AICSO AI安全官白皮书

Artificial Intelligence Cyber Security Officer Competency Framework (AICSO-CF V1.0)

V1.0



AICSO 人工智能安全官社区

版权声明：本白皮书为行业研究成果，仅供行业交流参考，不构成法律、合规、商业建议。

前言

本白皮书是由 AICSO 人工智能安全官社区发起的行业**公益性研究项目**，面向大众免费开放，旨在**推动 AI 安全官岗位体系、能力标准和企业实践方法的建设**。

本白皮书核心内容为 AI 安全官能力框架，实际更接近于一本 **AI 安全管理者的日常工作手册**，我们旨在回答 AI 系统在企业深度使用后，企业应该如何开展 AI 安全治理工作，如何定义 AI 安全领导职责、建立能力体系并形成可持续治理机制，从而有效控制风险。

本白皮书从 AI 安全官定义、个人能力、技术体系、治理合规和组织建设等维度，构建 AI 安全官工作统一描述语言，为企业 AI 安全人才培养、岗位设计、能力评估和 AI 安全体系建设提供参考和指引。

本白皮书为首次编制，编写时长跨度 3 个月，鉴于当前 AI 技术的快速更新，若有时效性的表述不准确，请以最新内容为准。同时，鉴于在各行各业的真实 AI 应用水平和真实安全风险考量，部分已知的未解决问题将在 V2.0 版本中修订，这一部分内容在第 11 章有详细的呈现。

本白皮书在编制过程中，得到了诸多安全行业从业者，尤其是国内外高等院校、大模型企业、互联网企业、金融企业一些安全同行的大力帮助和支持，在此表示深深地感谢！

感谢联通（四川）产业互联网有限公司王颉、 山东警察学院张璇，两位行业专家给白皮书的编制提供诸多宝贵建议！

但限于编者自身水平、AI 技术快速迭代以及安全场景的复杂性，文中难免存在疏漏、不足或有待精进之处。我们诚挚欢迎各位行业专家、从业人员及广大读者提出批评、指正与修订建议。如您有内容勘误、观点补充、案例共建或标准优化意见，均可发送邮件给 hello@aicso.ai 联系我们，也欢迎您扫码加入 V2.0 修订专家组，和我们一起将此白皮书建设得更加完善！



AICSO 人工智能安全官社区 钱君生

2026 年 9 月 20 日

AICSO AI 安全官白皮书

名称	《AICSO AI 安全官白皮书 V1.0》
英文名称	Artificial Intelligence Cyber Security Officer Competency Framework（AICSO-CF V1.0）
编制单位	AICSO 人工智能安全官社区
编写人员	邓诗智、路程、罗喆帅、朴杨鹤然、钱君生、涂宏伟、王立敏、王颉、张璇
版本	V1.0
日期	2026 年 9 月
协议	CC BY-NC-SA 4.0

目 录

第 1 章 概述 7
1.1 背景与目的 7
1.2 4-12-5 能力模型 8
1.3 本白皮书结构 8
第 2 章 AI 安全官定义体系 10
2.1 AI 时代的风险与挑战 10
2.1.1 传统安全与 AI 安全的差异 10
2.1.2 从网络安全到 AI 安全 11
2.2 AI 安全官角色定位 12
2.3 AI 安全官/CISO/CSO 及周边关系 12
2.4 AI 安全官工作重心：AI Security & AI Safety 13
2.5 AI 安全威胁 15
2.5.1 威胁分类矩阵 15
2.5.2 新兴威胁趋势 17
第 3 章 AI 安全官 4-12-5 能力模型 19
3.1 能力模型设计原则 19
3.2 四大能力域 19
3.2.1 能力域 D1：AI 技术与攻防安全 20
3.2.2 能力域 D2：AI 工程与架构安全 20
3.2.3 能力域 D3：AI 治理与风险合规 21
3.2.4 能力域 D4：业务与组织领导力 21
3.3 十二项核心能力 21
第 4 章 AI 安全官职业等级 24
4.1 AI 安全官五级划分 24

4.2 等级特征对比 25
4.3 各等级详细描述 26
L1 AI 安全助理工程师 26
L2 AI 安全工程师 26
L3 AI 安全架构师 27
L4 AI 安全总监 27
L5 AI 安全官 28
第 5 章 AI 安全官能力评估 29
5.1 能力评估方法 29
5.2 能力评估细则 29
C01 算法与数据安全 29
C02 AI 对抗防御 30
C03 AI 红队测试 32
C04 AI 基础设施安全 33
C05 AI 供应链安全 34
C06 AI 安全运营 35
C07 AI 治理体系建设 36
C08 AI 法规与合规 37
C09 AI 伦理与知识产权 38
C10 AI 业务风险评估 39
C11 AI 安全事件管理 40
C12 安全沟通与培训 41
5.3 企业 AI 采纳深度与能力适配 42
5.3.1 三种企业 AI 采纳类型 42
5.3.2 不同类型 AI 企业能力适配 43
5.3.3 风险优先级差异 44

