

版权声明

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。

式、插图、照片、方法、过程等内容，除另有特别注明外，其著作权或其他相关权利均属于北京启明星辰信息安全技术有限公司。未经北京启明星辰信息安全技术有限公司书面许可，不得复制、摘录、备份、修改、传播、翻译成其它语言、将其全部或部分用于商业用途。

本文档依据现有信息制作，其内容如有更改，恕不另行通知。

免责声明

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确可靠，但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损害承担责任。

信息反馈

如有任何宝贵意见，请反馈：

信箱：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

您可以访问启明星辰网站：www.venustech.com.cn 获得最新技术和产品信息。

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 缺少大模型资产使用的合规性持续监测评估 10

3 定义和建议路 13

3.1 AI-R-SOCC定义 13

3.2 建议路 14

4. 16

4.1. 16

4.1.1 16

4.2 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

内部私搭大模型.....20 4.2.3

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 基于用户画像的行为分析.....22

型安全图谱分析.....22 4.3.3 大模

.....23 4.4 智能告警降噪

事件.....23 4.4.1 安全

4.4.2 智能降噪.....24

4.5.1 大模型自身风险监测.....25

4.5.2 风险人员监测.....26

4.5.3 风险行为/事件监测.....27

4.5.4 智能治理建议生成.....27

.....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]
6.1 智能运营.....36
[illegible]
37
6.4 安全专家.....
37
6.5 全天候值守.....

1 公司简介

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、拥有

2010年6月23日，[illegible]

[illegible]

[illegible]

的多项空白。

大型活动提供全方位信息安全保障。

小学。

启明星辰公司将秉承诚信和创新精神，继续致力于提供具有国际竞争力的自主创新的安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。

背景与挑战 2

1 背景分析 2.

随着生成式大模型（LLM）技术的发展与行业应用，以及国产大模型现象级产品

等辅助等技术环节，数据分析、代码生成、

然而，大模型在企业内的多形态部署（如私有化部署、员工私搭、外部 API 调用）导致数据主权模糊、合规风险剧增、攻击面扩大等问题。据 Gartner 预测，至 2025 年，30% 的企业将因大模型滥用导致重大数据泄露事件。大模型在企业中大量应用过程中可能产生的

训练和应用中，数据含大量敏感信息，一旦泄露，会侵犯个人隐私、

行为和后果，引发法律风险。

攻击者可训练数据注入恶意样本，干扰后续训练，使模型性能下降、

。

通过输入数据微小扰动，让模型产生错误输出，在图像识别领域会

识。

像和视频，误导公众、影响社会稳定；

数据，若处理不当，易导致隐私泄露。

- 隐私泄露：大模型训练依赖大量用户数

- **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、风险可管、处置可溯。**

2.2 面临挑战

便捷的同时也带来了新的安全边界问题，例如大模型输入/输 大模型的应用给企业带来

的注入攻击等，应对这些新问题企业会新增新型的大模型安全 出的敏感信息泄漏、对大模型

防护手段，但这些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

1. 数据孤岛效应：

MASB无法独立识别跨系统攻击链。 覆盖的私有API通道注入恶意指令时，

2. 分析能力局限性：

隐蔽指令链）；

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体训练数据

违规访问行为。

有效性无法保障：

团队日均处理告警量超千条，有效告警识别率不足20%。

分散的子系统监控难以捕捉此类长期潜伏威胁。

4. 性能与安全的平衡困境：

- **监控影响业务：**MAVAS深度扫描导致模型推理延迟增加30%，企业被迫在安全

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

合规性持续监测评估

管不足：

2.2.2 缺少大模型资产使用的

1. 大模型合规评估监

时同步至所有子系统。

更新要求，企业难以及

2. 资产不可见性：

私自调用ChatGPT、Claude等公共模型处理客户数据（如

调查，67%的企业无法准确识别内部大模型

- 供应链风险隐患：外部模型（如阿里云、DeepSeek）的数据存储地、训练集

监管要求，可能违反《数据安全法》关于数据

3. 权责界定缺失：

）2/001 标准的完整审计报告。

符合 IS

1. 响应机制分级 协同：

- 单点响应的局限：人

以此作为风险行为给个别人员进行决策执行，如果没有发生 则不予处罚进行

- 策略冲突风险：各个系统对外可能采取不认为个别的策略行为；环节

作的合规。

2. 闭环治理缺失：

[illegible]

