

## 版权声明

北京启明星辰信息安全技术有限公司版权所有，并保留对本文档及本声明的最终解释权和修改权。

本文档中出现的任何文字叙述、文档格式、插图、照片、方法、过程等内容，除另有特

摘、翻译成其它语言，将其全部或部分用于商业用途，部分进行有限摘录、复制、修改、传

本文档依据现有信息制作，其内容如有更改，恕不另行通知。

## 免责申明

北京启明星辰信息安全技术有限公司在编写该文档的时候已经尽最大努力保证其内容准

北京启明星辰信息安全技术有限公司不对本文档中的错误、不准确、或错误导致

的损失和伤害承担责任。

## 信息反馈

如有任何问题，请联系：

公司：北京市海淀区东北旺西路8号中关村软件园21号楼启明星辰大厦 邮编：100193 电话：010-82779088

传真：010-82779000

公司网站地址：www.venustech.com.cn

# 目录


1 公司简介 ..... 6

2 背景与挑战 ..... 8

2.1 背景分析 ..... 8

2.2 面临挑战 ..... 9

2.2.2 缺少大模型资产使用的合规性持续监测评估 ..... 10

3 [illegible] ..... 13

3.1 AI-R-SOCC 定义 ..... 13

3.2 建设思路 ..... 14

4 [illegible] ..... 16

4.1 [illegible] ..... 16

4.1.1 [illegible] ..... 16

4.2 [illegible] ..... 19

4.2.1 大模型资产实体定义 ..... 19

4.2.2 企业/单位自建大模型 ..... 20

.3 内部私搭大模型.....20 4.2

4.3 大模型安全分析.....21

4.3.1 大模型风险关联分析.....21

4.3.2 基于用户身份的行为分析.....22

4.3.3 大模型安全图谱分析.....22

4.4 智能告警降噪.....23

4.4.1 安全事件.....23

.....24 4.4.2 智能降噪.....

.....25 4.5.1 大模型自身风险监测

.....26 4.5.2 风险人员监测

.....27 4.5.3 风险行为/事件监测

.....27 4.5.4 智能治理建议生成

.....28 4.6 行为审计溯源

.....28 4.7 大模型安全态势呈现

.....30 5 部署和典型应用场景

5.1 场景一：大模型应用中实时风险管控.....30

5.3 场景三：运行期间的大模型自身安全性.....33
5.4 场景四：影子大模型监测与治理.....35
6 [illegible].....36
6.1 智能运营.....36
6.2 [illegible].....37
6.3 [illegible].....37
6.4 安全专家.....37
6.5 全天候值守.....37