5.3.4 团队配置建议 45
第 6 章 AI 安全考核指标 47
6.1 考核指标概述 47
6.2 指标设计原则 47
6.3 核心考核指标参考案例 47
6.3.1 业务价值考核指标拟定参考 49
第 7 章 AI 安全官岗位 51
7.1 AI 安全官岗位说明书 51
7.2 AI 安全官 JD 模板 51
模板 1：AI 安全工程师（大模型安全方向） 52
模板 2：AI 安全工程师（AI 对抗与模型安全方向） 52
模板 3：AI 安全工程师（AI 应用与数据合规安全方向） 53
模板 4：AI 安全官 54
第 8 章 组织架构 55
8.1 AI 安全治理组织架构参考 55
8.2 AI 治理委员会 56
8.3 各企业规模的团队配置建议 56
第 9 章 企业案例与行业建设路线 58
9.1 按企业规模的建设路线 58
9.1.1 初创企业（50 人以下） 58
9.1.2 中型企业（50-500 人） 58
9.1.3 大型企业（500 人以上） 59
9.2 按行业的建设重点 60
9.3 典型企业案例 61
案例一：大模型研发型企业 A 61
案例二：互联网 AI 应用企业 B 62
案例三：金融企业 C 63

第 10 章 中国及欧美 AI 监管要求映射 65
10.1 AI 监管合规框架概览 65
10.2 中国 AI 监管体系 65
10.2.1 AI 安全监管相关法律法规 65
10.2.2 大模型备案/登记、安全评估与算法备案 67
10.3 ISO 4200 与中国法规的对应关系 69
10.4 欧美 AI 安全治理 70
10.4.1 欧美 AI 监管规则层级体系 70
10.4.2 AI 安全官能力模型与欧美监管要求映射 71
第 11 章 版本演进与迭代说明 73
11.1 版本管理 73
11.2 问题清单 73
11.3 社区生态 74
第 12 章 附录 75
附录 A：AI 风险评估检查表 75
附录 B：AI 安全能力自测问卷 77
附录 C：AICSO 术语表 77
附录 D：AI 安全工具推荐 80
附录 E：参考资料 81

第 1 章 概述

1.1 背景与目的

过去两年，人工智能从实验室走向生产环境的速度超出了大多数人的预期。大语言模型、多模态模型、AI Agent 以及正在兴起的 MCP 生态，正在以前所未有的速度重塑企业 IT 架构。与此同时，生成式 AI 与智能体扩大了企业的攻击面，提示词注入、训练数据投毒、模型窃取、对抗攻击、AI 供应链投毒.....这些之前很少见的威胁，如今已经成为安全团队日常面对的课题。

传统的网络安全体系是以企业信息化为核心的安全治理框架，擅长处理 IT 信息化系统的漏洞管理、入侵检测、数据泄露防护，对 AI 系统特有的风险类型缺乏系统性的应对方案。随着 AI 技术在企业内部的广泛应用，企业亟需一个能够跨 AI 技术应用、安全治理和业务风险的专职角色来开展工作。这个角色可以由独立新设岗位承担，也可以由首席安全官、AI 负责人或其他高级安全角色兼职承担。为了方便描述，本白皮书在现有安全治理框架基础上，提出此角色的岗位设计和能力发展参考模型，我们将承担这一核心职责的角色称为**AI 安全官**。考虑到业内习惯使用 CSO（首席安全官）或 CISO（首席信息安全官）描述企业安全负责人，本白皮书中沿用 AI+CSO 称呼 AI 安全官，即 **AICSO**。

然而，在角色确定后，当我们试图定义 AI 安全官应该具备什么样的能力，在企业内应该如何开展工作、如何评价工作时，却发现行业中缺乏一个被广泛认可的标准。现有的安全体系大多聚焦于传统网络安全，对 AI 特有的攻防技术、治理需求和合规要求均覆盖不足，所以我们编制了这一框架，试图填补这一领域的空白。

本白皮书试图阐述清楚以下几个问题：

1）为 AI 安全官这一角色划定清晰的能力边界，即在懂 AI+懂安全之上，有明确的能力域和技术职级体系。

2）要为企业提供从零开始建设 AI 安全团队的路线图，无论是 50 人的初创公司还是 5000 人的大型集团，都能在本白皮书中找到参考依据。

3）为每一位有志于 AI 安全领域的从业者，勾勒出从 AI 安全入门人员，到 AI 安全官的职业发展路径。

4）甚至，希望它能为整个行业提供一套统一的话语体系，让 AI 安全官不再是一个空泛的标签，更是有着实操落地的样例和典范。

1.2 4-12-5 能力模型

本白皮书对 AI 安全官的能力模型采用了一个简洁的数字结构：**4-12-5**，如下图 1-1 所示：



图 1-1 AI 安全官 4-12-5 能力模型

其中 **4** 代表四大能力域，分别是 AI 技术与攻防安全、AI 工程与架构安全、AI 治理与风险合规、业务与组织领导力；**12** 代表十二项核心能力，每个能力域下设 3 项能力，4 个能力域共 12 项；**5** 代表 AI 安全官职业发展的 5 个等级，从 L1 AI 安全助理工程师到 L5 AI 安全官。

采取这样的结构设计，初衷在于平衡专业深度与模型的可传播性。既想让它足够精细，覆盖 AI 安全的各个维度；同时又足够简洁，方便企业招聘、人才评估和个人成长等场景使用。当前该模型为社区提出的参考框架，尚需通过不同企业和岗位的试用检验，以确定覆盖及评估一致性。

1.3 本白皮书结构

本白皮书共分十一章正文，及 5 个附录。

前四章是基础章节，主要是建立 AI 安全官能力框架的认知共识，从背景和角色定义出发，逐步深入到能力模型和职级体系。第五章是整个 AI 安全官能力框架的核心，详细展开十二项能力在五个职级上的具体要求，并根据企业 AI 采纳深度（AI 全栈型、私有化

部署型、AI 消费型三种类型）提供差异化的能力适配建议。第六至第八章聚焦于组织层面，讲述 AI 安全考核指标如何定义、关键岗位如何设置、团队如何搭建。第九章通过三个行业案例，展示不同规模、不同行业的企业如何落地这一框架建议。第十章将本白皮书的能力框架，与中美欧现有 AI 监管法规进行映射，为合规落地提供指引。

