

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

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

有限公司进行复制、摘录、修改、传播、翻译成其它语言，将其全部

或部分内容用于商业用途，其内容如有更改，恕不另行通知。

本文档依据现有信息制

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力求尽可能准确，但其中内容

仅供参考，北京启明星辰信息安全技术有

限公司对本文档中的编辑、不准确、或错误等致

的疏忽和遗漏不承担法律责任。

信息反馈

如有任何宝贵意见，请反馈：

北京市海淀区中关村软件园21号楼启明星辰大厦 邮编：100193

信箱：北京市海淀区东北旺西路

100193

电话：010-82779000

传真：010-82779000

www.yanustech.com.cn 获得最新技术和产品信息

你可以访问启明星辰网站：



# 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
    - 2.2.1 大模型资产面临缺乏统一持续评估..... 10
    - 2.2.2 缺少大模型资产使用的合规性持续监测评估..... 10

## 3 定义和建设思路..... 13

### 3.1 AI-R-SOCC定义..... 13

### 3.2 建设思路..... 14

## 4 核心能力..... 16

### 4.1 合规性..... 16

#### 4.1.1 能力场..... 16

### 大模型资产实体定义..... 19 4.2.1

### 企业/单位自建大模型..... 20 4.2.2

4.2.3	内部私搭大模型.....	20
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.....	23
4.4	智能告警降噪.....	24
4.4.1	安全事件.....	24
4.4.2	告警降噪.....	24
4.4.3	告警降噪.....	24
4.4.4	告警降噪.....	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
4.6	行为审计溯源.....	28
4.7	大模型安全态势呈现.....	28
5	部署和典型应用场景.....	30
5.1	场景一.....	30
5.2	场景二.....	30
5.3	场景三.....	30
5.4	场景四.....	30
5.5	场景五.....	30
5.6	场景六.....	30
5.7	场景七.....	30
5.8	场景八.....	30
5.9	场景九.....	30
5.10	场景十.....	30
5.11	场景十一.....	30
5.12	场景十二.....	30
5.13	场景十三.....	30
5.14	场景十四.....	30
5.15	场景十五.....	30
5.16	场景十六.....	30
5.17	场景十七.....	30
5.18	场景十八.....	30
5.19	场景十九.....	30
5.20	场景二十.....	30
5.21	场景二十一.....	30
5.22	场景二十二.....	30
5.23	场景二十三.....	30
5.24	场景二十四.....	30
5.25	场景二十五.....	30
5.26	场景二十六.....	30
5.27	场景二十七.....	30
5.28	场景二十八.....	30
5.29	场景二十九.....	30
5.30	场景三十.....	30
5.31	场景三十一.....	30
5.32	场景三十二.....	30
5.33	场景三十三.....	30
5.34	场景三十四.....	30
5.35	场景三十五.....	30
5.36	场景三十六.....	30
5.37	场景三十七.....	30
5.38	场景三十八.....	30
5.39	场景三十九.....	30
5.40	场景四十.....	30
5.41	场景四十一.....	30
5.42	场景四十二.....	30
5.43	场景四十三.....	30
5.44	场景四十四.....	30
5.45	场景四十五.....	30
5.46	场景四十六.....	30
5.47	场景四十七.....	30
5.48	场景四十八.....	30
5.49	场景四十九.....	30
5.50	场景五十.....	30
5.51	场景五十一.....	30
5.52	场景五十二.....	30
5.53	场景五十三.....	30
5.54	场景五十四.....	30
5.55	场景五十五.....	30
5.56	场景五十六.....	30
5.57	场景五十七.....	30
5.58	场景五十八.....	30
5.59	场景五十九.....	30
5.60	场景六十.....	30
5.61	场景六十一.....	30
5.62	场景六十二.....	30
5.63	场景六十三.....	30
5.64	场景六十四.....	30
5.65	场景六十五.....	30
5.66	场景六十六.....	30
5.67	场景六十七.....	30
5.68	场景六十八.....	30
5.69	场景六十九.....	30
5.70	场景七十.....	30
5.71	场景七十一.....	30
5.72	场景七十二.....	30
5.73	场景七十三.....	30
5.74	场景七十四.....	30
5.75	场景七十五.....	30
5.76	场景七十六.....	30
5.77	场景七十七.....	30
5.78	场景七十八.....	30
5.79	场景七十九.....	30
5.80	场景八十.....	30
5.81	场景八十一.....	30
5.82	场景八十二.....	30
5.83	场景八十三.....	30
5.84	场景八十四.....	30
5.85	场景八十五.....	30
5.86	场景八十六.....	30
5.87	场景八十七.....	30
5.88	场景八十八.....	30
5.89	场景八十九.....	30
5.90	场景九十.....	30
5.91	场景九十一.....	30
5.92	场景九十二.....	30
5.93	场景九十三.....	30
5.94	场景九十四.....	30
5.95	场景九十五.....	30
5.96	场景九十六.....	30
5.97	场景九十七.....	30
5.98	场景九十八.....	30
5.99	场景九十九.....	30
5.100	场景一百.....	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 安全管理.....35

