

技术白皮书

AI 与 MCP 安全运营白皮书

如何让 AI 助手与外部 Agent 安全地参与防火墙运营，并管住企业自己的大模型与 AI Agent。

目录

- 01 摘要
- 02 AI 进入安全运营的机会与风险
- 03 NGFW One 的 AI 设计原则
- 04 内置 AI 助手
- 05 MCP：把防火墙接进 AI 工作流
- 06 管住企业自己的 AI
- 07 落地路线图
- 08 治理清单

关于本白皮书

本白皮书为草案版本，描述 NGFW One 产品的设计理念与规划能力。功能清单、版本归属、型号规格、菜单名称与参数均以产品正式发布版本和产品规格书为准。

01

摘要

AI 正在从「回答问题」走向「替人做事」。对安全团队而言，这意味着两件事：一是用 AI 提升运营效率，二是管住 AI 本身带来的新风险。

NGFW One 在这两方面同时布局：

- **用 AI 做安全：**内置 AI 助手支持自然语言查日志、一句话生成策略、告警降噪与自动周报；AI 自动研判对事件做上下文调查并按剧本处置。
- **让 AI 安全地参与：**内置 MCP Server 让 Claude 等 AI 客户端与自建 Agent 调用防火墙能力，写操作默认需人工确认，每次调用进入审计。
- **管住企业的 AI：**AI 应用防火墙管控员工使用大模型时的数据外泄；Agent 与 MCP 安全网关管控企业内部 AI Agent 的工具调用。

本白皮书介绍上述能力的设计原则、安全模型与落地路线，适合安全负责人、运维负责人与 AI 平台团队阅读。

02

AI 进入安全运营的机会与风险

机会

安全设备功能越来越多，真正懂全部配置的人很少。
AI 能把「翻十几个页面查日志」变成一句话提问，把重复的研判和周报交给机器，让小团队也能做到 7×24 响应。

风险

AI 可能误判；Agent 可能被网页或文档中的隐藏指令诱导；员工可能把敏感数据交给外部大模型；AI 调用如果不留痕，事后无从追责。

风险	典型表现	NGFW One 的应对
AI 误操作	AI 生成了错误策略或误封业务地址	写操作默认生成草稿、人工确认；保护白名单对 AI 同样生效；一键回滚
越权调用	AI 客户端拿到了超出需要的权限	API Key 按角色授权；只读 Key 看不到写工具；IP 白名单与频率限制
提示词注入	恶意内容诱导 Agent 执行危险操作	高风险操作强制审批；工具描述投毒检测；自建 AI 应用的注入防护
数据外泄	员工把合同、源码、客户信息提交给大模型	AI 应用防火墙识别、脱敏或拦截敏感数据
无法追责	不知道 AI 何时、为谁、做了什么	所有 AI 查询、MCP 调用与执行动作进入审计日志

03

NGFW One 的 AI 设计原则

原则 1

AI 给草稿，人做决定

AI 生成的策略、处置建议都以草稿呈现，确认后才生效。只有管理员为特定规则或 Key 明确授权时才自动执行。

原则 2

权限不放大

AI 助手继承当前管理员的权限；MCP 调用只能执行该 API Key 角色允许的操作，AI 不会获得比人更多的权限。

原则 3

一切留痕

每一次 AI 查询、生成、应用和 MCP 调用，都记录调用方、参数、结果与时间，可在审计日志中按 Key 筛选。

原则 4

护栏优先于智能

保护白名单、审批流、频率限制和回滚机制对人和 AI 一视同仁，AI 的判断再好也不能绕过护栏。

数据与隐私

AI 推理方式以产品正式说明为准。后台可关闭云端 AI 功能，关闭后其他安全功能不受影响。所在单位对数据出境或外发有合规要求的，请在启用前与我们确认部署方案。

04

内置 AI 助手

能力	示例	安全机制
自然语言查询	「过去 24 小时被封停最多的 10 个 IP 是哪些？」	转换为日志检索，结果附原始日志链接，便于核对
一句话生成策略	「只允许财务网段在工作日访问 ERP 服务器的 443 端口」	生成策略草稿，提示冲突，确认后应用或交他人审核
告警降噪	把上千条扫描告警归并为少数几个事件	每条聚合结果可查看判定依据，可一键建例外
自动周报	每周推送攻击趋势、封停统计、新设备与待办	支持邮件与 Webhook 推送
自动研判与处置	夜间事件自动调查，给出结论、证据链与剧本建议	按置信度分全自动 / 半自动 / 只建议，动作可回滚

