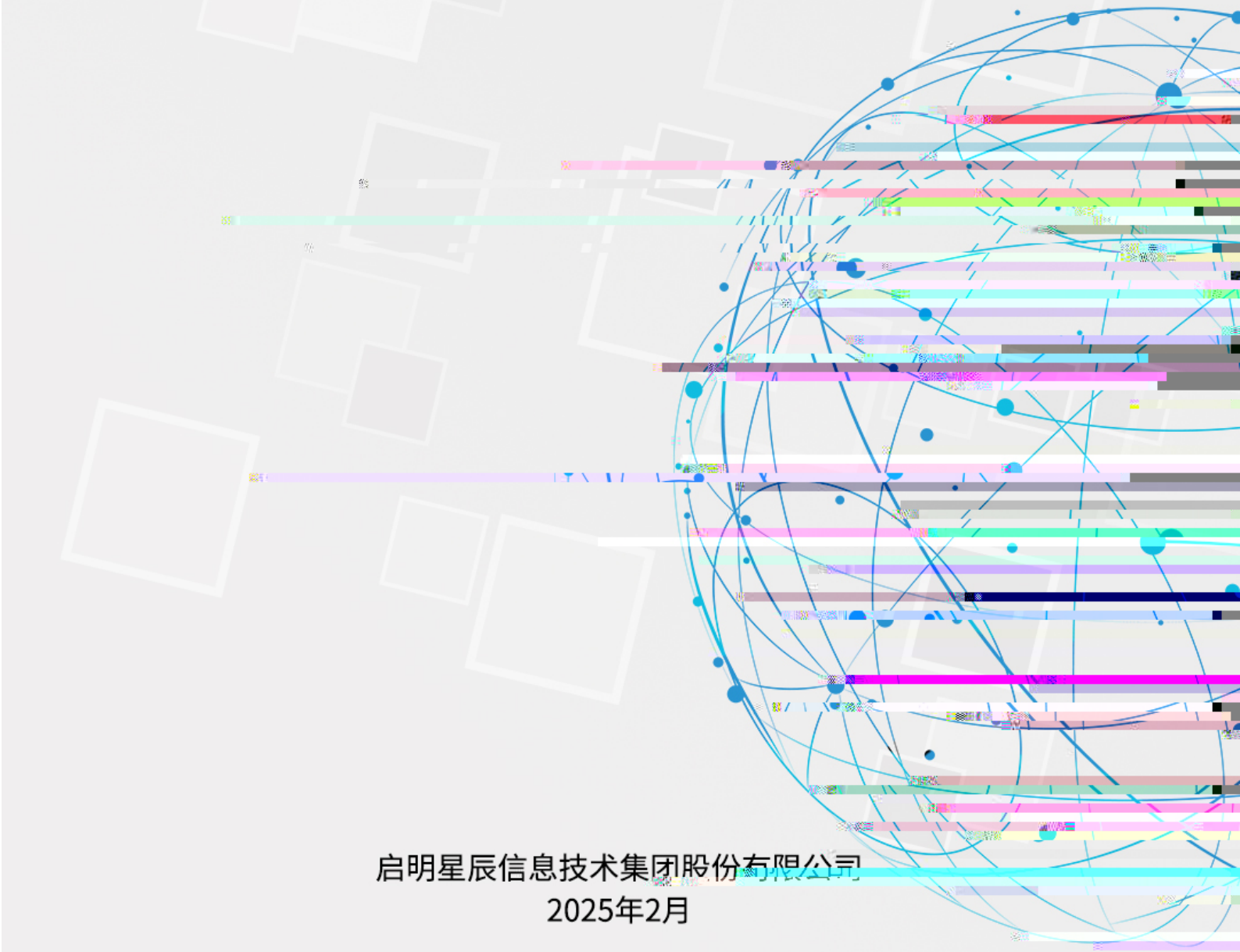
启明星辰

AI就绪的大模型身份与访问管理

AI-R-IAM v1.0

AI-Ready Identity and Access Management



启明星辰信息技术集团股份有限公司
2025年2月

版权申明

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。

未经北京启明星辰信息安全技术有限公司书面同意，任何人不得以任何方式或形式对本手册内的任何部分进行复制、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。

免责申明

本文档依据现有信息制作，其中资如有更改，

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遗漏、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：