6.1 智能运营.....36

35

6.2 事件处置.....36

36

6.3 持续改进.....36

37

6.4 安全专家.....37

37

6.5 全天候值守.....37



的多项空白。

大模型网络安全提供全方位有效安全保障。

合产品推出专项服务，帮助客户全面梳理其IT基础设施的安全性和自产效能，为保护和提升中国企业的民族网络安全产业第一品牌而不懈努力。

# 挑战

# 分析

## 2 背景与挑战

### 2.1 背景分析

来源: Gartner, 2024. 数据仅供参考, 不作为任何投资建议。 | 网络安全态势感知与威胁情报 | 2024年10月

随着数字化转型的深入, 企业数据资产日益丰富, 数据泄露和滥用风险显著增加。

网络安全威胁日益复杂, 攻击手段不断升级, 企业面临严峻的安全挑战。

数据泄露 (如私有化部署、员工私搭、外部API调用) 导致数据资产流失。

数据滥用 (如过度采集、违规共享) 引发合规风险, 损害企业声誉。

数据篡改: 大模型在企业中大量应用, 数据篡改可能引发的安全风险不容忽视。

模型窃取: 攻击者通过输入恶意数据, 窃取模型参数或训练数据。

- **数据泄露:** 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、损害企业竞争力和声誉, 引发法律风险。
- **数据投毒:** 攻击者向训练数据注入恶意样本, 干扰正常训练, 使模型性能下降、准确性降低。
- **模型窃取:** 通过对输入数据微小扰动, 让模型产生错误输出, 在图像识别领域会导致分类错误。

模型输出不可控: 模型输出内容可能包含虚假信息、歧视性言论, 甚至引发社会不稳定。

误导公众、影响社会稳定: 模型输出虚假信息, 误导公众, 影响社会稳定。

大模型训练依赖大量用户数据, 若保护不当, 易导致隐私泄露。

数据篡改: 攻击者通过输入恶意数据, 篡改模型输出结果, 影响企业运营。

模型窃取: 攻击者通过输入恶意数据, 窃取模型参数或训练数据。

放在企业全局视角的大模型安全治理面临五大核心矛盾：

- **数据孤岛化**：各子系统独立运行，缺少对大模型全生命周期（输入-推理-输出）的

监测与检测能力。

个规划中；

- **数据孤岛化**：安全日志分散存储，无法实现跨系统攻击链溯源与

攻击溯源与检测。各子系统独立运行，缺少对大模型全生命周期（输入-推理-输出）的



监测与检测能力。

环，实现大模型应用的全局可视、风险可管、处置可溯、攻击可防。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、攻击可防。

处置可溯。

## 2.2 面临挑战

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

信息泄露、对大模型的注入攻击等。

大模型

信息泄露、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模型安全

大模型

防护手段，但这些手段基本针对某一单一的风险点，对于企业管理者面对众

出的敏感

多大模型的安全

防护手段，但这些手段基本针对某一单一的风险点，对于企业管理者面对众

1. 数据孤岛效应：

目的数据保护企

MASB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。

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。

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。

## 2. 分析能力局限性:

链);

的隐蔽指令

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

● 缺乏因果推

问行为。

源或违规访

无法保障:

3. 告警有效性

不下: 各子系统独立告警导致重复通知 (如MAE和MAVAS同时报告)

● 误报率居高

均处理告警量超千条, 有效告警识别率不足20%。

一模型的异常行为), SOC团队日

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20%, 企业被迫在安全性

与业务连续性间取舍;

不足40%, 运维成本飙升。

● 资源浪费严重: 多套子系统独立运行, 硬件资源利用率不

## 模型资产使用的合规性持续监测评估

## 2.2.2 缺少大模

性评估, 亟需安全合规管理的结构化治理。

企业在多子模型部署

### 1. 大模型合规评估监管不足:

一 缺乏结构化治理: 企业部署模型或可训练, 但缺乏安全评估

的合规性评估, 导致“带病运行”风险, 加剧数据泄露、模型滥用等

数据安全) 特征

2. 数据合规滞后: 监管部门对大模型输出的新兴风险(如深度伪造)

要求, 企业难以及时同步至所有子系统。

更新要

可见性:

2. 资产不

AI" 泛滥: 员工私自调用 ChatGPT、Claude 等公共模型处理客户数据(如

"显

业无法完整识别内部大模型资产。

调查, 67%的企业

训练集

● 供应链风险隐患: 外部模型(如通义千问、DeepSeek) 的数据存储策略、

反《数据安全法》关于数据出境的要求

来源缺乏透明性, 可能违

3. 权责管理缺失:

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

含客户隐私数据的业务模型)

研发部门可访问通田模型, 但禁止调田

— 审计链条断裂: 模型使用记录分散在 MAF、MAGB 等策略库中, 无法快速生成

符合 ISO 27001 标准的完整审计报告。

— 模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

1. 响应机制缺乏统一协同:

● 单点响应的局限: 大模型的响应外置往往需要综合输出/输出数据风险、合规标准

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

模型训练数据合规: 需落实数据安全法第 34 条(个人信息处理)

○ 策略冲突风险: 各子系统独立外置可能, 导致各环节认为合规的策略存在另一环节

作的管控机制。

## 2. 闭环治理缺失：

“漏洞扫描工具在扫描设备时发现的漏洞，其修复的及时性直接影响设备的安全运行。漏洞扫描工具在扫描设备时发现的漏洞，其修复的及时性直接影响设备的安全运行。漏洞扫描工具在扫描设备时发现的漏洞，其修复的及时性直接影响设备的安全运行。”