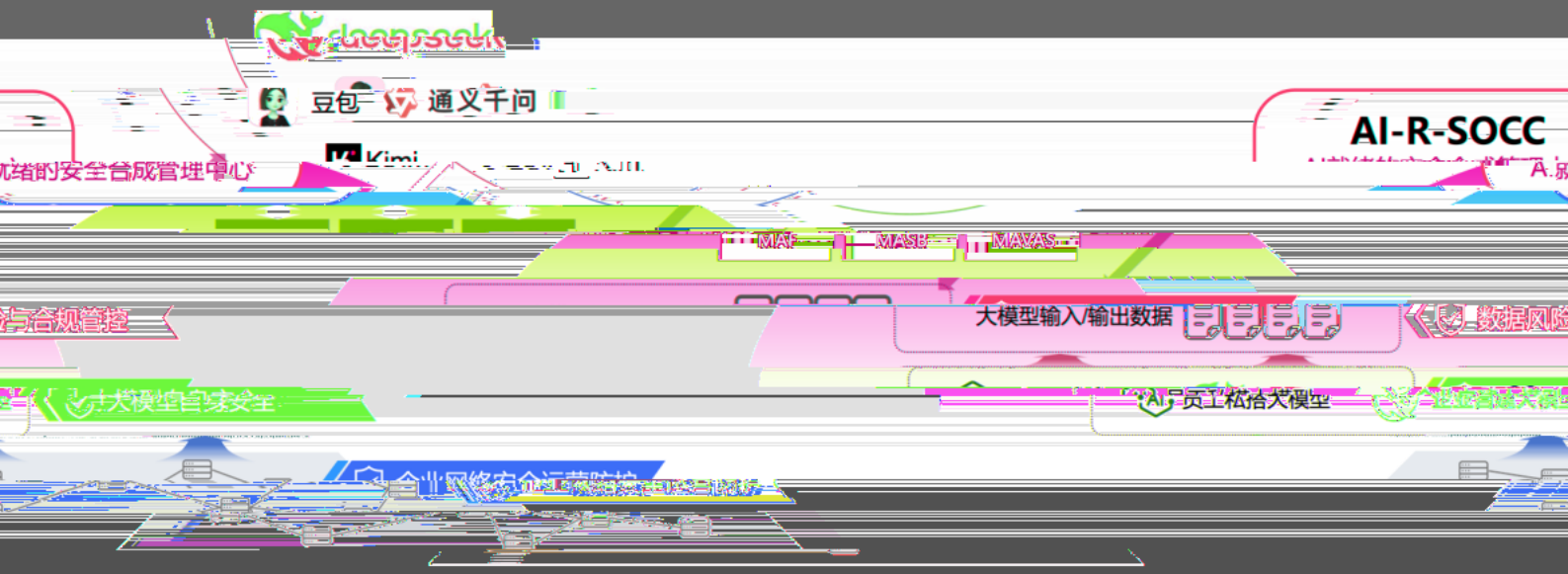
- **知识沉淀不足：**处置经验沉淀在人员头脑中，缺乏标准化剧本库供复用。

3 定义和建设思路

3.1 AI-R-SOCC 定义

管理基座融合大模型应用安全、数据安 面对大模型安全挑战，亟需三位一体的 AI 安全管

主。 全融合管理中心（AI-R-SOCC）应运而



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

C 依托 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-SOC

MAF、MASR、MAVAS 等大模型安全的管控手段，可形成对大

应用 将会被 AI-R-SOCC 放入企业网络安全整体视角中进行统一管理，形成大模型

3.2 建设思路

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 是企业级大模型应用

实时监测和阻断大 型、在保障大模型使用安全的场景中可以统一纳管合规审核、身份管理、

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

同期的智能化管控体系，例如通过MAVAS、MAF实时上线前以及投入应用的大模型进行

内容合规性的监测评估，通过 MAF+MASB+MAVAS 对 续的大模型自身安全以及输入输出内

风险性、合规性进行全面分析，发现威胁行为以及可能造 企业内部人员的大模型使用行为的风

成的企业损失

的核心定位包括：

- 作为企业大模型安全体系的“指挥舱”，聚合分散的安全能力，打 **上层管理平台**：作

台。 通常是安全与管

大模型上线前的安全合规审查、运行期的实时监测与阻断，

形成闭环管理。

险评估）、输出内容合法合规性（如内

AI 就绪的安全合成管理中心 AI-R-SOCC v1.0



提供多源异构的海量数据进行接入和治理，可接入融合包括大模型应用安全“新三

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

件套”

化分析引擎来支撑上层能力构建，包括处置海量多源异构信息的数据中台能力、提供安星智

能体辅助智研及智能决策的泰合安全大模型能力、以及多种关联分析引擎、行为分析引擎

析引擎。

提供了面向大模型智能体全生命周期的安全管控能力以及 AI 驱动

大模型智能体安全管控能力中集成了 MAF、MASB 和 MAVAS 的

安全评估能力、基于泰合分析的大模型智能体风险监测及安全事

联动响应能力。智能化安全运营能力中集成了 AI 驱动的智