5 个附录包含风险检查表、能力自测问卷、术语表和 AI 安全工具推荐等，供读者在实际工作中直接使用。

第 2 章 AI 安全官定义体系

2.1 AI 时代的风险与挑战

要理解 AI 安全官存在的必要性，要先从当前 AI 技术应用引入了哪些传统软件安全从未遭遇的风险说起。

2.1.1 传统安全与 AI 安全的差异

在传统安全中，主要防护服务器、网络、终端、应用、数据等 IT 基础设施与业务系统等，安全团队通过安全架构、代码审计、渗透测试、安全运营等体系化操作，建立相对稳定的防御机制。AI 安全的防护对象主要是模型、训练数据、提示词、推理服务、智能体等，其技术特点是概率性行为，即使代码无漏洞，模型本身也会产生风险，同一输入也可能得到不同结果。

下表 2-1，从 10 个方面列举了传统安全与 AI 安全的差异，这是两者之间的根本区别：

表 2-1 传统安全与 AI 安全风险差异对比

对比维度	传统安全	AI 安全
防护对象	服务器、网络、终端、应用、数据等 IT 基础设施与业务系统	AI 模型、训练数据、提示词、推理服务、智能体(Agent)、向量数据库、AI 工作流
核心威胁	漏洞利用、病毒木马、钓鱼、DDoS、权限越权、数据泄露、恶意代码注入	提示词注入、模型投毒、数据污染、模型窃取/逆向、越狱、对抗样本、智能体越权、幻觉导致错误决策
攻击入口	端口、接口、文件上传、账号密码、系统漏洞、邮件链接	Prompt 输入、训练数据集、微调数据、模型权重、上下文记忆、Agent 工具调用
安全假设	程序逻辑固定，代码行为可预期	模型是概率行为，输出具有不确定性，即使代码无漏洞，模型本身也会产生风险
安全目标	保障保密性、完整性、可用性（CIA 三元组），阻止未授权访问与系统破坏	除 CIA 外，还有：可靠性、可解释性、公平性、防误导、防止自主 Agent 执行有害动作
检测方式	特征匹配、规则引擎、行为基线、恶意代码哈希、日志审计	对抗样本检测、提示词风控、输出内容审核、模型漂移检测、数据污染识别、幻觉评估
风险触发	攻击是攻击者主动构造恶意载荷触发，正常输入不会破坏系统	风险可由正常输入诱导（提示词注入），模型自发产生幻觉，不一定存在恶意代码
权限边界	权限由权限矩阵确定，权限不会动态扩大	Agent 可自主调用工具，自主决策，权限边界动态变化

安全手段	防火墙、WAF、IDS/IPS、杀毒、访问控制、加密、漏洞扫描、渗透测试	提示词防护、模型水印、数据清洗、对抗训练、输出过滤、Agent 能力沙箱、模型评估、红队 AI 攻防
业务场景	网站防护、内网安全、主机安全、数据库安全、邮件安全	大模型服务、AI 智能体、RAG 知识库、模型训练平台、AI 生成内容

如上表 2-1 所示，上述的差异表明了 AI 安全与传统安全的根本性区别。最近诸多的 AI 供应链攻击事件，使得 AI 供应链安全尤为值得警惕。传统软件的供应链相对成熟，行业有成熟的漏洞数据库、依赖分析工具和补丁管理流程。但 AI 供应链要复杂得多，一个 AI 应用可能包含开源模型、第三方大模型 API、MCP 工具链、微调数据集、Skills、A2A、ANP 等多种新型协议或组件，这其中每一个都有可能成为新的攻击入口，而 AI 供应链大多缺乏传统软件供应链那样完善的安全审计机制，这导致安全风险的暴露面进一步增加。

2.1.2 从网络安全到 AI 安全

当我们着手处理这些安全风险，想把过去在传统领域积累的安全能力，迁移到 AI 安全领域时，发现 AI 安全与传统网络安全之间的能力差异同样显著。

下表 2-2 为传统网络安全的核心能力与 AI 安全领域新需求的能力映射，由此对比可见一斑：

表 2-2 传统网络安全与 AI 安全领域能力映射关系

序号	传统网络安全	AI 安全
1	网络边界防御	模型边界防御
2	身份认证与授权	IAM+Agent/MCP 身份认证+工具权限控制+命令行权限控制
3	数据加密	训练数据保护 + 输出过滤
4	漏洞管理	模型漏洞管理 + 对抗鲁棒性
5	安全审计	AI 行为审计 + 可解释性
6	合规检查	AI 专项合规 + 伦理审查
7	应急响应	模型回滚 + 输出熔断 + 舆情管理

从这张映射表可以看出，AI 安全是在传统网络安全之上，结合 AI 新技术的应用，做的**扩展和深化**。一个网络安全工程师，即使再优秀，他也不会自动具备 AI 安全的能力，也仍然需要学习才能具备 AI 新技术的能力。

而 AI 安全治理，需要的恰恰是这样的全新的复合型人才，它既要有扎实的 AI 技术理解，又要有深厚的安全专业功底。

这正是 AI 安全官这一角色存在的根本原因。

2.2 AI 安全官角色定位

既然如此，那么 AI 安全官究竟是一个什么样的角色？

简单来说，AI 安全官是企业中**专门负责 AI 系统安全治理、风险管理和合规保障的高级管理/技术角色**。它包含了五个相互关联的核心职责：

1）**建体系**是 AI 安全官的首要职责，也是其他一切工作的基础。它要求 AI 安全官在企业内部建立起完整的 AI 安全治理框架，从组织架构到管理制度，从流程规范到度量指标。没有这个基础，后续的风险管理、合规保障、攻击防御都将是空中楼阁。