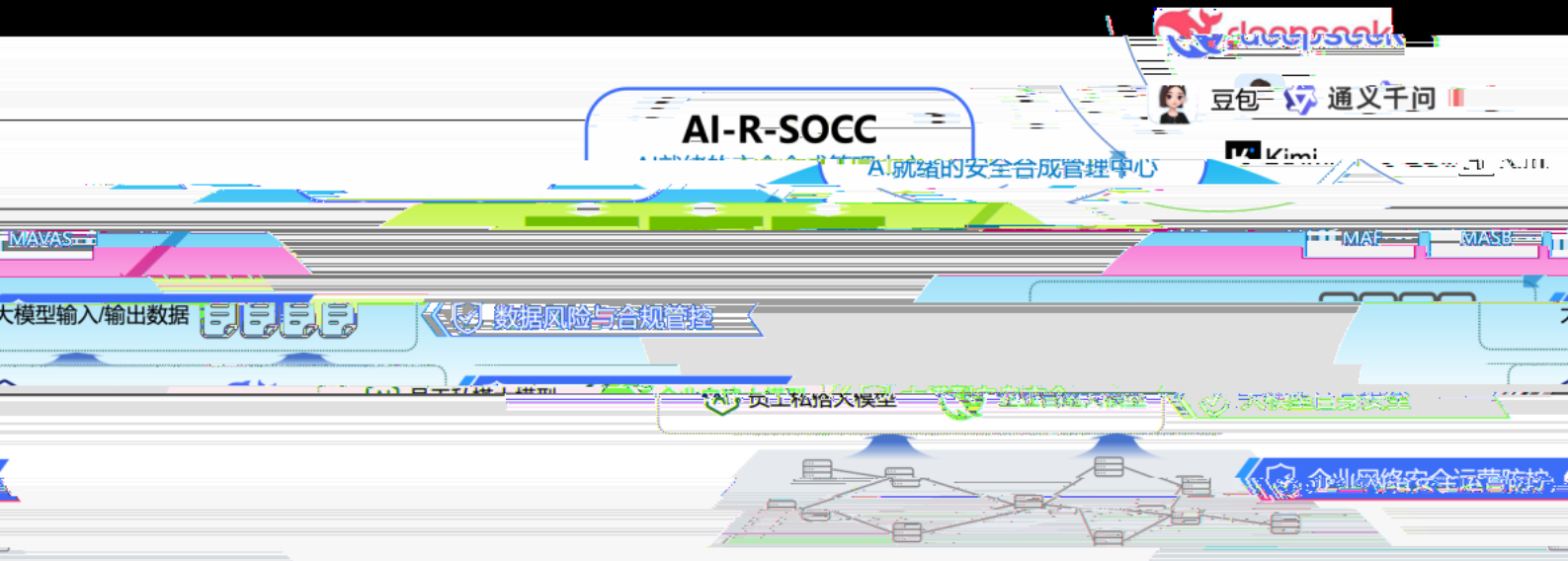
- 知识沉淀不足：大量经验沉淀在人员头脑中，缺乏标准化知识库供复用。

### 3 定义和建设思路

#### 3.1 AI-R-SOCC 定义

面对大模型安全挑战，亟需新一代的AI安全管理与融合大模型应用安全、数据安全

全融合管理中心 (AI-R-SOCC) 应运而生。



的趋势下,大模型也成为了企业中的信息交互中心,

在企业员工大量使用大模型辅助办公

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

这势必带来了新的安全边

一种最直接的消除隐患,在数据中识别风险,提升企业安全运营效率

MAE、MASB、MAVAS 等大模型安全的管控手段,可形成对大模型应用自身的安全性评估

MA

模型,并且这些过

什么?且自然评估大模型结果中是否存在输出数据的灵敏性和合规性的

放之于企业网络安全的具体视角进行统一运营,形成大模型应用

将会被 AI-R-SOCC

网络安全的一体化安全运营

融合、协同安全

### 3.2 建设思路

化安全治理以及智能化安全运营领域的中坚力量

AI-R-SOCC 是企业级大模型应用的集中

处置-审计”全生命

模型安全子系统（如 MAF、MASB、MAVAS）构建覆盖“准入-运行-退

用的大模型进行技

术的智能化管控体系，例如通过 MAVAS、MAF 实现对大模型输入及输出

及输入输出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

滥用行为的风险性，会周期性进行全面分析，发现威胁行为以及可能造

企业内部人员的大模型滥

成的企业损失

的核心定位包括：

管理平台：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

理平台与策略融合。

策略数据

策略管理：从模型上线前的安全合规审查到运行期间的策略调整与优化。

策略生命周期

管理：

形成闭环

多给应用做AS的，运营策略制定与执行，策略生命周期管理

从保护问题）三大核心维度。

输入内容合法合规性（如内容风险检测）、输出内容合法合规性（如内容风险检测）

运营策略制定与执行，策略生命周期管理，策略生命周期管理

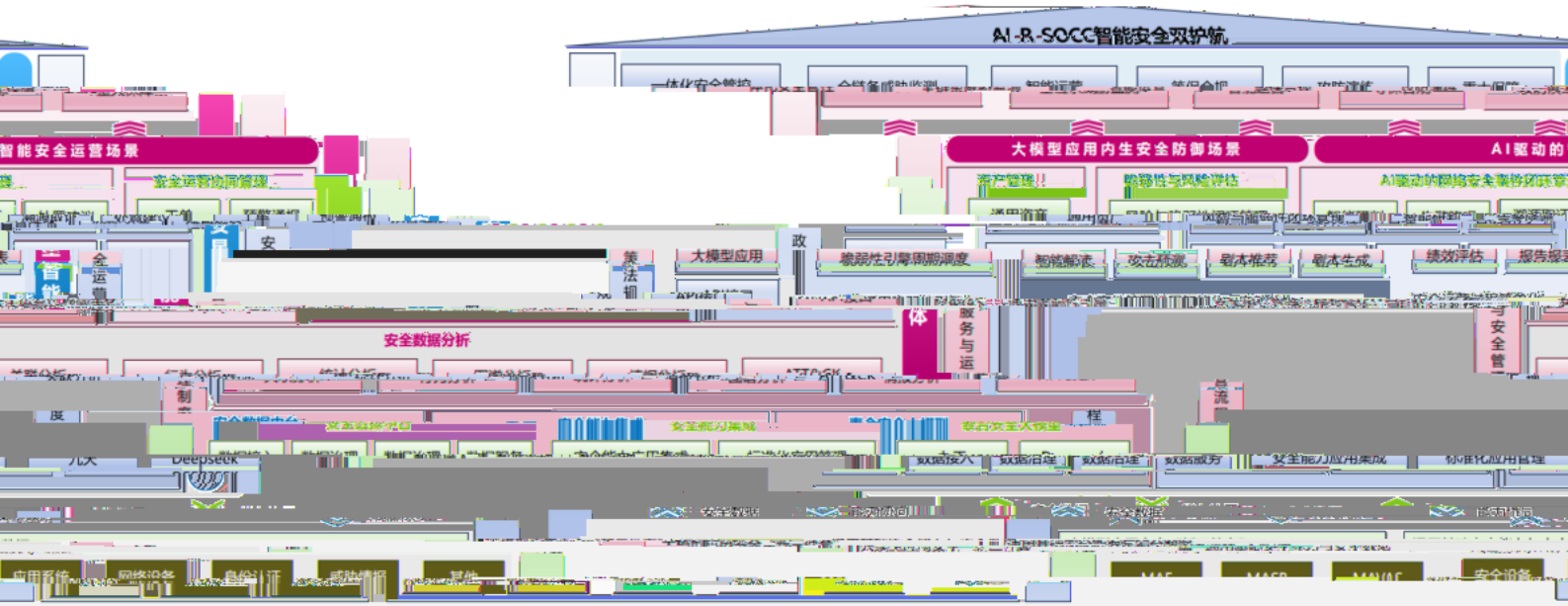
策略生命周期

运营策略制定与执行，策略生命周期管理，策略生命周期管理

策略生命周期

运营策略制定与执行，策略生命周期管理，策略生命周期管理

策略生命周期