邮箱：

地址：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：010-82779083

传真：010-82779000

目录

1 公司简介.....4

2.1 新质生产力“大模型”.....6

2.2 大模型应用中安全挑战.....7

2.2.1 身份与访问控制的安全挑战.....7

2.2.2 数据安全挑战.....9

3 身份与访问管理系统（IAM）.....11

4 AI 就绪的大模型身份与访问管理.....13

4.2 可信权限管理.....16

4.4 可信行为审计.....19

5 应用场景.....21

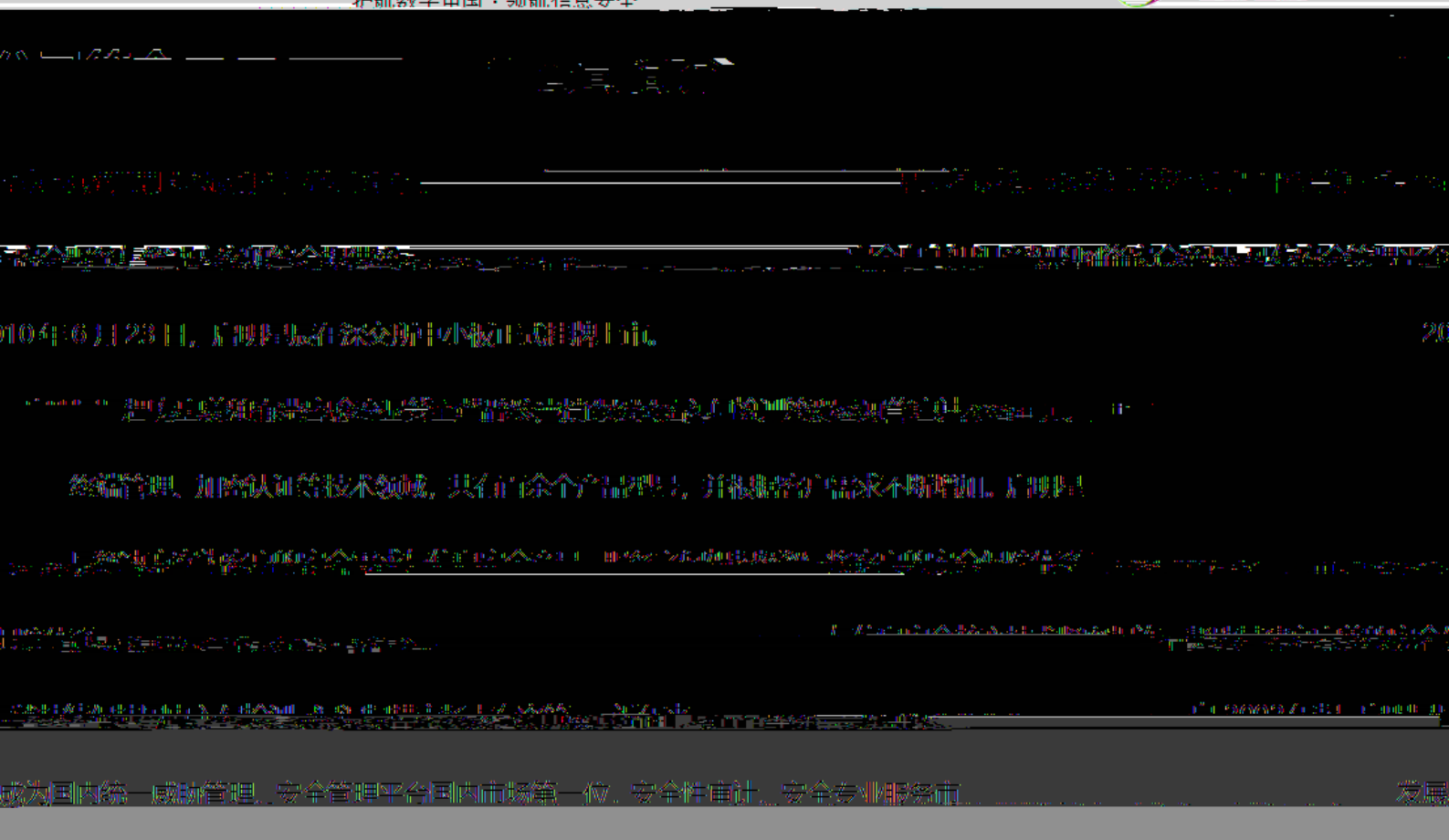
5.1 大模型应用场景数据流.....21

5.2 场景一：用户访问大模型应用场景.....21

5.3 场景二：应用访问大模型应用场景.....26

5.4 场景三：用户/AI智能体的安全应用管理场景.....30

6 发展趋势展望.....31



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

渠道和 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
售后服务体系。

泽民、 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
书记亲 李岚清、曾庆红等党和国家领导人亲切视察启明星辰公司；2003 年 1 月，胡锦涛总书
切接见了启明星辰公司 CEO 严望佳博士。