# 1 公司简介

的、拥有 启明星辰公司成立于 1996 年，由留美博士严望佳女士创建，是国内最具实力的

10 年 6 月 23 日，启明星辰在深交所中小板正式挂牌上市。 20

启明星辰拥有完善的专业安全产品线，横跨防火墙/UTM、入侵检测管理、网络审计、

终端管理、加密认证等技术领域，共有百余个产品型号，并根据客户需求不断增加。启明星 续

辰解决方案为客户的安全需求与信息安全产品、服务之间架起桥梁，将客户的安全保障体系 展

与信息安全核心技术紧密相连，帮助其建立完善的安全保障体系。

自 2002 年起，启明星辰就持续保持国内入侵检测、漏洞扫描市场占有率第一，近年来，

发展成为国内统一威胁管理、安全管理平台国内市场第一位，安全性审计、安全专业服务市

场领导者。目前，公司在全国各省市自治区设立三十多家分支机构，拥有覆盖全国的渠道和

售后服务体系。

长期以来，启明星辰公司得到了党和国家领导人的关怀与鼓励。2000 年 1 月，江泽民

切视察启明星辰公司；2003 年 1 月，胡锦涛总书记亲 李岚清、曾庆红等党和国家领导人亲

博士。 切接见了启明星辰公司 CEO 严望佳博

获得国家规划布局内重点软件企业、国家火炬计划软 凭借多年来的潜心开发，启明星

秀企业，中国电子政务 IT100 强等荣誉，及拥有最高级别的涉及国家秘密的计算 件产业优

品信息系

星辰目前是我国规模最大的国家级网络安全研究基地，完成包括国家发改委产业化、 启明星

科技部 863 计划、国家科技支撑计划等国家级科研项目近百项，创造了百 示范工程、国家

的多项空白。

小学。

全产品和最佳实践服务，帮助客户全面提升其 IT 基础设施的安全性和生产效能，为打造和提升国际化的民族信息安全产业第一品牌而不懈努力。

2 背景
2.1 背景

- **数据孤岛化**：安全日志分散存储，无法实现跨系统攻击链追溯与合规审计；

子系统，构建“监测-分析-处置-优化”智能闭环，实现大模型应用的**全局可视、风险可**

**处置可溯。**

## 2.2 面临挑战

大模型的应用给企业带来便捷的同时也带来了新的安全边界问题，例如大模型

出的敏感信息泄漏、对大模型的注入攻击等，应对这些新问题企业会新增新型的大

手段基本针对某一单一的风险点，对于企业管理者满对众多大模型的安全

防护手段，但这些

1. 数据孤岛效应：

覆盖的私有API通道注入恶意指令时，MASB无法独立识别跨系统攻击链。

的隐蔽指令链）；

- **缺乏因果推理：**当模型输出泄露客户隐私时，现有工具无法追溯至具

源或违规访问行为。

3. 告警有效性无法保障：

- **误报率居高不下：**多子系统独立告警导致重复通知（如MAF和MAVA