使用建议

- 提问时包含时间范围、对象（IP、站点、设备）与期望输出。
- 涉及封停和策略变更的结论，应结合原始日志核实。
- 可在「系统 → 管理员」中为特定角色关闭 AI 助手。

05

MCP：把防火墙接进 AI workflow

MCP（Model Context Protocol）是让 AI 调用外部工具的开放协议。NGFW One 同时提供 MCP Server 与 MCP Client，双向打通 AI 与防火墙。



图 1 双向 MCP：外部 AI 通过 MCP Server 管理防火墙；内置 AI 助手通过 MCP Client 调用外部工具

MCP Server 工具

工具	说明	类型
query_alerts	按时间、级别、来源查询告警	只读
search_logs	检索访问、攻击、审计日志	只读
list_blocked	列出当前封停的 IP / 域名 / 设备	只读
get_device_inventory	获取内网设备资产清单	只读
get_traffic_top	获取流量排行	只读
block_ip	封停指定 IP，可设时长	写入（默认需确认）
unblock_ip	解除封停	写入（默认需确认）
add_domain_rule	新增域名黑白名单规则	写入（默认需确认）
create_policy_draft	生成策略草稿，管理员审批后生效	写入（草稿）
generate_report	生成指定时间段的安全报告	只读

实际可用工具以设备版本和 API Key 角色为准，客户端连接后通过 list_tools 获取。

接入方式

方式	地址 / 命令	适用场景
远程 (Streamable HTTP)	<code>https://<设备IP>:9443/mcp</code>	支持远程 MCP 的客户端、服务器上运行的 Agent
本地 (stdio)	<code>ngfw-mcp --host <设备IP> --api-key <KEY></code>	Claude Desktop 等以本地进程方式加载 MCP 的客户端

安全模型

- **认证与授权：**每次调用使用 API Key 认证，只能执行该 Key 角色允许的操作，只读 Key 看不到写工具。
- **写操作人工确认：**封停、解封、修改策略默认进入「待确认」队列，管理员批准后才执行。
- **按角色放开：**SOAR 等确定性场景可为特定 Key 放开免确认，同时配置 IP 白名单、频率限制与较短的过期时间。
- **保护白名单仍然生效：**白名单中的地址无法通过 MCP 封停。
- **MCP Client 数据外发：**接入外部 MCP 服务时，AI 助手可能发送 IP、域名、告警摘要等查询数据；只接入可信服务，逐个工具授权，调用同样进入审计。

06

管住企业自己的 AI

除了用 AI 管防火墙，企业还需要管住员工与业务系统里的 AI。NGFW One 为此提供两项能力。

AI 应用防火墙

能力	说明
AI 应用识别与分级	识别主流大模型的网页、客户端与 API 调用，按允许 / 仅限部分人员 / 禁止分级管控
敏感数据脱敏	检测身份证号、手机号、银行卡号、源代码、合同关键字，脱敏后放行或拦截
提示词注入防护	保护企业自建的大模型应用，识别越狱、指令覆盖与间接注入
输出内容审查	对自建应用的回答做敏感信息与违规内容检查
用量与合规审计	谁、在什么时间、用了哪个 AI、提交了哪一类数据，全部可查

Agent 与 MCP 安全网关

能力	说明
MCP 服务登记	统一登记允许使用的 MCP 服务器，发现未登记的「影子 MCP」
细粒度授权	按 Agent、用户、工具和参数范围授权，例如只允许查询、不允许删除
高风险操作确认	删除、转账、外发、改权限等操作自动挂起，等待负责人审批
工具投毒检测	检测工具描述与返回内容中的隐藏指令和异常变更
调用审计与回放	每次调用的发起者、参数和结果完整留痕，可按会话回放