计划软 凭借多年来的潜心研发，启明星辰获得国家规划布局内重点软件企业，国家火炬计
计算 件产业优秀企业，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
机信息系统集成资质证书。

业化 启明星辰目前是我国拥有最大的国家级网络安全研究基地，完成包括国家发改委产

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，国家科

金项专利和软件著作权，参与制订国家及行业网络安全标准，填补了我国信息安全科研领域

[redacted]

[redacted]

[redacted]

2 大模型的发展与风险

2.1 新质生产力“大模型”

PT，近年来，大模型技术取得了突破性进展，成为人工智能领域的核心驱动力。从 ChatG

理、再到百度文心一言、阿里通义千问、九天大模型、DeepSeek 等，大模型在自然语言处

能问答、图像[illegible]够学习到丰富的语言知识、语义理解和逻辑推理能力，从而实现文本生成、知

。人[illegible]生成、代[illegible]

期）：○ 准备期（2022 年 11 月

在 ChatGPT，让[illegible]2022 年 11 月，OpenAI 发

在该细[illegible]人[illegible]

年）：○ 成长期（2023

[illegible]

问，为[illegible]（2）国内：[illegible]，[illegible]，阿[illegible]

技术储备。[illegible]模型发展提供

[illegible]

大模型社区[illegible]（1）国外：OpenAI 公司等[illegible]

[illegible]

（2）国内：2024 年 5 月 24 日，在中国移动人工智能生态大会上，中国移动正式发布

了“九天”人工智能基座，包含万卡级智算集群、千亿多模态大模型、汇聚百大要素的生态

- **大规模应用期（2025 年 1 月-）**

（1）国外：大模型深入医疗、金融、教育等多行业，助力疾病诊断、风险评估、个性化学习等。OpenAI 还在探索新应用领域。

（2）国内：DeepSeek 凭借创新算法和架构，展现出低成本、高效能以及开源优势，

应用中安全挑战

2.2 大模型

，都暴露出诸多安全问题。

Identity，NHI）

1. **人类身份安全问题：**人类身份代表自然人身份实体，在大模型应用交互场景中的背

2. AI Identity安全问题：

非法访问权限。

（2）访问权限放大：与人类身份不同，AI Agent不会以可预测的方式请求访问权限，

和操作敏感资源，造成数据泄露或系统破坏。

（3）自身安全：AI Agent身份系统通常存储大量用户的个人信息、行为数

系统遭受黑客攻击，这些数据可能被窃取，导致用户隐私泄露，进而可能被

定向广告骚扰等。

3、非人类身份（NHI）安全问题：非人类身份定义为企业技术内的应

2025年2月发布《大大NHI风险2025》，如下图所



应用中，大多数存在于大模型基础设施交互过程中，NHI 的身份生命周期管理薄弱，存在诸多安全风险与问题：

NHI 在大模型

(1) [illegible]

被内外部威胁者发现。

被

(2) 影子账号：大模型应用迭代快，加之生命周期流程弱、账号使用信息不可见，许多 NHI 账号不活跃，增加攻击面。

N

(3) 缺乏账号所有权：多数参与大模型应用的组织中，NHI 无明确所有权信息，而明确管理者对安全维护和问题补救很关键。

(4) 缺乏环境隔离：在许多情况下，相同 NHI 会在生产和非生产环境中使用，或者相同的逻辑 NHI 在每个环境中都有相同的密码，从而增加了横向移动的风险。

[illegible]