—模型的异常行为），SOC团队日均处理告警量超千条，有效告警识别率不足20%。

分散的子系统

4. 性能与安全的

与业务连续性间取舍；

- **资源浪费严重：**多套子系统独立运行，硬件资源利用率不足40%，运维成本飙升

### 2.2.2 缺少大模型资产使用的合规性持续监测

1. 大模型合规评估监管不足：

步至所有子系统。

更新要求，企业难以及时同步

2. 资产不可见性：

调查，67%的企业无法完整识别内部大模型资产。

3. 权责管理缺失：

1. 响应机制缺乏统一协同：

2. 闭环治理缺失：

- **知识沉淀不足：**处置经验沉淀在人员头脑中，缺乏标准化剧本库供复用。

OCC 定义

3.1 AI-R-SO

个挑战，哈尔滨 依托AI安全管理与智能人机协同应用安全、数据安

面以模型为

个融合管理中心（AI-R-SOCC）如图：




在企业员工大量使用大模型辅助办公的背景下，大模型也成为了企业中的信息交互中心，

这势必带来了新的安全边界问题，给企业安全运营工作带来了新的挑战，AI-R-SOCC 依托

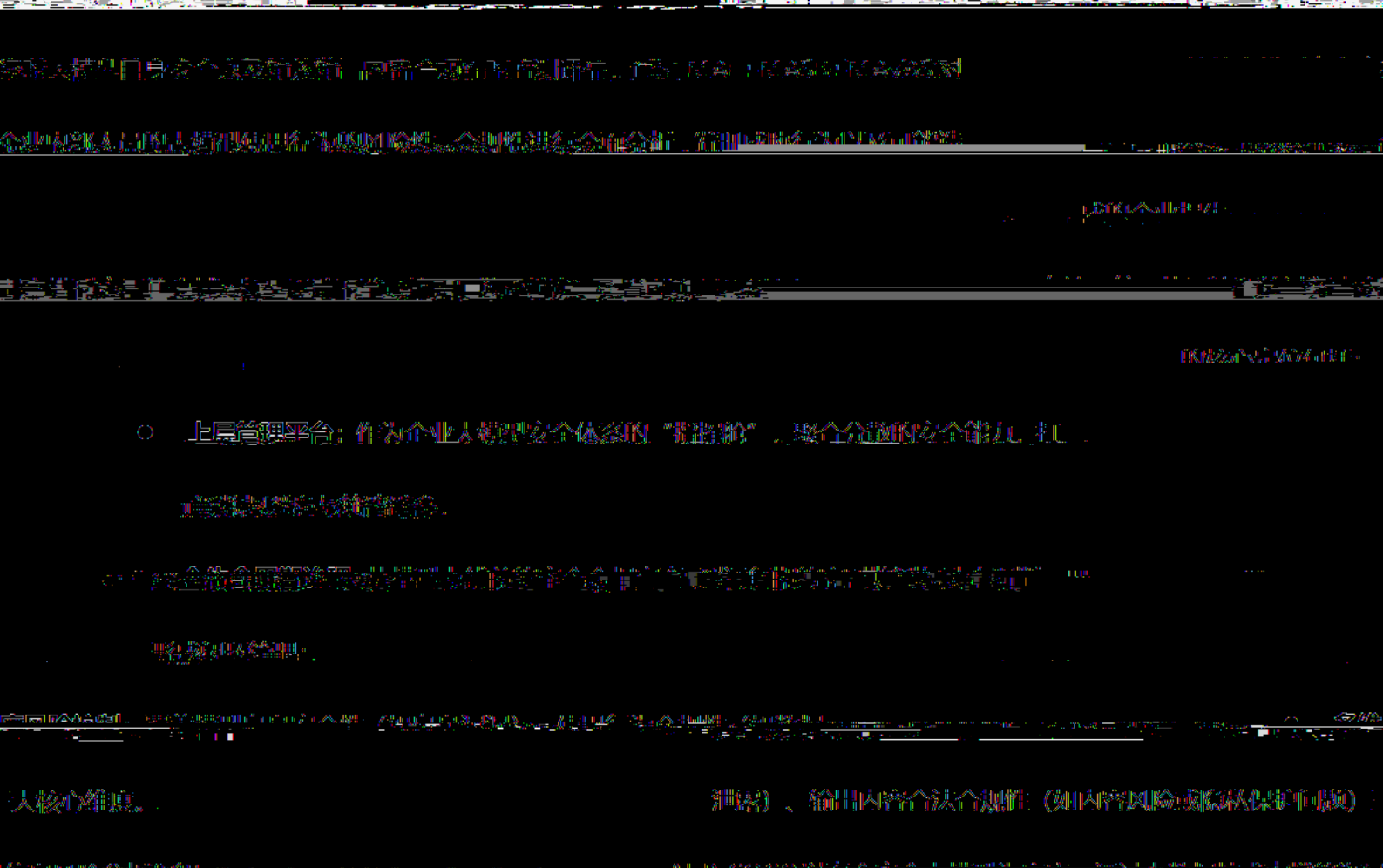
MAF、MASB、MAVAS等大模型安全防护技术，可形成对大模型应用自身安全

和公安AI-R-SOCC放之于企业网络安全整体视角中进行统一运营时，形成人机协同

## 3.2 建设思路

用的集中化安全治理以及智能化安全运营保障的中枢基 AI-R-SOCC 是企业级大模型应

入-运行-处置-审计”全生命 模型安全子系统（如 MAF、MASB、MAVAS）构建覆盖“准



- 上层管理平台：作为企业人机协作体系的“指挥部”，

人核心维度。 测）、输出内容合法合规（如内容风险、隐私泄露下钻）

# 4 核心能力

能体使安全人员能够在统一的界面上高效完成指令传递和执行，同时调用预设安全剧本，大

和执行安全指令，结合历史事件和知识库，智能体还能自动推荐处置剧本，帮助安全团队在

[illegible]

同时，系统提供对剧本执行情况和纳管设备的实时监控，智能识别设备的连通性、性能状态及潜在安全隐患，帮助用户提前发现问题并采取防范措施，提升整体安全运营的效率。

### 4.1.2 任务管理

任务管理提供了全面的任务调度与管理功能，涵盖人工任务、周期任务和脆弱性任务三[illegible]

- 人工任务管理

通过直观的界面展示待办任务详情，用户可以快速查看、分配和处理各项任务，确保任[illegible]

- 周期任务管理

提供自动化的任务调度功能，支持用户根据需求设置任务执行周期，并配置具体的执行动作，帮助用户高效管理日常定期检查与维护任务，减少人工干预，提升任务执行的自动化[illegible]

