AI 已成必然趋势，安全运营才是真正挑战

依托工程管理理念，让 AI 在安全、可控的环境中稳定发挥效能

人工智能全面落地已是行业定局，如今行业面临的核心难题，不再是是否应用 AI，而是如何实现安全、合规、可控的 AI 常态化运营。本文结合网络运维领域现状，剖析 AI Coding、AI 智能体带来的全新安全风险，并指明行业优化方向。

现状：AI Coding 已规模化投入生产

由人工智能生成的代码，如今已正式走进企业生产环境，这一趋势无法逆转。即使在超过 15 年、有着极为严格的开发管控标准的开源项目 OpenStack 平台中，开发者也开始大量提交 AI 辅助编写、甚至完全由 AI 生成的代码补丁，部分代码已经随版本完成上线交付。目前，全球各类开发团队都在普遍使用 AI 生成代码。

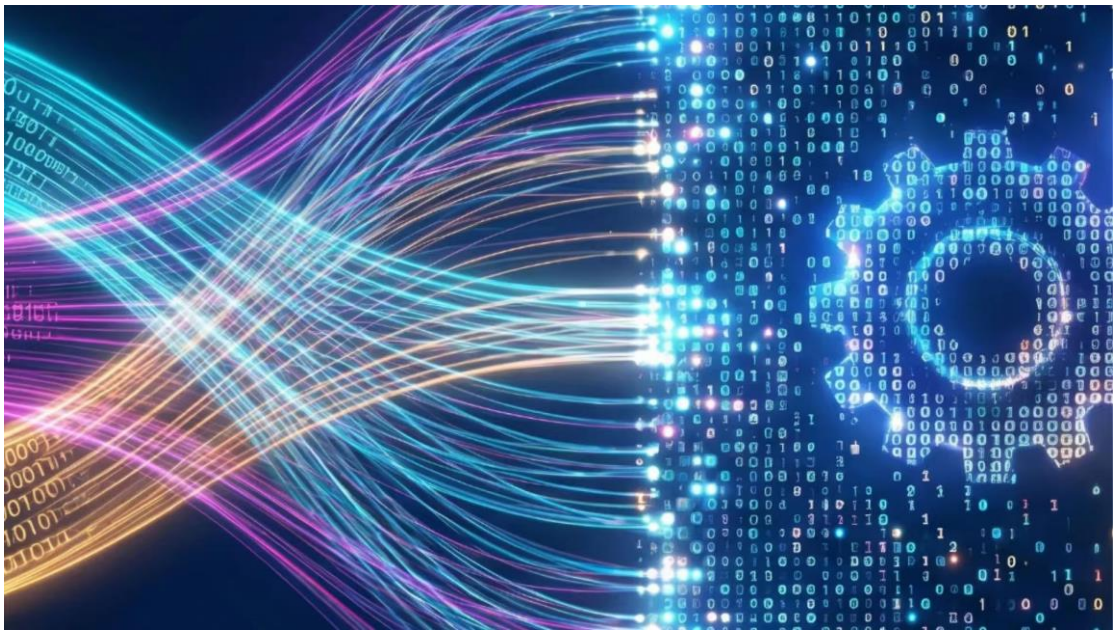
从技术角度来看，代码语法规则固定、逻辑框架具备可预测性，恰好契合生成式 AI 的技术优势，因此 AI 编写代码的效率十分突出。但效率提升的背后，隐患也随之而来：所有 AI 产出的代码补丁，都必须经过人工审核，核验代码逻辑正确性、运行安全性以及长期可维护性。AI 大幅降低了代码编写门槛，待审核代码与补丁数量急剧增加，作为代码上线前最后一道防线的人工审核团队，正承受着巨大的工作压力，传统审核流程已难以适配 AI 时代的代码产出规模。

核心风险：AI 智能体颠覆传统安全信任体系

AI 生成代码带来的审核压力只是安全挑战的一部分，AI 智能体的普及，正在彻底打破行业沿用多年的传统安全信任机制，这也是当下最严峻的风险点。

长期以来，IT 行业建立了成熟的安全容错机制，例如核心操作需多重密钥授权、重要审批流程设置多人联签等，以此规避单点风险。但现阶段，企业在落地 AI 智能体时普遍较为仓促，常常直接向智能体开放邮箱、数据库、生产环境等核心资源的全部访问权限。这类智能体仅在初期接收人类简单指令，后续便全程自主运行，一旦出现操作失误或被恶意利用，往往要在损失造成后才能被发现，属于典型的事后补救型风险。

目前行业存在明显的技术脱节问题：AI 工具的功能迭代速度，远远领先于配套安全技术的发展。针对 AI 智能体的行为隔离、操作审计、故障回滚、精细化权限分配等核心安全能力均不完善。多数企业没有为智能体设置分级权限，而是直接授予全量操作权限，安全边界形同虚设。补齐这一系列安全短板，已成为行业亟待探索的全新创新方向。



行业标杆实践：诺基亚的 AI 安全运营思路与落地方案

针对 AI Coding 泛滥、AI 智能体权限失控、安全管控缺失等行业共性难题，诺基亚结合自身网络、运维、研发技术积淀，形成了完整的理念、技术框架与商用实践，为 AI 安全运营提供了可落地的参考路径。

（一）核心理念：拒绝黑盒运行，建立可控自治体系

诺基亚提出玻璃盒（Glass Box） 安全准则，明确所有 AI 及智能体的自主行为必须

做到可观测、可溯源、可解释、可回滚，坚决杜绝无边界的黑盒操作。同时严格推行最小权限原则，根据职能对 AI 智能体进行分级授权：普通运维智能体仅开放查询、监控权限；故障处置类智能体仅可执行预设修复动作，严禁跨域修改核心配置；网络扩容、业务切片下线、核心网变更等高风险操作，必须经过人工二次确认，AI 无法自主执行。