杂，难以掌握所有依赖关系。

2.2.2 数据安全挑战

大模型应用在数据安全方面呈现如下挑战：

(1) 用户提示词输入输出风险：用户提示词可能包含敏感信息，若大模型系统对输入提

机制不完善，敏感信息可能在输入环节直接暴露。在输出结果时，若对输

示词的验证和过滤机

(2) 大模型 API 调用风险：API 调用

这个插件，为其设置、调整参数和选项。例如，行为可能会获取 API 调用权限，

非法获取金融大模型平台调用广告广告数据。

(3) RAG 知识库投毒风险：RAG 知识库接入大模型，若知识库

(4) 数据投毒风险：在大模型

(5) 个

可信环境基于大数据安全能力，实现网络风险、行为风险、数据风险、业务风险、设备风险的全链路安全风险审计。

访问可信与数据可信作为一体两面，将防控范围扩展到安全设备、系统资源、应用资源、API 资源等，通过“ID+身份”的细粒度可信身份与跨终端设备、跨平台系统操作、可

制的融

组合性和可扩展性，使企业能够以更少的资源实现更好的安全性，搭配一体化安全机制

合协同，实现全程全网、端到端的安全保护。

个人中应用越来越深入，需要针对其身份、

大模型技术特点和业务场景的一体化全程

访问管理。

随着大模型作为新质生产力，在政府、企业、个

访问管理、数据安全等方面的需求，建立一套基于

可信的数字身份基座，即 AI 就绪的大模型身份与

4 AI 就绪的大模型身份与访问管理

M)上为人机混合 AI 智能体(AI

AI 就绪的人机型份访问管理(以下简称:AI R IA

型为人机型份训练、测试和出厂检查多个环节

全从“单点可控”迈向“一体化全程可信”保护体

属性支撑。

构建一体化全程可信能力，构建大模型安

系，同时为其它安全能力提供身份管理属



型。系统不仅支持传统

先进的身份认证机制，确保只有经过授权的用户和系统能够访问大模型

名认证能力，有效防

的用户名和密码验证，多因素认证，还引入基于中国移动号卡特性的实

止未经授权的访问和潜在的安全威胁。

管理

- 可信访问权限

对多个大模型访问、AI 接口访问、RAG 访问提供了集中的精细化的权限

AI-R-IAM 针

用户的操作体验。

限管理不仅提高了系统的安全性，还增强了用

- 可信审计管理

业务调用过程进行安全审计和实时监

AI-R-IAM 对大模型使用过程和大模型内部组件

IP 地址、访问的大模型资源、执行的

控，记录用户的所有操作行为，包括用户登录时间、

操作等信息，实时监测用户和组件的访问行为，及

信能力，能够有效应对当前大模型面临的安全挑战，

AI-R-IAM 通过构建一体化的全程可

全发展奠定了坚实的基础。

工智能技术的安

4.1 可信身份治理



份和属性的治理和管理，AI-R-IAM 的“身

AI-R-IAM 支撑基于大模型特性的多维度身

周期

针对用户、智能体、API 应用程序建立集中一体的身份标识，对身份和属性进行全生命周

的管理，是数字世界秩序的底层支撑。

证能

AI-R-IAM 采用了先进的身份认证机制，同时引入中国移动基于号卡特性的实名认证

力，有效防止未经授权的访问和潜在的安全威胁。

AI-R-IAM 不仅优化传统人类身份从入职到离职的全生命周期管理，还创新性地将 NHI

和 AI Identity 纳入统一治理范畴，助力企业实现数智化转型，有效降低安全风险，提升合

规性。在具体流程如下：

(1) 用户身份在入职、转岗、离职时，进行信息收集验证、权限调整及账号注销等操作；

4.2 可信权限管理

算法工程师	✓	✓	✓	×	×	×
					✓	✓
业务用户	×	受限调用	×	×	×	×

- AI 接口权限管理

户名单等进行权限参数管理。

权限隔离

只能访问其业务相关的知识库片段；

的段

- RAG（检索增强生成）

文档级权限控制：不同部门

实时内容过滤：结合语义分析引擎，自动屏蔽无权限的检索结果（如涉及竞对信息



4.4 可信行为审计



数据采集

- 多维数

构建全面的用户行为基线分析提供支撑

- 安全元数据管理

构建多维分析模型，反映大模型用户的特征、行为习惯、个人便好，生成用户画像、模

型画像、数据画像等，并根据风险画像评分，智能量化风险水平。

知识图谱

- 数据

多源融合、关系排序、深度去重等深度治理及分析，实现不同实体之间的关联

通过数据多

供从“关系”的角度分析大模型风险的能力。

关系构建，提供

深度分析

- 数据

理全生命周期各个环节进行合规风险分析，重点数据审计、敏感数据和异常场

大模型数据

对大模型安全风险进行评估，识别已知风险和未知风险。

景分析等，

智能追踪溯源

- 智

实现风险出现后个体的智能追踪溯源，采用路径计算、路径探索算法，根据风险事件发生前后的

析，实现风险智能追