- 脆弱性任务管理

[illegible]户快速识别和修复系统中的安全漏洞，全面保障资产安全。

### 4.1.3 大模型安全风险智能响应

智能响应围绕脆弱性监测、核心任务、告警智能监测、智能研判和预判任务三大功能，

构建了高效的安全监控体系。系统通过实时分析和动态跟踪，协同实现对脆弱性、告警事件

的全面覆盖。在脆弱性监测能力方面，智能响应支持对资产的脆弱性进行持续监控，支持用户创建自定义监

测任务，并结合最新漏洞情报快速识别资产和业务系统中的潜在脆弱性风险，确保安全隐患

得到及时发现与处理。

- 智能监测

围绕脆弱性管理提供实时监控与任务跟踪能力，通过趋势图与列表展示脆弱性监测任务的执行情况，帮助用户高效掌控任务进展与资产状态。系统结合大模型的智能分析能力，将任务执行效果与关联脆弱性资产进行可视化呈现，为用户持续监测与动态分析提供数据支撑，确保脆弱性问题得到全面覆盖与及时跟进。

- 研判任务

警有效性判断、全路径分析以及画像分析，全面评估告警的威胁级别与潜在风险，帮助用户快速识别和优先处理高风险事件，聚焦关键威胁，减少不必要的干扰，显著提升事件响应效率。