两者的区别

NGFW One 的 MCP Server 让 AI 来管理防火墙；Agent 与 MCP 安全网关则站在企业所有 AI Agent 与 MCP 服务之间，管理这些 Agent 能做什么。前者是「AI 用防火墙」，后者是「防火墙管 AI」。

07

落地路线图

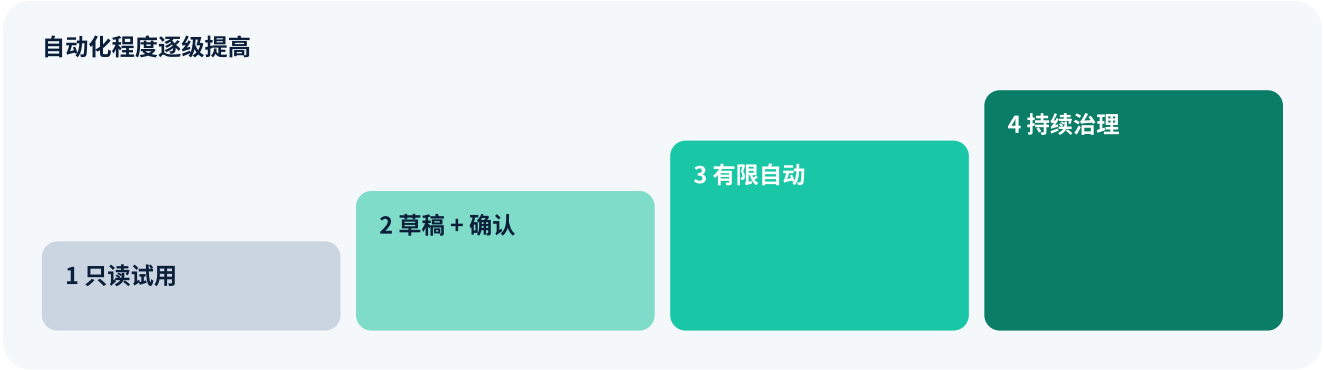


图 2 AI 接入安全运营的四个阶段：先只读，再草稿确认，最后对少数场景放开自动执行

阶段	目标	关键动作
1 只读试用	让团队熟悉 AI 助手与 MCP	开启 AI 助手查询与周报；为 Claude Desktop 创建只读 Key 并限制来源 IP
2 草稿 + 确认	AI 参与日常操作但不越过人	为运维 Key 开放写工具，所有写操作走待确认队列；AI 生成的策略先交他人审核
3 有限自动	对确定性场景秒级处置	对蜜罐即封等少数场景放开免确认；启用 AI 自动研判的半自动剧本
4 持续治理	防止权限与风险累积	每月复盘审计日志；回收不用的 Key；同步上线 AI 应用防火墙与 Agent 安全网关

08

治理清单

以下清单可作为 AI 参与安全运营的上线与复盘检查项：

- ☐ 每个 AI 客户端、Agent、SOAR 集成使用独立的 API Key，按最小权限分配角色
- ☐ 每个 Key 配置 IP 白名单、过期时间与调用频率限制
- ☐ 保护白名单覆盖出口网关、DNS、监控、堡垒机与合作方接口
- ☐ 写操作默认需人工确认；免确认仅限书面批准的少数场景
- ☐ MCP Client 只接入可信服务，逐个工具授权，评估数据外发合规
- ☐ 明确云端 AI 的启用范围；有数据出境要求的单位已确认部署方案
- ☐ 员工使用大模型的规则已制定，AI 应用防火墙策略与之一致，并已告知员工
- ☐ 企业内部 AI Agent 已在安全网关登记，高风险工具已设置审批
- ☐ 每月复盘审计日志中的 AI 查询、MCP 调用与自动处置记录

了解更多

AI 助手配置：ngfw.cn/docs/ai-assistant/ MCP 接入：ngfw.cn/docs/mcp/ AI 安全配置：ngfw.cn/docs/config-ai-security/ API 在线调试：ngfw.cn/docs/api-console/