可疑 IP->访问数据分布->可疑路径->可疑日志链路分析，层层递进分

踪。

5 应用场景

份与访问管理，为大模型的访问、使用、调用、训练等多 AI-R-IAM AI 就绪的大模型身

可信权限、可信审计、可信数据能力。相关场景包括：用 种场景提供可信身份、可信认证、

户访问私有大模型、公域大模型场景；业务系统调用智能体、RAG 知识库、私域大模型场

景；大模型数据训练场景、大模型数据运维场景等。



5.2 场景一：用户访问大模型应用场景

AI-R-IAM AI 就绪的大模型身份与访问管理能够为用户访问大模型应用场景提供可信

私域大模型、运营维护人员访问私域大模型等。 户访问公域大模型、用户访问私



1. 普通用户访问公域大模型

时，可信身份认证方面，构建了一套严谨的身份验证机制。当用户首次访问大模型应用时

保护用户的数据不被未授权访问。

持续监控整个系统的运行状况。它会记录下所有用户的操 在安全审计方面，AI-R-IAM

了大量不符合常规模式的数据访问请求，就可以触发警报提示管理员进行进一步调查，从而

有效地维护系统的稳定性和安全性。

2. 普通用户访问私域大模型

普通用户通过终端设备访问部署在企业内部的私域大模型应用（例如知识问答、文档生成等

运营维护人员访问私域大模型应用服务器，负责日常维护和故障排查。由于运维人员权

上，还需要重点加强精细化权限管理、高危操作、金库管理等方面能力。

而，如果是在非工作时间或者从企业外部网络尝试访问时，就需要

大模型应用服务器，然

。

发更加严格的验证机制

就需要立即启动阻断机制，这种阻断机制可以分为多个

权限配置等，一旦识别出高危操作，

维人员尝试执行高危操作时，系统会弹出明显的警告信

层次，在第一层次是警告提示，当

5.3 场景二：应用访问大模型应用场景



应用访问大模型场景主要包括业务应用访问公域大模型、业务应用访问私域大模型两种。

应用场景，业务系统智能体安全管理和访问公域、私域大模型、RAG知识库、AI-R-IAM

对业务应用电脑以下安全措施：

1 业务身份管理

身份冒访，

如密钥可能被硬编码在代码中，或者存储在不够安全的地方，攻击者可能通过伪造

证卡请求服务，把系统配置为使用该证书信任

作，

AI-R-IAM业务身份创建，修改和删除的全流程管理，实现业务身份

最小访问权限。

务系统身份信息和权限，确保合规并遵

身份修改：需根据功能和任务变更情况，更新业

原则。

循最小权限

毁：删除业务系统身份信息并销毁凭证，同时进行关联资源清理和安全检查，确

身份销

全稳定运行。

保系统的安

2. 业务权限管控

业务系统在访问公域、私域大模型时，如果没有基于最小权限原则来分配访问权限，如

某些功能可能只需要读取权限，却被赋予了写入或管理权限。由于权限划分不清晰，可能导

致内部人员误操作或恶意操作，进而影响系统整体安全。

[redacted]

3. 业务行为审计

[redacted]

整、合规性检查不足、数据隐私泄漏等安全问题。

[redacted]



获有限权限，同时通过身份绑定对双方行为审计监管，保障访问数据和训练数据的安全。通过 AI-R-IAM 能力解决如下问题：

1. 用户到智能体全链路身份可信

（AI

通过 AI-R-IAM 提供可信身份管理，赋予用户可信身份（user）与智能体身份

Identity）中统一身份ID进行绑定管理，明确在审计层面明确一般智能体身份

，方便界定相应责任。

项操作，系统可追溯到是谁让智能体进行操作，

智能体会与特定的代码及模型紧密绑定，运用

为了有效防止智能体被恶意掉包或仿冒，智

中预先指定的模型和代码构建，并成功部署

数字签名或证书技术，明确证明某个智能体是

，同时评估其可信度，从而确保系统运行的安

格验证，以此来精准确认该智能体的真实身份

全性和可靠性。

(如运行环境指纹，地理位置等）进行增强认证。

3. 关联审计与监管

通过身份绑定机制，对普通用户使用智能体行为，以及智能体自身的行为进行审计和监

管，当智能体发生越权操作或异常行为时，AI-R-IAM 系统应有能力快速定位并冻结该代理

依据何种规则。

泄

在企业内部，智能体易成为攻击目标，被攻击后可能执行恶意操作、造成训练数据外

新风险，如自动提交敏感数据到外部。

务器，RAG 知识库，

AG 的数据访问权限

体异常行为等。避免

在此场景下，AI-R-IAM 提供用户和智能体，智能体与数据存储服

内外部大模型之间的数据权限边界，实行最小化策略，制定智能体对 R

范围，辅以模型护栏、上下文检测等安全控制，可及时阻断和审计智能

因智能体自主决策引发大规模安全事故。

6 发展趋势展望

异常行为。

信；借助 5G/6G 网

术，保障大模型在物联网场景下与海量设备交互时的数据安全和身份可