台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

平

件套”（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

能辅助辅助以及智能决策的融合安全建模能力，以及多种关联分析引擎、行为分析引擎

图谱分析引擎。

提供了面向大模型应用生命周期的安全管理能力以及 AI 驱动

在此基础上 AI-R-SOCC

面向大模型的安全管理能力深度融合 MAF、MASB 和 MAVAS 的

的智能化安全运营能力。面

的综合评估能力、基于综合分析的大模型应用风险监测及安全行

大模型应用安全性及合规性

策略联动响应能力。智能化安全运营能力深度融合 AI 驱动的情

件管理能力以及风险行为

## 4 核心能力

### 4.1 安全智能体

安全智能体是安全运营中心（SOC）的核心引擎，负责接收来自各种安全设备的告警信息，并进行智能分析和关联分析，以识别潜在的安全威胁。

安全智能体通过统一的界面，提供统一的告警展示、事件关联、溯源分析、研判处置等功能，帮助安全人员快速响应和处理安全事件。

安全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

大提升安全事件处理效率。

图 4-1

安全智能体通过统一的界面，提供统一的告警展示、事件关联、溯源分析、研判处置等功能，帮助安全人员快速响应和处理安全事件。

安全智能体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

短时间内快速制定精准应急响应策略。

安全智能体还可以通过群聊@安全智能体，实现智能体自动回复功能，对工单单位

进行快速响应和处理，提升安全运营效率。

安全智能体利用自然语言处理技术，实现智能体自动回复功能，提升安全运营效率。

安全智能体通过统一的界面，提供统一的告警展示、事件关联、溯源分析、研判处置等功能，帮助安全人员快速响应和处理安全事件。

#### 4.1.1 能力市场

能力市场为安全运营中心提供了一个整合自身能力的平台，将各种安全能力封装成能力

包，通过能力市场进行发布和订阅，实现能力的快速集成和调用。

通过 Python、Java 等语言，借助鲲鹏 SDK 开发并集成应用，将安全基础设备能力封装成

能力包，通过能力市场进行发布和订阅，实现能力的快速集成和调用。

图 4-2

### 复杂场景下的自动化响应流程

提升设备资产的准确性、性能

同时，系统提供对网络设备上配置文件的实时监测，实时

发现问题并采取防范措施，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