此外，诺基亚认为传统静态安全防护、纯人工审核模式已无法匹配 AI 的迭代速度，主张构建以 AI 防御 AI 的闭环安全体系，依托分级智能体完成常态化安全巡检、行为监控与风险处置，形成 “AI 操作、AI 监督、AI 兜底” 的全新防护逻辑。

（二）技术产品方案：全方位化解两大核心风险

1. 应对 AI 智能体权限失控、操作不可控问题

依托 NSP 智能体 AI 框架与 AgenticOps 智能体运维体系，搭建全域管控底座：

- 安全边界隔离：为所有 AI 智能体设置策略围栏，提前划定可操作范围与禁止操作清单，一旦指令越界将被系统直接拦截，避免智能体随意访问数据库、生产设备等核心资源；
- 全链路可解释追溯：完整留存智能体的推理逻辑、决策依据与每一步操作指令，运维人员可清晰查看行为链路，实现全程审计；
- 事务级一键回滚：将 AI 批量变更封装为原子事务，校验发现异常时可全网自动回退，快速止损；
- 多智能体分工监督：按监控、排障、安全巡检、资源调度划分专用智能体，单一主体不具备全链路权限，同时设置上层监督智能体，实时拦截越权行为并触发告警。

配套 NetGuard 安全智能体实现主动威胁防御，7×24 小时比对 AI 操作行为基线，精准识别异常访问、批量篡改配置等恶意行为；同时校验容器镜像，拦截存在漏洞、被篡改的程序文件，防范 AI 被恶意利用的风险。

2. 缓解 AI 生成代码 / 配置审核压力

针对海量 AI 产出内容上线难题，诺基亚打造多层级校验与审核体系，融入研发运维流水线：

- 数字孪生前置仿真：基于 EDA AIOps 平台，让 AI 生成的配置、脚本、代码先在虚拟环境中完整运行测试，自动排查逻辑漏洞、策略冲突与安全隐患，从源头拦截风险内容，大幅削减人工初审工作量；
- 智能安全筛查：在计费、网络编排等平台内置大模型校验模块，自动识别代码中高危接口调用、数据未脱敏、越权访问等问题，并生成风险报告，标注隐患点位；
- 分级审核机制：低风险配置、常规代码经自动化校验后可自动放行，核心底层代码、关键网络配置则强制人工终审，兼顾效率与安全；
- 版本化管理：对所有 AI 生成资产进行版本留存，实现全生命周期溯源，便于问题复盘与追溯。

（三）商用落地案例

相关安全运营方案已在国内外多家运营商落地，并验证了实际价值：

国内运营商落地：联合 A 省某运营商部署数据安全智能体与数字孪生底座，严格限制 AI 智能体的数据访问权限，自动审计批量操作行为，每年大幅削减人工安全审核工作量，保障千万级用户数据安全；携手 B 省某运营商上线 MAS 多智能体协同系统，各智能体权责分离、无法单独改动核心参数，所有操作留痕可追溯，在工单处理效率大幅提升的同时，实现 AI 零误操作故障。

海外运营商落地：海外运营商 Reddot 采用 Deepfield Genome Shield 防护方案，依靠 AI 智能体抵御 AI 驱动的新型 DDoS 攻击，在网络边缘阻断恶意流量，替代传统人工值守；诺基亚旗下 Altiplano、Corteca 等宽带网络平台，也全面嵌入受控 AI 工具，网络规划、故障诊断类 AI 仅可执行查询与轻度优化，设备割接、下线等高风险操作必须人工确认。

（四）标准化落地步骤

诺基亚总结出一套通用落地流程，分为三大阶段：

- 前置约束：上线 AI 工具与智能体前，明确安全边界、黑白名单与访问权限，从源头杜绝无限制授权；
- 中间校验：依靠数字孪生、自动化检测工具完成前置风险筛查，仅将高风险内容转交人工复核；
- 后置兜底：搭建全量审计日志、行为追溯体系，搭配监督智能体与一键回滚能力，形成完整安全闭环。

解决思路：沿用工程规范，适配 AI 安全体系

面对 AI 带来的全新安全挑战，行业无需从零搭建体系，可借鉴数十年来保障复杂软件系统稳定运行的成熟工程规范，并结合 AI 的特性进行迭代优化，打造适配 AI 场景的安全运营模式。

强化全流程审核机制——针对 AI 生成代码，进一步细化审核标准，区分常规业务代码、核心底层代码的审核强度，借助工具辅助筛查基础漏洞，减轻人工负担，同时保留人工终审环节，守住代码安全底线。

完善 AI 智能体管控规则——摒弃“全权限开放”的粗放模式，推行最小权限原则，根据业务需求为智能体划分权限范围；搭建全流程行为审计系统，实现智能体操作可追溯、行为可监控；配套故障一键回滚功能，在异常发生时快速止损。

构建专属安全防护体系——围绕 AI 全生命周期搭建安全架构，覆盖代码生成、智能体部署、日常运行、故障处置等各个环节，把安全要求前置到 AI 应用设计阶段，从源头降低安全风险。

结语

人工智能融入研发、运维、生产全链路已是大势所趋。技术本身并非阻碍行业发展的难题，安全运营能力的缺失，才是制约 AI 价值释放的核心瓶颈。当下，整个科技行业需要正视 AI Coding、智能体带来的新型安全威胁，依托成熟的工程管理理念，加快安全技术与管控规范的落地，让 AI 在安全、可控的环境中稳定发挥效能。