- 大模型风险处置

对恶意使用、违规输出等场景，例如，预置联动剧本，覆盖大模型应用安

- **提示词攻击处置**：MAF 检测到恶意指令注入 → AI-R-SOCC 触发 MASB 冻结账号

敏感文件上传 → MAF 实时脱敏输出内容 → MAVAS

险 → 隔离高风险模型并告警数据治理部门。

- **数据泄露阻断**：MASB 识别

扫描关联模型训练集残留风

误封合法用户）。

## 4.2 全局资产管理

平台提供全面的资产管理管理能力，帮助企业“摸清家底”，通过多种适配器的有机融

资产状况，为风险管控、全局监测提供基础支撑。

资产主要包括：

资源和组件。这类实体资

的网络资源，提供完善的管理能力。

- **应用软件**：对大模型应用相关的应用、组件进行统一管理，包括数据库、数据

API服务管理：对大模型应用对外提供的服务进行全面的管理，支持对API访

问形式为进行合理的。

### 4.2.2 企业/单位自建大模型

提供大模型全生命周期管理，便于对此类大模型进行合理的管理。

映射企业员工人，支持实时监测（如最小化发送至审批审核）。

安全信息包括：大模型的基本安全性、合规性检测指标，使用中的风险行为及安全性

等维护工作。

监测。

全信息、运行性能监控信息的维护管理如上。

大模型涉及的基本信息、安

公网大模型进行记录并在平台中进行

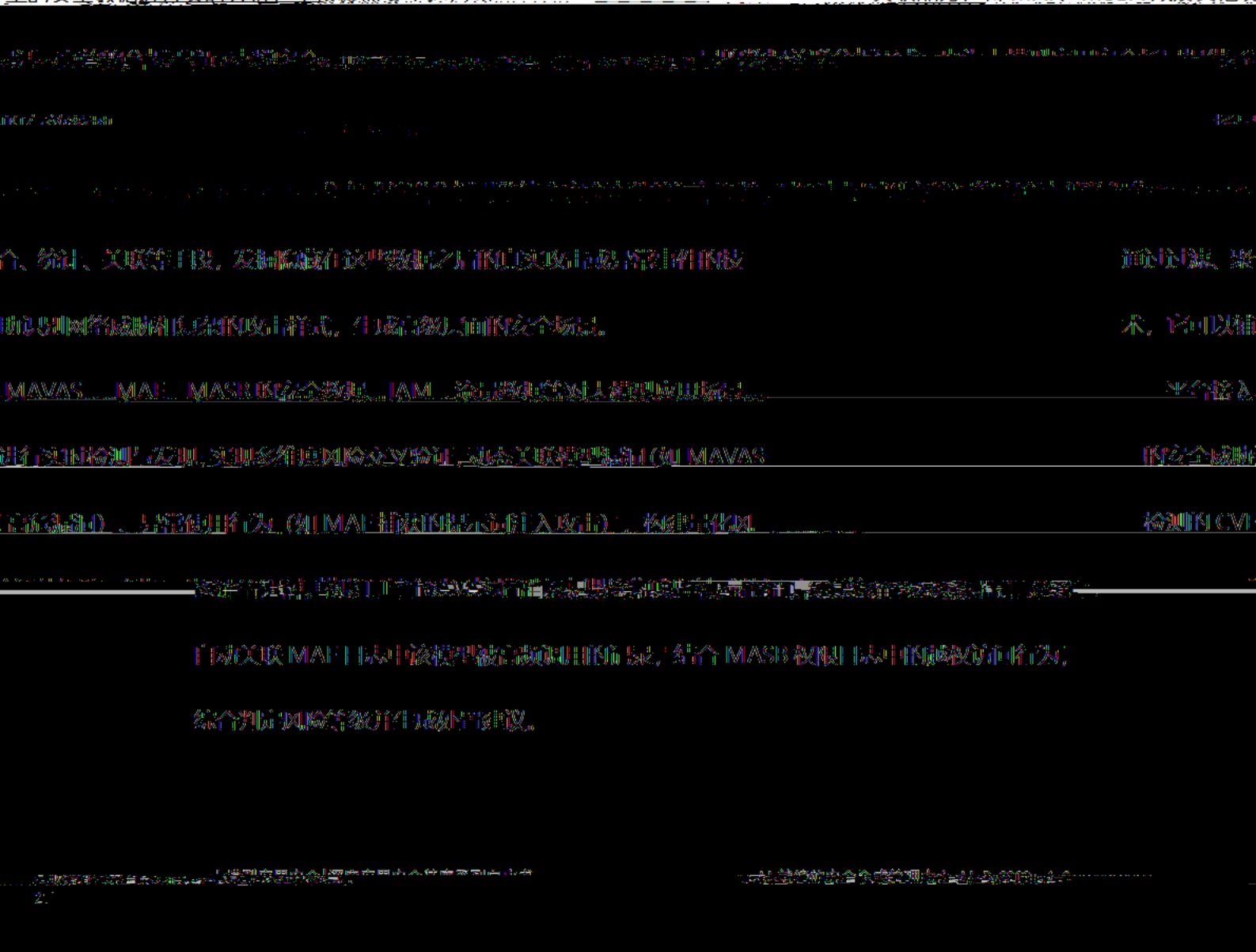
的风险信息，以便于对该类型大模型

通过监测企业员工的公网大模型访问，将所涉及的

大模型实体维护，包括大模型的基本信息、员工应用中

## 4.3 大模型安全分析

## 4.3.1 大模型风险关联分析



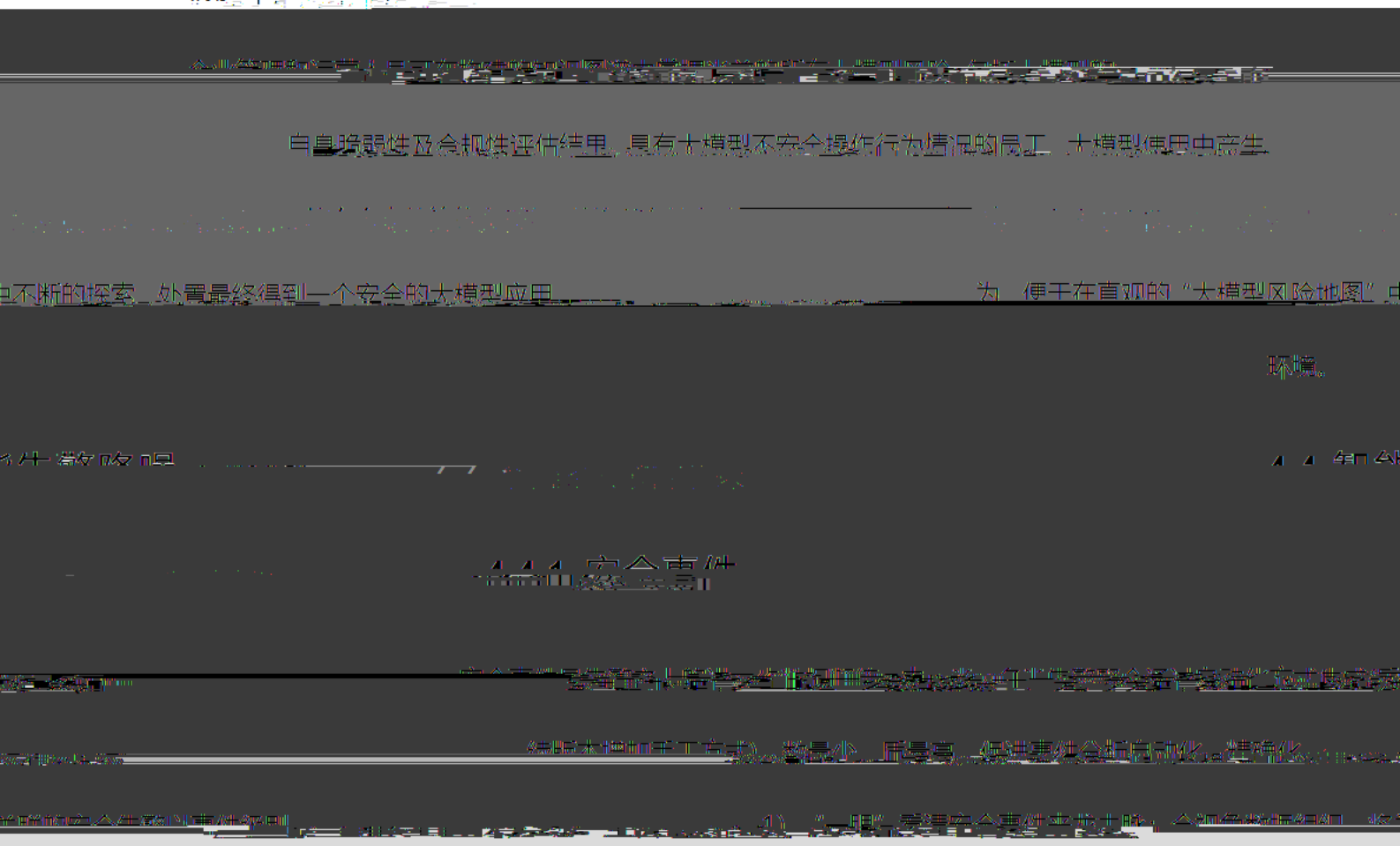
### 4.3.2 基于用户身份的行为分析

件。