## 4.1.2 任务管理

系统涵盖人工任务、周期任务和脆弱性任务三

任务管理提供了全面的任务调度与管理功能

- 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

理

- 周期任务管理

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

提供自动化的任

任务，减少人工干预，提升任务执行的自动化

动作，帮助用户高效管理日常定期检查与维护

- 脆弱性任务管理

户快速识别和修复系统中的安全漏洞，全面保障资产安全。





- **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

实时脱敏输出内容 → MAVAS

并告警数据治理部门。

封合法用户)。

## 司资产管理

提供全面的资产管理能力，帮助企业“摸清家底”，通过多种适配器的有机融

景下，新增了针对大模型的资产管控，使组织能够从全局掌控大模型应用相关的

产状况，为风险管控、全局监测提供基础支撑。

，通过云上和线下相结合的方式

资源和组件。这类实体资产主要包括：

库、数据湖、

- **应用软件：**对大模型应用相关的应用、组件进行统一管理，包括数据湖

模型

4.2.2 企业/单位自建大模型

提供基本信息和安全管理，便于对此类大模型进行全生命周期的

映射企业负责人，支持变更审批流程（如版本升级需安全团队审核）。

安全信息包括：大模型的总体安全性、合规性检测指标，使用中的风险行为及安全告

警事件信息等。

监测。

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

大模型实体维护，包括大模

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

大模型实体维护，包括大模

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

大模型实体维护，包括大模

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

大模型实体维护，包括大模

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

大模型实体维护，包括大模

大模型涉及的基本信息、安全管理、运行和部署等安全管理事项。

通过监测企业员工的

## 4.3 大模型安全分析

通过公在式关联分析引擎、异常行为分析引擎

平台其工多项目物的个是安全数

ATT&CK分析、知识图谱分析、并结合告警智能降噪、事件自动溯源等