在此基础上 AI-R-SOCC

的智能化安全运营能力。面向

大模型智能体安全管控及全生命周期的

件管理能力以及风险行为的

4 核心能力

4.1 产品智能化安全运营

ALP-2000产品智能化升级，分析下发的自然语言处理技术，标准了一个集中

统供应处置，安全指令生成，剧本执行了一体的交互平台，通过自然语言交互，实现智

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

事件处理效率，

恒提升安全

产品智能化支持7x24小时不间断响应，具备实时识别和应对网络威胁能力，可自动检测

和执行安全指令，结合历史事件和知识库，智能推荐处置剧本，帮助安全团队在

复杂场景中快速制定精准应对策略。

用户可以通过群聊@安易智能体唤醒其自动回复功能，对于简单任务，系统内置的小模

型能够直接响应，无需调用大模型，从而显著降低

智能体界面，利用自然语言对话或鼠标操作完成动

反馈。用户还可以通过点击悬浮图标进入智

作调用、剧本执行和文件上传等任务，实现对安全场景的全方位支持与快速响应。

4.1.1 能力市场

能力市场为安全人员提供了一个集中的平台，将各种安全的功能模块和自动化

为Python、Java等语言，帮助用户SDK开发并集成到，将安全知识沉淀转化为可执行

的

杂场景下的自动化响应流程。

设备的连通性、性能 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

题并采取防范措施，提升整体安全运营的效率。 状态及潜在安全隐患，帮助用户提前发现问题

4.1.2 任务管理

三。 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

- 人工任务管理

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

- 周期任务管理

行周期，并配置具体的执行 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

动化。 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

- 脆弱性任务管理

专注于系统脆弱性管理，利用大模型的调度能力驱动漏洞扫描设备完成包括漏洞扫描

安全。 户快速识别和修复系统中的安全漏洞，全面保障资产

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这一核心任务，依托智能监测、智能值守和研判任务三大功能，

[illegible]

- 智能监测

[illegible]