体分析。通过关联 MAF 的输入输出审计日志、MASB 的访问控制记录及 MAVAS 的合规检测结果，系统可精准识别高风险行为模式，分析能力例如：

- **风险行为评估**：整合各大模型安全子系统的数据，对用户的大模型使用行为进行风险和合规性分析。
- **行为画像**：建立用户安全使用画像，识别高风险用户、高频风险行为、关键违

- 图谱实体涵盖：
- **风险实体：**用户（身份/权限）、大模型、业务资产、敏感数据信息元素、安全告警/事件、合规策略等。
- **风险关系：**使用关系、会话关联、输入关系、输出关系、事件归属、策略违规映射等



关键路径、ATT&CK 等方式进行呈现，一眼看出安全告警的重要信息，同时在安全事件的概览页面将 IP、资产等信息进行展示，清晰展示安全事件的影响范围。

2）“一图”展示安全事件攻击故事线：以拓扑形式展示安全事件的攻击关系，方便用户的安全事件进行研判。拓扑图以 IP/资产为节点，以关联的告警为边，并支持数据的下钻，

数据）。

3）安全事件处置与抑制：在安全事件的实体页签对关联的 IP、域名、账号、进程、文

安全事件的处置、抑制。

### 4.4.2 智能降噪

告警疲劳中解脱出来，投入到安全运营更关键的任务中。

全息降噪引擎将安全专家的知识与 AI 能力进行深度融合，覆盖常见的安全事件类型的




## 4.5 大模型风险监测管控



### 4.5.1 大模型自身风险监测

1) 上线前风险评估：[illegible] MAVAS [illegible]

[illegible]

[illegible]

、对抗攻击防护、鲁棒性测试等多个维度的全

针对这些安全隐患提供涵盖伦理对齐

进行集中展示，闭环管理。

[illegible]

环境中，基于 MAVAS 的大模型应用场景

2) **安全性动态评分**：可在大模型运行的实网

[illegible]

[illegible]

[illegible]

[illegible]