2）**控风险**要求 AI 安全官能够识别、评估和控制 AI 系统全生命周期中的各类风险。这不是一次性的评估工作，而是贯穿 AI 系统从需求定义、模型开发、测试验证到上线部署、持续运行的全过程。每一个阶段都有其特有的风险类型和控制措施。

3）**保合规**在当下的监管环境中尤为重要，特别是国内大模型厂商，《生成式人工智能服务管理暂行办法》《算法推荐管理规定》《深度合成管理规定》等法规提出了明确的合规要求，包括算法备案、生成式人工智能服务备案或应用/功能登记、安全评估以及生成合成内容标识等。AI 安全官需要确保企业的 AI 应用，在法规框架内安全稳定地运行。

4）**防攻击**要求 AI 安全官具备深厚的 AI 安全攻防技术能力。在面对提示词注入、越狱攻击、数据投毒、模型窃取等新型攻击手段，AI 安全官不仅要能够识别威胁，还要能够构建有效的防御体系。

5）**促发展**是体现 AI 安全官角色的战略价值。AI 安全的外延很广，AI 安全官在业务发展的护航的同时，要站在战略角度考虑科学伦理、生命安全、算法公平等方面，在安全与创新之间找到平衡点。

2.3 AI 安全官/CISO/CSO 及周边关系

在实际工作中，AI 安全官这一角色还没广泛地被使用，它的身份定位与企业的已有的安全管理人员，比如企业信息安全负责人（CSO 或 CISO）之间的关系可能会被误解，有人认为 AI 安全官是他们的下属，也有人认为 AI 安全官将来会取代他们。

站在当下 AI 安全实践和未来发展趋势的角度看，这两种理解都是片面的。下表 2-3 是它们两者在职责范围、技术深度和组织关系上的对比，由此可见它们之间存在清晰的分工：

表 2-3 AI 安全官与企业信息安全负责人的关系

维度	企业信息安全负责人	AI 安全官
核心关注	企业整体信息安全	AI 领域安全
技术深度	网络、系统、应用安全	模型、数据、推理、Agent 安全
法规焦点	传统安全与数据保护相关法律法规	AI 专项法规 + 传统法规中的 AI 条款
风险类型	漏洞、入侵、泄露	对抗攻击、数据投毒、输出失控、伦理风险
组织关系	向 CEO/CTO 汇报	可独立设置，也可作为 CISO 下设分支
协作对象	IT、研发、法务	AI 团队、数据团队、产品经理、AI 治理委员会

在实际案例中，AI 安全官与企业信息安全负责人之间是强协作关系，他们的协同关系强依赖于企业应用 AI 技术的规模和深度，这在后续 AI 安全组织架构章节将展开详细描述。总的来说，AI 安全官专注于 AI 特有的风险识别和治理，企业信息安全负责人统筹企业整体安全架构。两者是互帮互助、互相补充，共同推进企业安全和企业 AI 治理工作的开展。

而在更大的视野中，AI 安全官是 AI 治理体系中的一个关键角色。AI 治理涵盖公平性、透明度、问责制、伦理等多个维度，包含多个角色，通常是虚拟型的组织来承接治理工作，AI 安全官和企业信息安全负责人都属于其中一员，AI 安全官参与其中，聚焦于 AI 安全维度，与公平性专家、伦理委员、合规团队等其他人员一起，协同协作，共同推进和完善 AI 治理工作。

2.4 AI 安全官工作重心：AI Security & AI Safety

在安全领域，Security 和 Safety 这两个词在中文中都被翻译为**安全**，它们关注重点不同，前者侧重于信息化层面的安全，后者侧重于物理层面的安全。

在 AI 技术中，**AI Security 与 AI Safety** 之间存在大量交叉和相互转化。理解这一点，对于 AI 安全官准确识别和管理 AI 风险至关重要。

AI Security 侧重于信息安全层面，关注的是对抗性威胁，即恶意攻击 AI 系统的行为。它回答的核心问题是：如果有人想破坏 AI 系统，他们是怎么攻击的？

典型的 Security 威胁包括提示词注入、模型窃取、数据投毒、对抗样本攻击、越狱攻击等。这些问题，通常更强调恶意利用、越权访问、系统完整性和对抗性威胁，但并不意味着所有 Security 风险都必须存在可识别的外部攻击者。

AI Safety 则侧重于功能安全/对齐安全， 关注的是非对抗性风险，即 AI 系统在没有恶意攻击者的情况下，是否产生不可预测、不一致或有害的行为。它回答的核心问题是：即使没有人攻击，AI 系统自己会出问题吗？

典型的 Safety 问题包括模型幻觉、对齐失败、分布偏移下的性能退化、偏见放大、以及高风险场景下的错误决策。这些问题，通常是 AI 系统在正常运行条件下，可能出现的失效、伤害、失控和不符合预期的行为，导致错误发生的原因可能是模型局限，也可能是数据、环境、人机交互设计不足等。如下表 2-4，详细列出了它们之间的区别和差异点。