UFBA 画像分析

的个是安全数

的个是安全数

的个是安全数

的个是安全数

的个是安全数

的个是安全数

### 4.3.1 大模型风险关联分析

的个是安全数

的个是安全数

其他类型的安全数据进行实时分析，不仅覆盖安全运营场

关联能力，将MAF、MSBA 以及其

分析引擎、告警安全

的数据关联分析需求，为云上堆栈应用安全场景提供了灵活、高效的

拓展的右效落地

自这些数据之后，其更实交易若到者的技

的个是安全数

，生成洞察层面的安全场忌。

术，它可以辅助识别网络威胁的自身的攻击样式

的个是安全数

平台接入 MAVAS、MAE、MASB 的安全个

的安全威胁进行实时检测与发现，实现多维度的交叉验证，动态关联模型和（如 MAVAS

检测的 CVE 漏洞信息）、异常用行为（如 MAE 捕获的恶意注入攻击）、恶意软件

的个是安全数

的个是安全数

的个是安全数

的个是安全数

综合判定风险等级并生成处置建议。

### 4.3.2. 基于用户身份的行为分析

行为分析系统对用户和实体（如主机、应用、网络流量和数据库）基于历史

数据或日志构建行为特征模型或基线进行分析，分析模型可识别异常特征行为或

异常行为，且能通过分析用户基线模型的打包分析来帮助发现威胁和潜在的恶意事

件。

对用户和实体的行为进行收集、处理和分析，主要关注网络中

行为管理负责

检测。通过对网络中的行为深度解析和识别，将行为数据

的行为模型和异常行为

管理，构建行为分析模型

和特征数据提供给模型

在大模型应用安全场景由一平台基于 MASB 用户的细粒度身份信息（包括用户角色、

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”三位一

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

和合规分析。

身份、行为结果综合评估画像，输出风险等级、可信度等级、异常行为特征

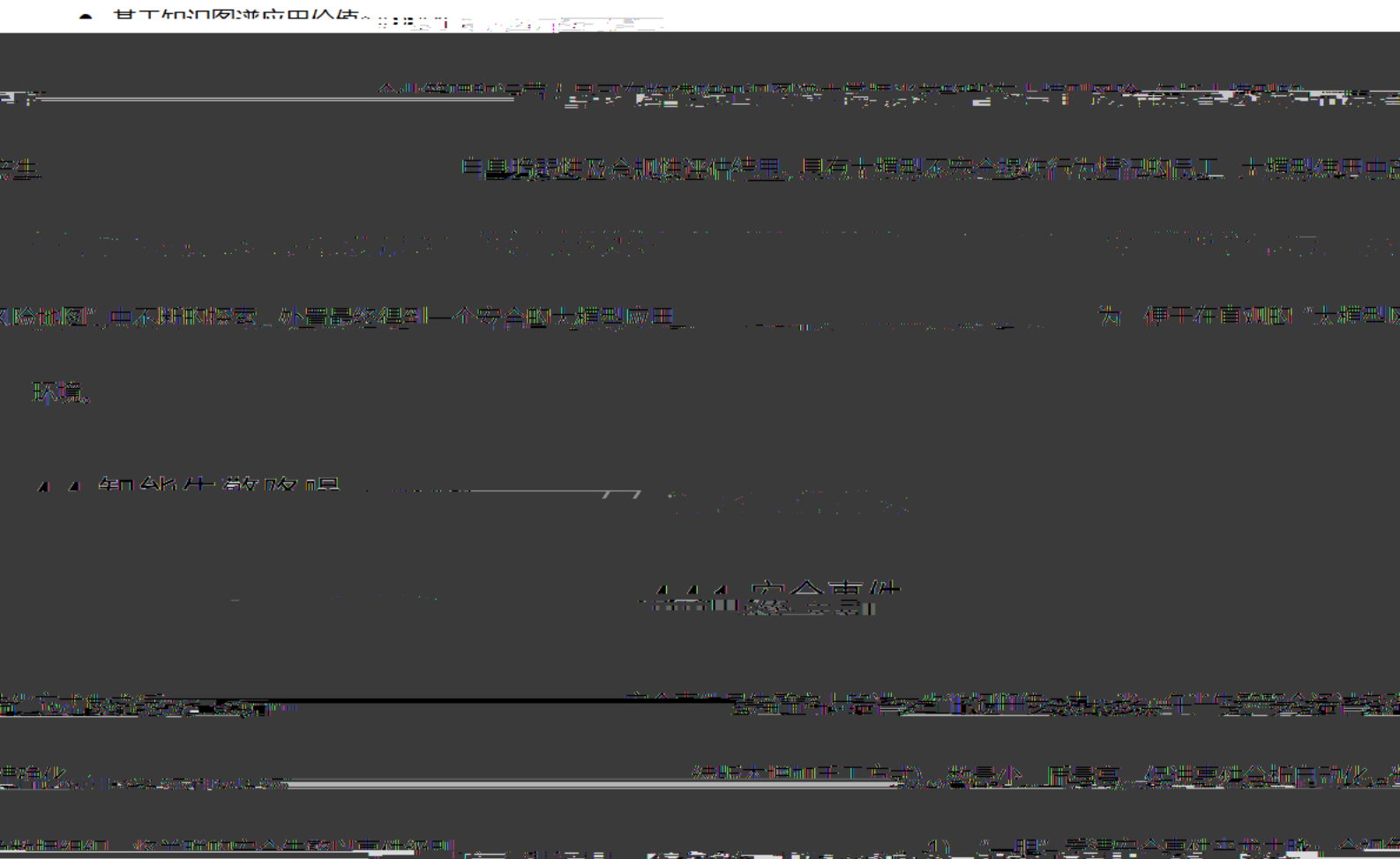
模型、模型应用数据

安全网络分析

AI 模型应用数据

- 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，

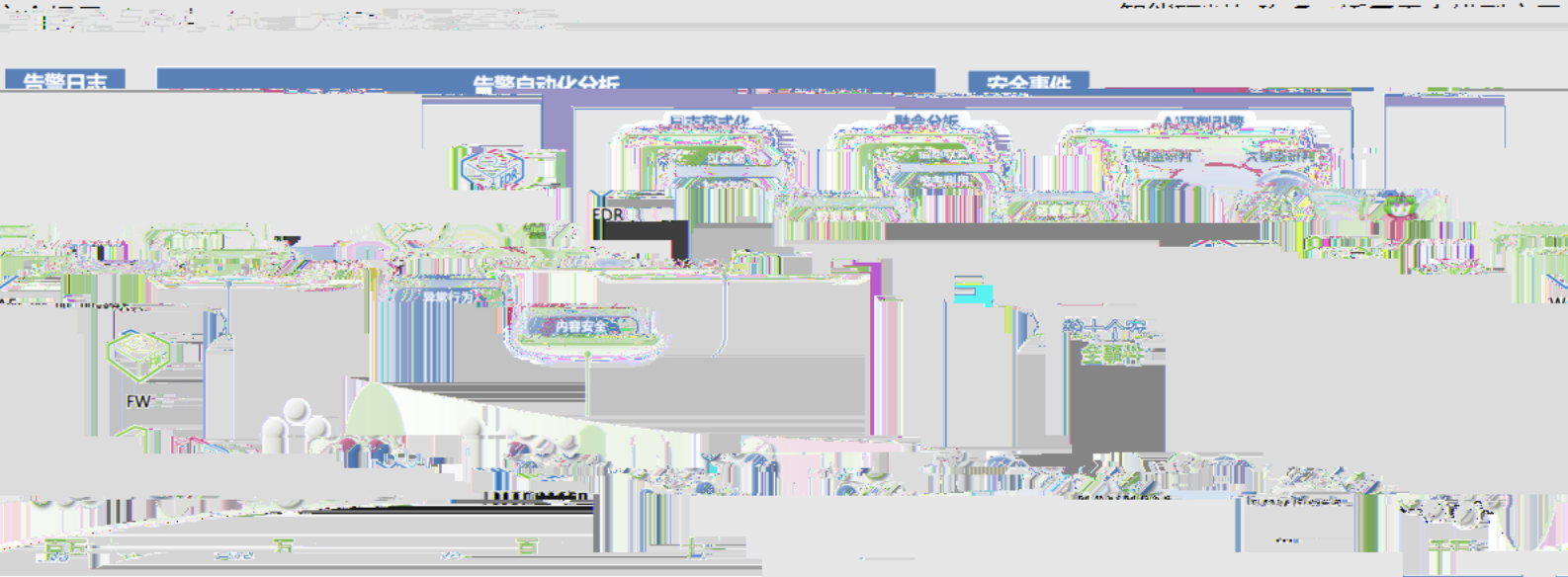
告警、告警、告警行为、以及进程信息（需两格）空报后FDR日志... 可以本系主机... 数据)。

告警的文本内容中的IP、域名、端口、进程、文... 3) 安全事件处置与抑制... 安全事件的外置... 抑制...