**内容合规实时检测**：动态维护国家到行业再到企业单位的大模型使用政策及规范 3) **内**

响应延迟、资源利用率，可与企业内多大模型的服务

断或负载均衡策略。

确环境风险（如部署服务器的系统漏洞或相关组件的

4) **服务能力监控**：监测大模型的响

接口联动进行服务自动触发熔断

5) **系统层风险预警**：扫描模型依赖

### 4.5.2 风险人员监测

户行为审计记录，实现大模型风险责任到人且问题事件可追溯，有效提升企业中应用大模型的风险可控。

2) **高危用户画像**：基于人员行为风险数据对企业中应用大模型的所有人员构建用户风

员分布”、“风险行为分布”等，便于由整体到个人掌握人员风险情况。

### 4.5.3 风险行为/事件监测

1) **多维度告警聚合**：将 MAF、MASA 所监测的大模型攻击行为、风险使用行为、恶意指令、越权访问记录等的告警事件，MAVAS 的合规性、风险性评估，AI-R-SOCC

表查询和详情呈现

2）**溯查溯源分析：**AI

入排查并减少大模型应用风险。

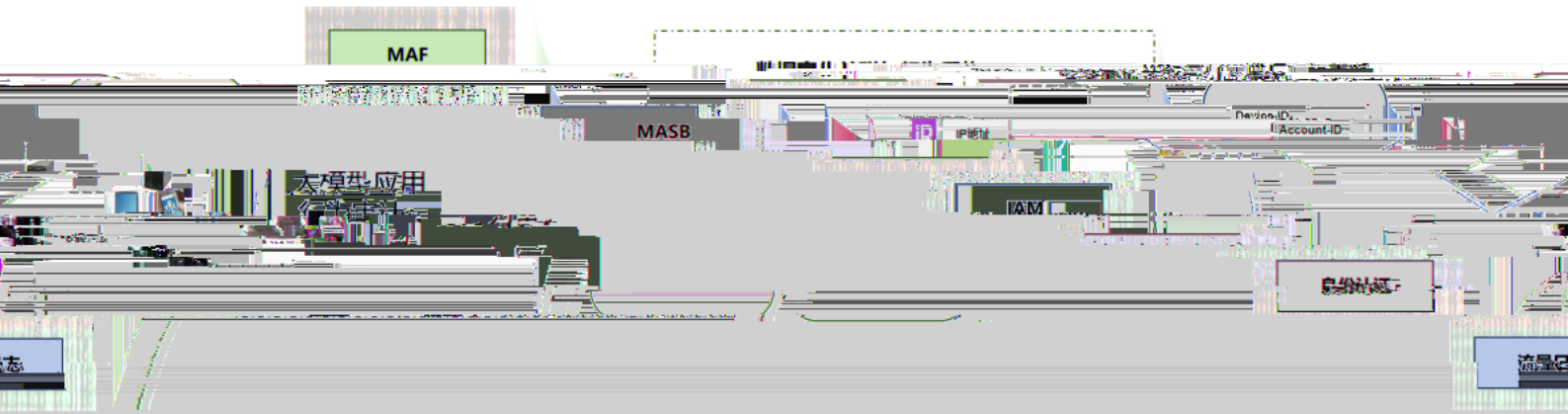
统一的大模型安全合规网关，都助力客户进入大模型安全合规化

外部协议，方

体”包括：

## 2.5 行为审计与溯源

用户、业务协同对大模型应用的访问行为进行全程记录，




## 型安全态势呈现

## 4.7 大模

多种安全风险评估、监测、管控的过程中将产生大量的监测数据，AI-R-SOCC ... 在大模型多

通过**多源数据融合引擎与 2D+3D 可视化技术的应用**，将来源的合规探测、行为监测、风

AI-R-SOCC 提供了大模型综合风险态势、大模型资产风险态势、大模型使用风险监测

户行为安全的直观

总体安全健康指数以及大模型自身风险、使用中的安全威胁、用

指标，便于全局掌控安全风险和安全态势。

融合分析，便于快速识别高危人员并阻断违规行为。

- **大模型智能运营态势：**智能运营态势大屏集成了总览信息、智能运营任务完



# 5 部署和典型应用场景

安全“新三件套” AI-R-SOCC 作为大模型应用安全方案的核心基座，对接大模型应用

型安全评估系统），（MAF 大模型应用防火墙、MASB 大模型访问安全代理、MAVAS 大模

部署等全生命周期的复杂安全需求，提供针对性和高效 全运营，针对大模型在训练、推理、

○ 解决方案

1. 敏感数据泄露防控

2. 生成内容滥用阻断