表 2-4 AI Security 与 AI Safety 之间区别和差异点

维度	AI Security	AI Safety
关注焦点	对抗性威胁	非对抗性风险
核心问题	有人想攻击系统吗？	系统自己会出问题吗？
威胁来源	外部攻击者（恶意意图）	模型局限性、数据偏差、环境变化
典型威胁	提示词注入、模型窃取、数据投毒	幻觉、对齐失败、偏见放大
防御手段	红队测试、安全护栏、访问控制	安全对齐、持续监控、人工审核
评估方法	对抗测试、漏洞扫描	安全评测基准、行为测试、A/B 测试
责任主体	AI 安全官（Security 维度）	AI 安全官 + AI 伦理团队

在实际的 AI 安全治理过程中，AI Security 与 AI Safety 之间既有区别也存在关联，有时甚至难以区分，杂糅在一起。比如，一个模型的对齐失败，看上去是 Safety 问题，但根本原因可能被攻击者实施了越狱攻击才导致的问题发生，其实它是一个 Security 问题。反之，提示词注入攻击是 Security 问题，但注入成功后可能导致模型产生有害输出内容，表现出来的是 Safety 问题。作为管理者，AI 安全官需要同时关注这两个维度，并理解它们之间的相互作用。

实际情况是，现行基模企业 AI 安全管理人员职责主要覆盖 Security 维度，同时需要与 AI 伦理/对齐团队协作，覆盖 Safety 维度。如果企业没有 AI 伦理/对齐团队，则 Safety 维度的职能通常也会落到 AI 安全管理人员头上。本白皮书建议，在未设专门 AI 伦理/对齐团队时，应由管理层明确 Safety 相关风险的责任归属，由业务、研发、安全及法务共同承担相应职责。涉及医疗、自动驾驶等场景时，应引入具备相应专业能力的负责人。

一个成熟的 AI 安全治理体系，需要同时包含 Security 和 Safety 两个方面的控制措施，缺一不可。因此，AI 安全官 在构建 AI 安全治理团队时，不仅需要招募具备传统红蓝对抗经验的 Security 专家，也需要引入懂得数据伦理、偏见评估、大模型微调对齐的 Safety 与合规专家，两者不可偏废。

2.5 AI 安全威胁

AI 安全官的核心工作内容是识别威胁防范风险，为了统一威胁描述，映射处置威胁所需要的专业能力，本白皮书基于 OWASP LLM Top 10、MITRE ATLAS 和 NIST AI RMF 等国际主流框架，建立了一套适用于 AI 安全官工作的 AI 安全威胁分类，帮助 AI 安全从业者建立系统化的威胁认知框架。

2.5.1 威胁分类矩阵

按照 AI 应用分层对风险进行划分，分层后的 AI 安全威胁与 AI 安全官所具备的能力映射关系，详细如下表 2-5 所示：

表 2-5 AI 安全威胁与 AI 安全官所具备的能力映射关系

威胁类别	典型威胁	攻击层	严重度	对应 AICSO 能力
输入层攻击	提示词注入（直接/间接/跨上下文）	输入层	高	C02
	越狱攻击（角色扮演/编码绕过/多轮攻击）	输入层	高	C02
	对抗样本攻击	输入层	中	C01，C02
数据层攻击	训练数据投毒	数据层	极高	C01
	数据提取攻击	数据层	高	C01
	隐私推断攻击	数据层	中	C01

模型层攻击	模型窃取	模型层	极高	C01, C05
	模型逆向攻击	模型层	高	C01
	模型后门植入	模型层	极高	C01, C05
输出层攻击	有害内容生成	输出层	高	C02, C09
	敏感信息泄露	输出层	高	C01, C06
	不安全代码生成	输出层	高	C02, C06
Agent 层攻击	Tool-call 注入	Agent 层	高	C02, C05
	Agent 自主性升级	Agent 层	极高	C02
	记忆投毒	Agent 层	高	C01, C02
	Agent 逃逸与劫持	Agent 层	极高	C02
供应链攻击	恶意开源模型	供应链	极高	C05
	恶意 MCP Server	供应链	高	C05
	恶意训练数据集	供应链	高	C01, C05
	推理 API 劫持	供应链	高	C05
运营层攻击	AI 系统滥用（社会工程/欺诈辅助）	运营层	中	C06, C09
	Shadow AI	运营层	高	C07
	模型服务拒绝（Model DoS）	运营层	中	C04, C06

上述表中的分类，便于威胁发生时快速识别攻击发生在哪一层，以找到对应的 AI 安全专业能力。比如，该问题的解决是应该在算法与数据安全上加大投入，还是在 AI 供应链安全上加大投入，都可以从上表中找到对应能力项。

但考虑到当前阶段，全球性的 AI 技术爆发，新技术新产品不断涌现，对于复杂 AI Agent 系统的问题处置，建议在具体威胁建模时增加资产×生命周期×信任边界×攻击面四维视角，方便在评估 Agent 风险时，检查是什么身份在运行、凭据在哪里、能够横向访问什么、是否能主动出网、是否能修改自己的日志、是否存在不可变审计，以及异常情况下是否可以立即切断执行权限等。如下表 2-6 所示。

表 2-6 资产×生命周期×信任边界×攻击面四维视角

维度	典型分析对象	关键问题
资产	数据、模型、Prompt、RAG 知识库、Agent、工具、凭据、日志	什么身份在运行？凭据在哪里？哪些资产一旦被篡改、泄露或滥用会造成实质影响？
生命周期	数据准备、训练/微调、评测、部署、运行、退役	风险在哪个阶段产生，在哪个阶段可以被发现和阻断？
信任边界	用户与模型、模型与 RAG、Agen 与工具、企业与供应商、Agent 与 Agent 之间	数据、指令和权限跨越了哪些主体或系统边界？是否信任链断裂？是否能跳出边界出网？是否能修改自己的日志？是否存在不可变审计？异常情况下是否可以立即切断执行权限？
攻击面	输入、输出、API、身份、记忆、工具、编排、供应链、基础设施	攻击者实际从哪里进入，并形成影响到哪些链路？

