Cyana Security Model: Executive Brief

A two page summary of how Cyana, SimpleHelp's AI assistant for remote support and RMM, is controlled. Written for owners, managers and decision makers. The full architecture, including every control's limits, is documented in the white paper *Cyana: Security and Control*.

Version: 1.0. Applies to SimpleHelp 6.0 and later.

What Cyana Is

Cyana gives technicians a natural language interface to the fleet, to live sessions and to Toolbox scripts, inside the SimpleHelp platform you already run. It gathers information, proposes actions and drafts scripts, and the technician reviews and decides what actions should be taken. It has no independent autonomous path to any endpoint: it reaches machines only through SimpleHelp's existing remote access service infrastructure, sees only the machines the technician can see, and honours the same permission model as every other SimpleHelp feature.

What It Changes

A question asked across a customer site (which machines are low on disk, which services failed overnight, which endpoints are behind on updates) replaces minutes of dashboard filtering and manual checks with one simple question and a structured summary. Inside a live session, a technician can have Cyana investigate a failing service and propose a fix for review, freeing the technician to work on other issues or perform critical checks and maintenance. In the Toolbox editor, Cyana drafts, adjusts, reviews and explains scripts from a plain language description. The companion document *Cyana: What It Does For Your Team* walks through these workflows in detail, while the rest of this brief covers how they are controlled.

The Question That Matters

Letting an AI act across an estate of customer machines is not acceptable without technician oversight and control. Cyana's answer is a layered architecture in which no layer depends on the model behaving well, and every layer is enforced in server code.

Two Layers Doing Two Different Jobs

Permissions are the security control. Administrators grant each technician group a set of Cyana permissions: file read, file write, command execution, process and service control, screenshots, tags, session control, web research. A capability that is not granted is not merely refused at run time, but is never offered to the model at all, and does not exist in Cyana's world for that group. This envelope binds everyone and everything, including tired technicians, junior technicians, misbehaving models, and compromised accounts, because it merges with the platform permissions that bound the account already. It is enforceable, auditable, centrally administered control.

Approval is the attention control. Inside the envelope, every meaningful action Cyana proposes is presented to the technician for explicit approval before it runs, and every chat starts in minimum capability / maximum approval mode. A technician may grant auto approve for specific categories, but only for the current chat, only after an explicit warning acknowledgement, never beyond what the administrator

permitted, and never persisted, so the next chat starts at maximum approval again. This layer exists to place a human decision, with full attention, in front of exactly the actions where a model error or an injected instruction would otherwise land.

Cyana deliberately keeps these two layers separate. There is no server switch to force manual approval of everything because decades of real world evidence (Windows UAC, browser TLS warnings, hospital alarm fatigue) show that mandatory routine prompting collapses into reflex clicking and destroys the attention it was meant to guarantee. Organisations that require a stricter posture should tighten the permission envelope rather than overload the technician with prompts, since the envelope is enforced in server code while excessive prompts collapse to a reflex under load.

The Additional Safeguards

- **Automated review of higher risk actions.** Commands, scripts, writes, deletes and service control are reviewed by a separate AI agent with no tools of its own, which returns a green, yellow or red rating. A red rating forces explicit technician approval even over an auto approve grant, and if review is configured but unavailable the system fails closed to red. The action review agent deliberately never sees tool results from the conversation, as this means the same injected content cannot reach both agents directly. The reviewer instead checks whether the proposed action is consistent with what the technician asked, which is where erroneous and injection driven actions reveal themselves.
- **Scope enforcement.** Every conversation has a technician chosen scope (a machine group, a subset, a single session), and every tool call is validated against it on the server before dispatch.
- **Outbound data loss prevention.** Arguments to any outbound capable tool are scanned for known protected server configuration values before dispatch, and detections block the call and write to the administrator activity log.
- **Isolation of web content.** Web research runs in a sandboxed agent whose only tool is a single page fetch, with no access to the conversation, machines or files, together with a strict URL policy and an output integrity check.
- **Full audit trail.** Every prompt, proposal, review rating, approval, denial and result is persisted in SimpleHelp's standard history, subject to standard view permissions, and exportable to PDF per conversation.

What the Controls Do Not Do

Cyana's documentation deliberately states the limits of each control. In brief: the outbound scan matches exact known server secrets rather than general secrets or PII, and is one layer among several rather than a sufficient control on its own. Marking tool output as untrusted is a signal to the model rather than a parser level guarantee and automated review is a secondary layer of protection rather than a decision maker. Web research is compartmentalised rather than claimed to be injection proof and the audit record does not currently capture which model produced each individual turn. Section 12 of the full paper enumerates every such boundary and the administrator responsibility that pairs with it. These limits are inherent to LLM driven tooling generally, and the difference here is that they are documented rather than left for the reader to discover.

Your Model, Your Data Path

Cyana runs against a model endpoint the organisation chooses: frontier providers such as Anthropic, OpenAI and Google Gemini natively, or any OpenAI compatible endpoint, including fully self hosted models on your own infrastructure, enabling fully private local data handling in which no prompt content ever leaves your network. Chat and review can use different models from different providers, so no single model is a single point of failure in the safety model.

Adopting It Safely

There is no migration involved in adopting Cyana. Cyana ships in SimpleHelp 6.0, a trial AI licence can be requested from the licence panel of a linked server, and there is nothing new to deploy to endpoints. Enable Cyana for a small group of experienced technicians, start in full approval mode with read only workflows such as fleet health, inventory and service state, and keep automated review on from day one. Expand what runs unprompted only as the team's observed confidence in the specific model grows, and broaden the administrator envelope last, deliberately, and per group. The full paper's Section 11 sets this out step by step, including the stricter variant for regulated environments.

Answering Client and Auditor Questions

Clients, insurers and auditors increasingly ask whether AI touches their systems and how it is controlled. Cyana lets you answer with a documented control architecture, a per action audit trail, optional full self hosting and a comprehensive security paper, rather than a policy document written after the fact.

Summary

Cyana cannot reach a machine the technician cannot reach, cannot be asked to do anything the administrator has not permitted, cannot act without technician approval except within bounds the technician has explicitly and temporarily granted, has every higher risk action rated by an independent reviewer, and writes everything it does into the audit record. AI can bring enormous benefits and leverage to your technicians, but Cyana's security model does not depend on the AI behaving well at all times.