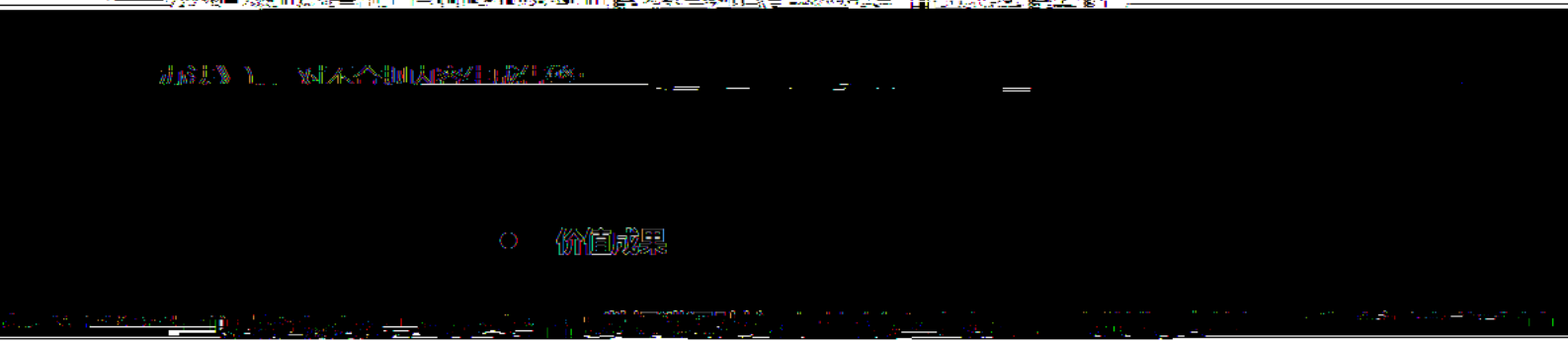
**大模型投毒攻击防御** 3.

模型对话敏感数据防泄漏：通过对输出数据的全面解析，识别大模型对话交互

程中的敏感信息输出实时识别、封堵或脱敏，保障大模型服务合规。

✓ 大模

过程



大幅降低；

低；

效的品牌声誉损害与用户流失。

3. **减少业务损失：**避免因模型滥用导致

## 5.2 场景一：模型上线准入

- **痛点**

泄露漏洞）；

“带病上线”风险（如训练数据污染、隐私

- **解决方案**

1. **自动化安全审查：**

管理暂行办法》）自动校验模型输出模板是否符合监管要求。

2. 动态准入决策：

○ 价值点

○ 痛点

能力退化风险：模型API响应延迟飙升（P95>1000ms）、资源过载（GPU利 ✓ 服务

用率>95%），引发业务中断，且缺乏实时预警与自愈机制。

扫描：联动 MAVAS 每小时扫描模型依赖环境（如 CUDA 版本、容器镜像），✓ 漏洞

高危漏洞（如 CVE-2023-1234）

✓ 自动化补丁：对低风险漏洞自动推送修复建议（如升级 TensorFlow 至 2.12）

漏洞触发模型隔离并告警。 高风险

**动态适配：** **2. 合规性**

✓

✓ 输出内容校验：通过 NLP 引擎比对模型输出与合规知识库，违规内容实时脱敏或

阻断。 阻

**务能力保障：** **3. 服**

性能基线监控：设定 API 响应延迟（<500ms）、错误率（<1%）、资源利用率 ✓ 性

至轻量化模型版本。

- **价值成果**

合规率提升至 100%。 2. 政务问答模型输出内容

## 5.4 场景四：影子大模型监测与治理

- **痛点**

✓ 企业大模型资产清单分散在多个部门，存在未登记的“影子模型”；

✓ 员工私搭模型（如 Llama 2 代码生成工具）成为数据泄露与攻击跳板；

- **解决方案**

行风险探测和输入/输出监测，对风险进行提示并在必

建私搭模型指纹库。

✓ 通过对发现的影子大模型进行

要时实时阻断。

- **价值成果**

- **大模型资产黑洞消除：**识别并

并治理若干个未登记的私搭模型，收缩企业内大模型攻

击面风险。

# 6 能力优势

## 6.2 集中调

## 6.3 快速响应

专家

## 6.4 安全

的大模型知识图谱与数据分析能力，提供针对大模型安全的常用安全知识智能

依托强大