2.5.2 新兴威胁趋势

AI 技术发展日新月异，不断演进的新威胁趋势，是值得每一个 AI 安全官密切关注的内容，尤其是下面的 4 类趋势：

1）多模态攻击面扩展

随着 GPT-5.6 Sol、Gemini 3.8 等多模态模型的普及，攻击面从纯文本扩展到图像、音频、视频等多个模态。图像中的对抗性扰动、音频中的隐写指令、视频中的对抗样本等新型攻击手法正在涌现，多模态输入扩大了不可信内容进入系统的路径。企业应持续核查模型输入、模型来源、许可证及适用监管要求，在防御体系中构建跨模态的安全分析能力。

2）RAG 安全风险

检索增强生成 RAG 是当前企业 AI 应用最广泛的架构模式之一，但它引入了特有的安全风险，比如检索内容投毒、知识库篡改、检索结果操纵等。攻击者可能通过在公开数据源中植入特定内容，使其在 RAG 检索时被选中，从而间接影响 AI 系统的输出。另外，Embedding 模型与向量索引的可信性、向量数据库访问控制、跨租户向量泄露、检索结果权限继承等，都是 RAG 安全需要保护的内容。

3）AI 生态系统供应链风险

开源模型生态的快速发展，比如 HuggingFace、ModelScope 社区的兴起，开源模型的数量在不断增加。同时，开源模型的能力也越来越强，用户使用也越来越多。再叠加开源生态的监管弱于商业化产品，它带来了类似传统软件供应链风险越来越凸显。模型文件篡改、

许可证陷阱、隐藏后门等问题急需治理，AI SBOM 实践正成为 AI 供应链安全管理的基础能力。

4）企业级 AI 安全法规加速

全球范围内 AI 安全法规正在加速出台，《欧盟人工智能法案》已经生效，中国的 AI 监管法规体系日趋完善，美国的 AI 行政命令也在推进中。AI 安全官需要建立持续的法规跟踪机制，确保企业的 AI 安全实践与法规要求保持同步。这要求 AI 安全官的工作重心，必须从单纯的技术防御向合规治理纵深发展，通过建立持续的合规审查、合规审计与可解释性评估机制，确保企业 AI 应用生命周期与全球多地法域的要求保持动态同步，防范高额的违规罚金与声誉风险

第 3 章 AI 安全官 4-12-5 能力模型

3.1 能力模型设计原则

在设计 AI 安全官能力模型时，我们遵循了下面的 6 项核心原则：

- **能力导向** 模型以实际工作中的能力需求为驱动，相比于关心 AI 安全官用的机器学习框架是 PyTorch 还是 TensorFlow，我们更聚焦于其能否在模型训练等具体场景下，设计并落地端到端安全方案的实战能力。
- **可度量性** 模型对每项能力的等级描述与评估标准均有明确定义，提供结构化的量化指标，杜绝模棱两可的模糊表述。
- **循序渐进** AI 安全官的五个职级之间是互相递进的，有清晰的职业成长路径，帮助从业者明确自身定位、明晰晋升目标与能力迭代路径，知道自己在哪里、要去哪里、怎么去。
- **注重实用性** 模型面向企业，在设计时要考虑模型的应用落地，可以直接用于岗位设计、团队搭建和能力评估。
- **保持前瞻性** 当前 AI 技术发展太快，在覆盖当前需求的同时，要保持一定的前瞻性，为 AI 技术的快速演进预留定义和承接空间。
- **可衍生性** 模型设计完成之后，需要天然适配 AI 技术路线的发展。新出现的 Benchmark 或新产品形态，要能与 AI 生态形成强有力地贯穿，保障个人职业参照与企业技术演进的长期协同。

3.2 四大能力域

本白皮书将 AI 安全官的能力划分为四大能力域，分别为 **AI 工程与架构安全、AI 治理与风险合规、AI 技术与攻防安全、业务与组织领导力** ，如下图 3-1 所示。

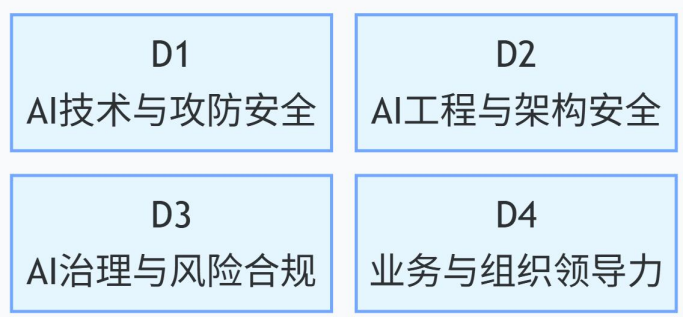


图 3-1 AI 安全官四大能力域

AI 技术与攻防安全能力域是 **AI 安全管理措施的落地基础**，没有扎实的攻防能力支撑，一切管理措施都是空中楼阁；AI 工程与架构安全能力域是 **AI 安全官技术能力的底座**，它们回答的是能不能做，怎么做更好的问题，没有工程能力和架构安全保底，安全工作是难以落实的；AI 治理与风险合规能力域是 **AI 安全官的治理能力框架**，它回答的是是否合规、AI 治理体系是否完备以及如何把握 AI 治理成熟度的问题，没有制度保障，再强的技术能力也无法规模化；业务与组织领导力能力域是 **AI 安全官的组织能力呈现**，它回答的是 AI 安全工作如何推动、AI 安全如何为业务创造价值的问题，没有组织层面的保障、沟通机制、培训赋能和事件管理，AI 治理体系就无法真正落地。