### 智能降噪

### 4.4.2

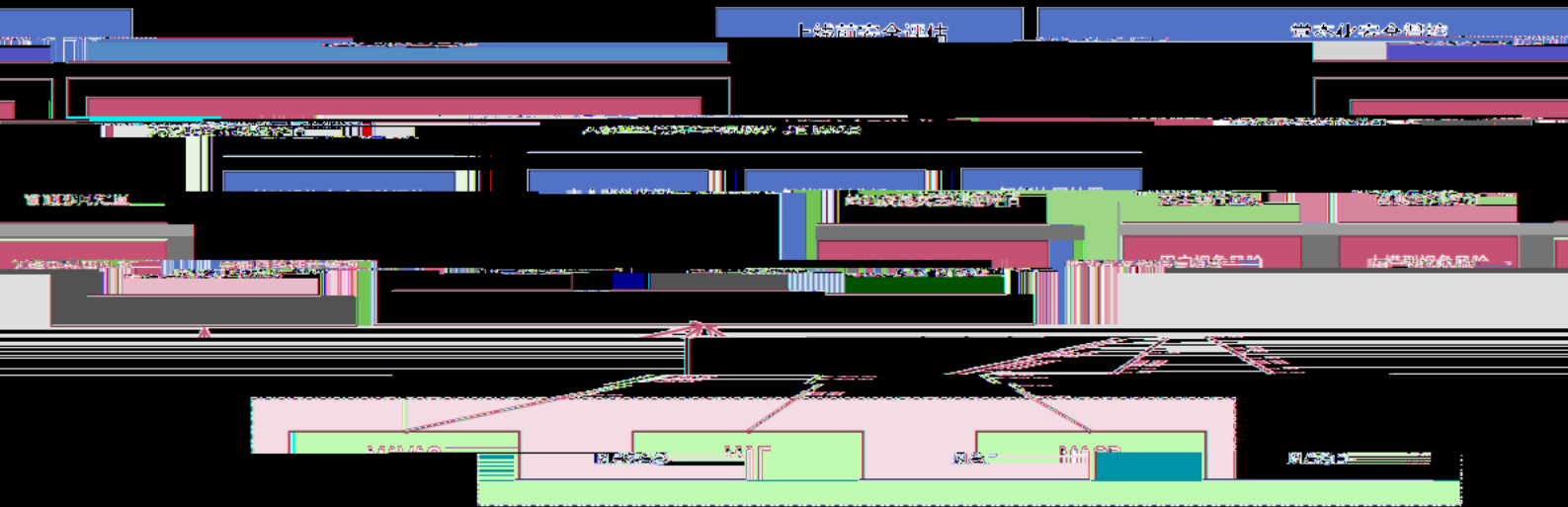
告警更关键的任务中。告警疲劳中解脱出来，投入到安全运营... 与 AI 能力进行深度融合，覆盖常见的安全事件类型的... 全息降噪引擎将安全专家的知识...



### 模型风险监测管控

### 4.5 大模

AI-R-SOCC 通过**全景式风险感知技术**，构建覆盖“模型本体-使用行为-业务环境”的... 多维度监测体系，实现从风险发现的全局可见，并可指导后续的联动处置形成大模型... 化的全链路闭环，助力企业实现大模型应用的**风险可视化、处置精准化、治理智能**...



### 4.5.1 大模型自身风险监测

#### 1) 上线前风险评估：采用 MAVAS 模型对模型输出在大模型上运行数据进行风险评估。

风险评估通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

对齐、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

平台进行集中展示、闭环管理。

全方位安全评估，风险评估结果在

实网环境中，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

输出中识别潜在攻击向量，通过 MAVAS 模型对模型输出进行风险评估。

检测评估，生成评估报告，MAVAS 的输出

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

通过生成攻击向量并分析模型输出，检测模型输出中是否存在与已知攻击向量相似的模式。

### 3) 内容合规实时检测：动态维护国家到行业再到企业单位的十措型禁用政策及规范

### 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

### 2) 高危用户画像：基于人员行为风险数据对企业中应用

## 4.5.3 风险行为/事件监测

### 1) 多维度告警聚合：将 MAI、MASA 所监测的大模型攻击行





提升数据流动风险的实时监测

● 大模型赋能风险检测态势：通过大模型赋能检测流程提升异常行为

精准识别异常行为、敏感信息泄露、异常操作等威胁维度，通过多源行为日志

挖掘

融合分析，便于快速识别高危人员并阻断违规行为。

信息，智能运营任务完成情况

● 大模型智能运营态势：智能运营态势大屏集成了总览

态势、重点人员活跃度TOP10和关键事件TOP10、威胁类型分布TOP10、威胁事件

态势、重点人员活跃度TOP10和关键事件TOP10、威胁类型分布TOP10、威胁事件

有效优化态势感知平台业务流程，为快速响应和精准溯源提供了坚实基础。



## 5 部署和典型应用场景

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“新三件套”

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

和高效

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

暂行办法》)、伦理准则(如歧视性言论),引发法律处罚与品牌危机等。

管理

● 解决方案

数据泄漏防控

1. 敏感数

MASD的 AI G... MASD引擎... 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)...