得到及时发现与处理。

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

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

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，

- 研判任务

像分析，全面评估告警的威胁级别与潜在风险，帮助用户

聚焦关键威胁，减少不必要的干扰，显著提升事件响应效

警有效性判断、全路径分析以及画

快速识别和优先处理高风险事件，

率。

- 大模型风险处置

预置联动剧本，覆盖大模型应用安全、恶意使用、违规输出等场景，例如：

- **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

- **数据泄露阻断**：MASB 识别敏感文件上传 → MAF 实时

敏输出内容 → MAVAS

数据治理部门。

扫描关联模型训练集残留风险 → 隔离高风险模型并告警

户）。

误封合法用户

4.2 全局资产

管理

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

平台提供全面的

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

自研人员进行审计。

模型

4.2.2 企业/单位自建大模

批准人员在该合作平台管理，但对此类人员进行合作合规监管。

映射企业员工人，支持如下监控：（如版本升级后安全审计队列）。

安全信息包括：大模型系统安全性、合规性检测指标，针对中重风险行为及安全性

控制在平台。

监测。

生成工具）实现 针对员工未经审批私自部署的模型（如研发人员搭建的Llama 2代码

大模型涉及的基本信息、安全信息、运行性能监控信息的维护管理如上。

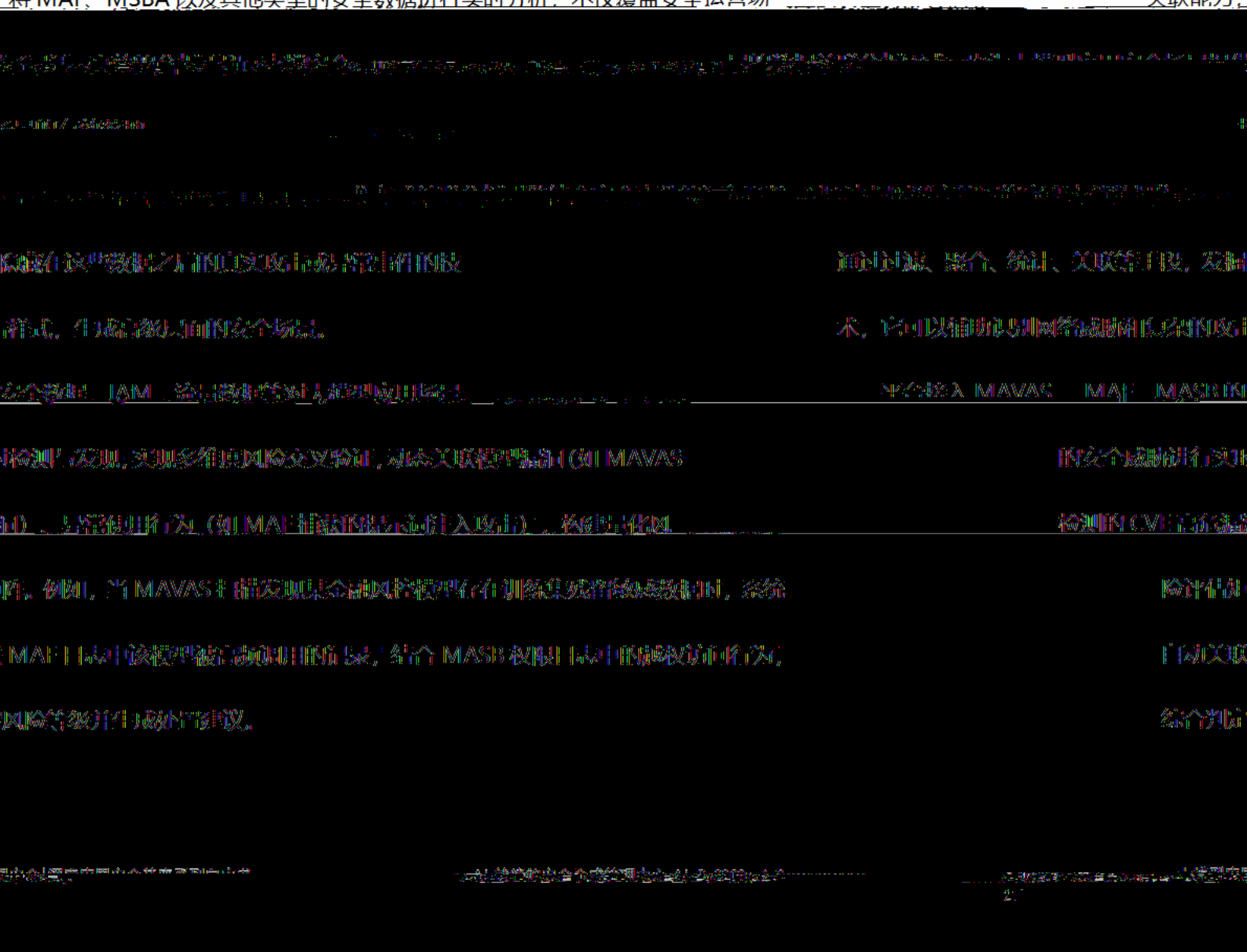
通过监测企业员工的公网大模型访问，将所涉及的公网大模型进行记录并在平台中进行

型实体维护，包括大模型的基本信息、员工应用中的风险信息，以便于对该类型大模型

4.3 大模型安全分析

关联分析

4.3.1 大模型风险



4.3.2 基于用户身份的行为分析

件。

的行为模型和异常行为检测。通过对网络中的行为深度解析和识别，将行为数据

立一 部门归属、权限等级），AI-R-SOCC 构建动态用户画像，实现“身份-行为-数据”三

制记录及 MAVAS 的合规检

体分析。通过关联 MAF 的输入输出审计日志、MASB 的访问控

测结果，系统可精准识别高风险行为模式，分析能力例如：

户的大模型使用行为进行风

- **风险行为评估**：整合各大模型安全子系统的数据，对用

险和合规性分析

。

用户安全使用画像，识别高风险用户、高频风险行为、关键违规场

- **行为画像**：建立

4.3.3 大模型安全图谱

AI-R-SOCC 基于知识图谱技术，将大模型安全生命周期监测管控过程中的所有主体一客

体、行为关系一事件，以

使用和自身风险全景视图，并通过关联层开放下钻的方式便于用户从总

体到局部的掌控大模型安全风险状况。

- 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，

信息、告警、日常行为、以及进程信息（需要接入安服易云EDR日志）

数据）。

在安全事件的实体页签对关联的 IP、域名、账号、进程、文

3）安全事件处置与抑制：

抑制。

安全事件的处置

4.4.2 智能降噪

的任务中。

告警疲劳中解脱出来，投入到安全运营更关键

力进行深度融合，覆盖常见的安全事件类型的

全息降噪引擎将安全专家的知识与 AI 能力