3.2.1 能力域 D1：AI 技术与攻防安全

能力域 D1 是 AI 安全的技术基础，也是大多数人看到 AI 安全时首先想到的领域，它主要是 AI 系统面临的直接技术威胁的识别、防御和红队能力，包含 C01 算法与数据安全、C02 AI 对抗防御、C03 AI 红队测试 3 项核心能力。

C01 算法与数据安全 是整个能力体系的基石之一，它关注的是 AI 全生命周期中数据和算法的保护，主要是隐私保护技术、模型水印技术和 AI Agent 新技术方向保护等。**C02 AI 对抗防御** 关注的是针对 AI 系统的主动攻击防御，提示词注入、越狱攻击、安全护栏技术以及 Agent 安全，包括权限控制、工具调用安全和逃逸防御等。**C03 AI 红队测试**是通过系统化的攻击测试，发现 AI 系统的安全弱点，是构建有效防御体系的前提，涵盖了从组织管理到红队工具使用的完整方法论。

3.2.2 能力域 D2：AI 工程与架构安全

如果说能力域 D1 关注的是攻击与防御，那么能力域 D2 关注的就是基础设施的韧性，它主要关注支撑 AI 系统运行的底层环境的安全保障，包含 **C04 AI 基础设施安全**、**C05 AI 供应链安全**、**C06 AI 安全运营** 3 项核心能力。

C04 AI 基础设施安全 涉及 GPU 集群安全、云原生环境、容器化安全等。这些基础设施层面的安全控制，是上层 AI 应用安全的物理基础。**C05 AI 供应链安全**在 AI 时代变得尤为复杂，开源模型、第三方推理 API、MCP 工具链和微调数据集、供应商安全审计、MCP Server 安全评估、AI SBOM 管理，这些构成了供应链安全的核心能力。**C06 AI 安全运营**是持续安全的保障，从日志采集和审计追踪，到模型行为异常检测，甚至紧急回滚机制，再到 AI SOC 的建设和度量指标运营，安全运营是 AI 安全治理的最后一公里，通过持续改进不断提高企业的 AI 安全水平。

3.2.3 能力域 D3：AI 治理与风险合规

能力域 D3 从技术层面转向制度层面，确保 AI 应用在合规框架内安全运行，包含 **C07 AI 治理体系建设**、**C08 AI 法规与合规**、**C09 AI 伦理与知识产权** 3 项核心能力。

C07 AI 治理体系建设关注企业 AI 治理的顶层设计，AI 治理委员会、管理制度、AI 系统全生命周期的治理流程，是将 AI 安全从个人能力转化为组织能力的关键环节。**C08 AI 法规与合规**要求从业者需要熟悉中国 AI 服务相关法律法规要求，同时了解《欧盟人工智能法案》和 ISO/IEC 42001 等国际标准的要求。**C09 AI 伦理与知识产权**关注偏见检测、公平性审计、AI 内容版权、责任分配等问题。

3.2.4 能力域 D4：业务与组织领导力

能力域 D4 超越了技术和制度层面，关注的是 AI 安全的组织推动力，包含 **C10 AI 业务风险评估**、**C11 AI 安全事件管理**、**C12 安全沟通与培训** 3 项核心能力。

C10 AI 业务风险评估要求从业者能够将技术风险翻译为业务语言，通过风险量化和业务影响分析，为企业管理层提供决策依据。**C11 AI 安全事件管理**是事件发生时快速响应的保障机制，包含应急预案设计、模型回滚、舆情管理、事后复盘等。**C12 安全沟通与培训**也是 AI 安全官工作的一部分，从技术团队到管理层，从安全竞赛到全员培训，逐步形成体系化安全培训与沟通机制。

3.3 十二项核心能力

在四大能力域的基础上，每个域下设 3 项核心能力，形成完整的十二项能力矩阵。这十二项能力清单，构成一个相互关联的能力生态，描述清楚 AI 安全官的能力水平和发展趋势的 12 个方面，像一系列支柱，支撑和帮助 AI 安全官走向职业成功。

