials / vulnerability exploitation	HIGH	LOW	CIO

THE DECISION

What we're asking you to sign

Three points. Together they make the controls a condition of go-live and put the residual risk on the record.

01 Controls are release blockers
Treat the controls above as mandatory. No go-live without them.

02 Accept the residual risk
The "after controls" column is the risk you are signing for.

03 Re-open on any change
Whenever the agent gains a new tool or connector, we run this analysis again.