4.5 大模型风险监测管控

体-使用行为-业务环境”的

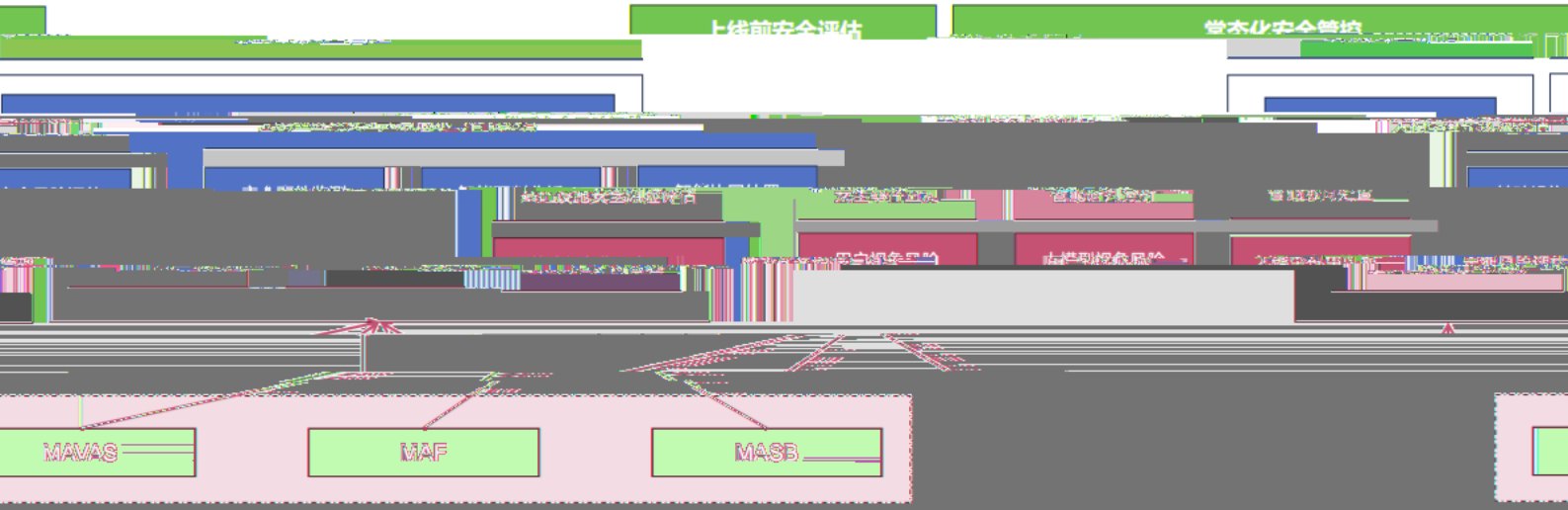
AI-R-SOCC 通过**全景式风险感知技术**，构建覆盖“模型本

后续的联动处置形成大模型治理优

多维度监测体系，实现从风险发现的全局可见，并可指导

化、处置精准化、治理智能化。

化的全链路闭环，助力企业实现大模型应用的**风险可视化**



1 大模型自身风险监测

上线前风险评估：平台与 MAVAS 进行深度能力协同，在大模型上线前提供相应的

针对这些安全隐患提供涵盖伦理对齐、对抗攻击防护、鲁棒性测试等多个维度的全方位安全评估，风险评估结果在平台进行集中展示，闭环管理。

2） **安全性动态评分**：

4) **服务能力监控**：监测大模型的响应延迟、资源利用率，可与企业内多大模型的服务接口联动进行服务自动触发熔断或负载均衡策略。

5) **系统层风险预警**：扫描模型依赖环境风险（如部署服务器的系统漏洞或相关组件的

4.5.2 风险人员监测

件可追溯，有效提升企业中应

户行为审计记录，实现大模型风险责任到人且问题事件

用大模型的风险可控。

大模型的所有人员构建用户风

2) **高危用户画像**：基于人员行为风险数据对企业中应用大

员分布”、“风险行为分布”等，便于内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）**调查溯源分析：** AI-B-SOC 记录有所有的大模型告警事件信息以及告警背后相关

为分析甚至流量数据，可根据安全运营的诉求对告警事件进行深入 的原始日志、行

入排查并减少大模型的应用风险。

4.5.4 智能治理建议生成

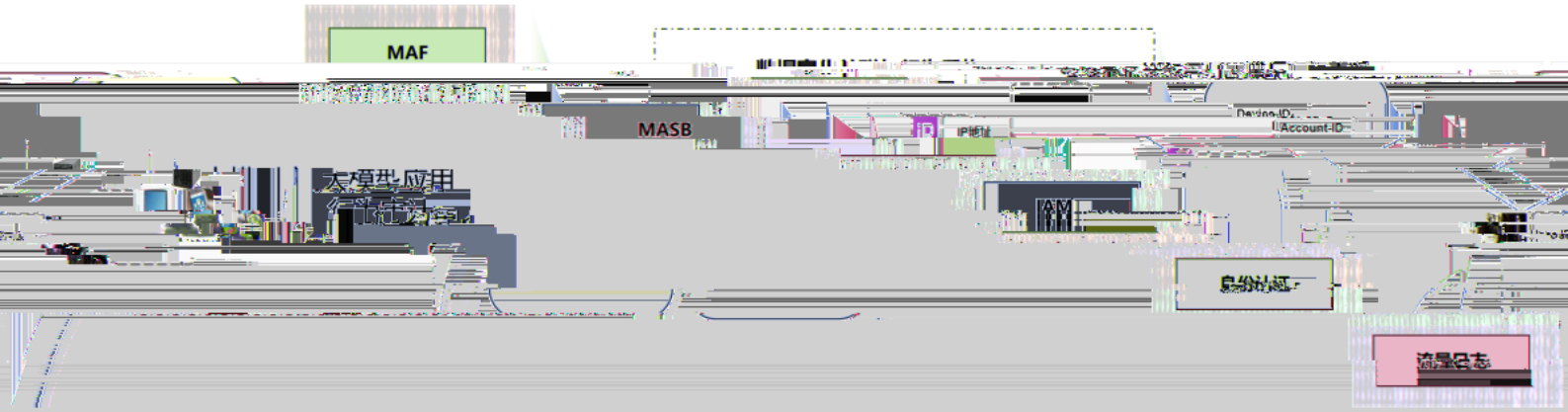
，帮助用户推进大模型的安全合规化 处理建议，并综合企业的大模型应用环境给出治理建议