动态拦截高风险内容;

去违抑内... 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)...

实时拦截违规内容:

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)...

及时抑制风险的产生。

2. 生成内容滥用阻断

语义深度解析:基于多轮对话上下文理解,识别隐晦攻击指令(如“生成绕过审

核”);

核的贷款申请话术

实时拦截违规内容: 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)...

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),

大模型投毒攻击防御

3.

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 MAS 阳

源 IP;

断污染源

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

大模型训

认证数据源。

## 大模型对话交互

✓ 大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

过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

## 大模型内容审核

✓ 动态合规监测，输出内容实时比对监管要求（如《生成式人工智能服

务管理办法》），对

不合扣内容生成告警。

### ● 价值成果

大幅降低

低；

3. 减少业务损失：避免因模型滥用导致的品牌声誉损害与用户流失。

## 5.2 场景一：模型上线准入管控

### ● 痛点

数据污染、隐私泄露漏洞)；

“带病上线”风险（如训练数

### ● 解决方案

1. 自动化安全审查：



资源过载 (CPU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, 早触发实时预警与自动

修复策略

### 1. 漏洞实时监测与修复:

境 (如 CUDA 版本、容器镜像)

✓ 漏洞扫描: 联动 MAVAS 每小时扫描模型依赖环境

识别高危漏洞 (如 CVE-2023-1234)

自动推送修复建议 (如升级 TensorFlow 至 2.12)

✓ 自动化补丁: 对低风险漏洞

高风险漏洞触发模型隔离并告警。

### 2. 合规性动态适配:

动态适配: 实时同步法律法规更新, 自动调整策略

✓ 策略引擎: 支持策略热更新, 确保合规策略实时生效

(如禁止生成敏感数据输出)

✓ 策略引擎: 支持策略热更新, 确保合规策略实时生效

引擎实时扫描模型输出与合规知识库, 违规内容实时敏感

✓ 输出内容校验: 通过 NLP

阻断。

### 3. 服务能力保障:

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

(CPU > 90%) 动态调整: 低资源时触发降频或扩容策略

至轻量化模型版本。

### ● 价值成果

1. 模型合规率提升: 通过动态策略引擎, 模型输出内容合规率提升至 100%。

2. 政务问答模型输出内容合规率提升至 100%。

## 5.4 场景四：影子大模型监测与治理

✓ 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”：

企业员工私搭模型（如 Llama 2 代码生成工具）成为数据泄露与攻击跳板

### ● 解决方案

包括容器化部署的私搭模型），构建动态资产库；

✓ 自动发现企业内部模型实例（包

含特征）与代码扫描（检测私有模型仓库），自动构

建动态资产库；

搭建模型指纹库。

建私

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

✓ 通过

实时阻断。

要时

### ● 价值成果

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

● 大模

击面风险。

## 6 能力优势

### 6.1 模型安全

大模型安全运营平台通过持续运营发现漏洞、漏洞安全修复运营等安全运营手段，保障大模型

模型训练数据安全监测、推理过程行为合规等任务，实现大模型安全运营的日常工作的自动化处

理。利用先进的可视化技术，让数据流向、风险检测点、任务执行进度等关键

信息一目了然，帮助用户快速定位问题环节，高效跟踪任务进度，保障大模型安全运

的环境下持续运行。

### 6.2 集中调度

模型数据安全防护、模型算

充分考虑大模型安全涉及的多维度安全能力，将各类针对大

能力“烟囱式”隔离，把

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

分散在不同系统的安全功能碎片化，统一形成统一的安全能力中心，主动防御对抗归属

安全运营效率，为大模型构建全方位的安全防护网。

### 6.3 快速响应

大模型安全运营平台通过持续运营发现漏洞、漏洞安全修复运营等安全运营手段，保障大模型

模型训练数据安全监测、推理过程行为合规等任务，实现大模型安全运营的日常工作的自动化处

理。利用先进的可视化技术，让数据流向、风险检测点、任务执行进度等关键

信息一目了然，帮助用户快速定位问题环节，高效跟踪任务进度，保障大模型安全运

## 6.4 安全专家

数据解析能力，提供针对大模型安全的党田安全知识智能

依托强大的大模型知识图谱与类

进行行为等，生成详细的安全事件分析报告，帮助安全运营人员快速识别威胁并

提供精准的威胁情报，帮助安全运营人员快速识别威胁并

提供精准的威胁情报，帮助安全运营人员快速识别威胁并

体作战实力，从容应对各类大模型安全威胁。

警技术，实现 7×24 小时全年无间断的大模型安全运

采用先进的自动化监控与智能预

合的运营流程，以及精细化的模型访问管理的数据流等，立即触发预警并启动应急响应

合的运营流程，以及精细化的模型访问管理的数据流等，立即触发预警并启动应急响应

合的运营流程，以及精细化的模型访问管理的数据流等，立即触发预警并启动应急响应