使用。

4.5.5 大模型风险管控处置

4.5.6 行为审计

用户、业务协同对大模型应用的访问行为进行全程记录，



全态势呈现

4.7 大模型安全

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

在大模型多种安全

多源数据融合引擎与 2D+3D 可视化技术的应用，将来海量的合规检测、行为监测、风

通过

险。

-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

AI-R

大模型综合风险态势：

安全的高纬

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、用户行为

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、智能运营任务完成



5 部署和典型应用场景

三件套”

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

系统），

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

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

4. 输出内容合规风险：模型输出违反行业监管要求（如不符合《生成式人工智能服务

则（如歧视性言论），引发法律处罚与品牌危机等。

管理暂行办法》）、伦理准

- **解决方案**

1. 敏感数据泄漏防控

[illegible]

动态拦截高风险内容；

✓ 实时监测输出内容，识别敏感信息泄露（如客户住址、商业秘密）、违法违规内

及时抑制风险的产生。

2. 生成内容滥用阻断

[illegible] 识别隐蔽攻击指令（如“生成绕过审

✓ 语义深度解析：基于多轮对话上下文理解

）；

核的贷款申请话术”

实时拦截违规内容；

求合规策略（如禁止生成“保本保收益”承诺），实

[illegible]

3. 大模型投毒攻击防御

✓ 输入数据验毒：通过 MAE 检测输入的异常数据（如噪声注入），联动 MASR 阻

断污染源 IP；

✓ 大模型训练或投喂数据批量认证：对模型训练集进行 MAVAS 安全扫描，拦截未

认证数据源。

数据的全面解析，识别大模型对话交互

脱敏，保障大模型服务合规。

✓ 大模型对话敏感数据防泄漏：通过对输出

过程中的敏感信息输出实时识别、封堵或

办法》），对不合规内容生成告警；

- **价值成果**

大幅降低；

低；

业务损失：避免因模型滥用导致的品牌声誉损害与用户流失。

3. 减少业

一、模型上线准入管控

5.2 场景

- **痛点**

“带病上线”风险（如训练数据污染、隐私泄露漏洞）；

- **解决方案**

1. **自动化安全审查：**

充分实战验证。

/API 敏感数据泄漏、大模型应用层漏洞攻击等方面进行

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

✓

个人汇总：

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

用率>95%），引发业务中断，且缺乏实时预警与自愈机制。

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

别高危漏洞（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）

风险漏洞触发模型隔离并告警。 高风

性动态适配： 2. 合规

络同步引擎：实时对接监管机构数据库（如网信办的合规库），自动解析最新 ✓ 策略

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

阻断。

： 3. 服务能力保障

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

至轻量化模型版本。

- **价值成果**

容合规率提升至 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**
  - ✓ 自动发现企业内部模型实例（包括容器化部署的私搭模型），构建动态资产库；
  - ✓ 通过流量探针（识别非标准 API 特征）与代码扫描，建私搭模型指纹库。
  - ✓ 通过对发现的影子大模型进行风险探测和输入/输出监测，对风险进行提示并在必要时实时阻断。
- **价值成果**
  - **大模型资产黑洞消除**：识别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

的环境下持续运行。

6.2 集中调度

充分考量大模型安全涉及的多维度安全能力，将各类针对大模型数

法保护、运行环境加固等安全平台能力统一管理调度。打破安全能力

安全运营效率，为大模型构建全方位的安全防护网。

安全运营

6.3 快速响应

6.4 安全专家

依托强大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