这 12 项能力，任何一项能力的缺失，可能会出现短板效应，在其他能力上产生连锁反应。它们的详细信息，如下表 3-1 所示：

表 3-1 AI 安全官十二项核心能力

能力域	编号	核心能力	英文名称	简述
D1 AI 技术与攻防安全	C01	算法与数据安全	Algorithm & Data Security	保护 AI 模型训练和推理过程中的算法安全与数据安全，保护 AI 应用，如 Agent 上下文、RAG 数据、长期/短期记忆、用户交互数据和向量数据的安全性
	C02	AI 对抗防御	AI Adversarial Defense	防御针对 AI 模型的对抗攻击、越狱、注入等威胁，防御针对 AI 应用的攻击，如 Agent 劫持、间接提示词注入、工具滥用、身份与权限滥用、越权等
	C03	AI 红队测试	AI Red Teaming	通过主动攻击测试的方式，发现 AI 系统的安全弱点，同时完成红队技术管理工作、前沿红队技术研究，比如 Agent 目标、规划、工具、记忆、身份、跨 Agent 通信、失败级联和人工确认绕过等
D2 AI 工程与架构安全	C04	AI 基础设施安全	AI Infrastructure Security	保护 GPU 集群、推理服务、API 等 AI 基础设施的安全，约等于传统基础设施架构安全和 AI 基模底层架构、AI 底层服务的安全性
	C05	AI 供应链安全	AI Supply Chain Security	管控 AI 开源模型、第三方服务、MCP 工具链等供应链风险，管理模型的使用、接入以及插件/MCP Server/A2A 组件/依赖包/第三方 Agent 服务/skills 等完整链路的安全性
	C06	AI 安全运营	AI Security Operations	AI 系统的持续监控、告警、审计与事件响应，记录 Agent 和工具调用链、授权、目录读写、日志和运行时异常等，在可观测数据之上完成 AI 安全运营策略。
D3 AI 治理与风险合规	C07	AI 治理体系建设	AI Governance System Building	建立企业 AI 治理组织、制度和流程，从组织和资源层面，保障 AI 安全工作的落实
	C08	AI 法规与合规	AI Regulation & Compliance	确保 AI 应用符合国内外法规和行业标准，主要是各个国家的上位法、行业规范、技术标准、管理要求等
	C09	AI 伦理与知识产权	AI Ethics & IP	处理 AI 公平性、可解释性、版权和责任归属问题
D4 业务与组织领导力	C10	AI 业务风险评估	AI Business Risk Assessment	识别和评估 AI 应用场景中的业务风险，按自治程度、可执行权限、数据敏感度、外部影响、可逆性和人工监督程度，对 AI 应用或 AI Agent 风险进行风险评估和分级量化，为风险处置提供依据

	C11	AI 安全事件管理	AI Security Incident Management	AI 安全事件的应急预案、响应处置和事后改进，比如建立智能体紧急停止开关、凭据吊销、工具隔离、记忆隔离、任务中止、应急切换、证据保全机制
	C12	安全沟通与培训	Security Communication & Training	跨部门沟通、安全文化建设与全员培训

第 4 章 AI 安全官职业等级

4.1 AI 安全官五级划分

能力模型告诉我们 AI 安全官需要具备什么能力，但一个人的能力本身有高低之分。一名刚进入 AI 安全领域的从业者和一个承担企业级 AI 安全治理责任的负责人，可能都在参与 AI 安全工作，但他们能够解决的问题、承担的责任和影响范围，是完全不同的。

本白皮书采用 5 级分级模型，将 AI 安全从业者的个人职业能力，划分为 5 个等级，采用互相递进的关系，形成一条清晰的职业成长路径。如图 4-1 所示：



图 4-1 AI 安全官五个等级之间关系

在每一个等级中，有明确的能力门槛、典型岗位、经验要求和认证建议，让从业者能够清楚地知道我现在在哪里、我将要去哪里、我需要做什么才能到达那里。

5 级划分的详情，如下表 4-1 所示：

表 4-1 AI 安全官五级职业等级划分

等级	岗位名称	核心定义
L1	AI 安全助理工程师	具备 AI 安全基础认知，并能在指导或既定流程下完成基础检查、测试、配置和报告任务
L2	AI 安全工程师	能针对具体系统设计和实施 AI 安全方案，解决复杂技术问题
L3	AI 安全架构师	能设计企业级 AI 安全架构、标准和工程体系，统筹跨系统技术风险

L4	AI 安全总监	能领导企业 AI 安全团队与治理体系，将技术、合规和业务风险统一纳入管理决策
L5	AI 安全官	能制定企业级 AI 安全战略，参与最高管理层、监管及重大风险决策，并对企业 AI 安全承担最高层级责任

如上表 4-1 所示，每一个等级对应的中英文名称、核心定义，明确了当前等级的核心能力和工作职责。这样的划分，一目了然，既降低了入门层级的复杂度，使企业招聘、人才评估和个人发展更容易使用，也保留了从工程实践、架构设计到企业治理与战略领导的完整成长路径。

4.2 等级特征对比

从每一个等级的核心定义可以看出，从 L1 到 L5，除了技术深度，还有工作性质、影响范围、决策责任和可验证成果的逐级扩展。L1 属于入门级，承接基础认知和标准化实践，重点是能否按照既定流程完成基础工作的能力。从 L2 开始，个人能力逐步向方案设计、架构建设、治理领导和战略责任过渡，直到具备 L5 等级的职业能力，成为真正的行业专家或领袖。

下表 4-2 从工作内容、影响范围、决策层级、外部协作、举证材料类型、工作经验 6 个方面，列出了 L1~L5 的核心差异点。

表 4-2 L1~L5 等级特征对比

维度	L1	L2	L3	L4	L5
工作内容	认知与标准化执行	方案设计与实施	架构与标准设计	治理与组织领导	企业战略与最高层级责任
影响范围	个人/项目基础任务	系统/平台	企业技术线/多业务	企业全组织	集团/生态及关键外部利益相关方
决策层级	标准任务执行决策	技术决策/方案决策	架构与技术标准决策	管理、资源和风险接受决策	企业战略、重大风险和外部监管决策
外部协作	偶尔	频繁	跨组织	行业级	生态级
举证材料类型	学习记录、检查记录、测试报告	项目方案、实施证据	架构、标准、平台化成果	治理体系、重大项目和管理决策	企业战略、重大事件、监管/审计/最高管理层证据

维度	L1	L2	L3	L4	L5
建议工作经验	0-3 年相关经验	3-5 年相关经验	5-8 年相关经验	8-12 年相关经验	12 年以上相关经验，重实际能力而非年限

其中，判断一名从业者为什么等级，**工作经验仅作为参考建议**，不能以工作年限作为硬性门槛。等级晋升主要要看实际能力，从 L3 到 L4 的晋升，是从技术执行者到架构设计者的能力转变。从 L4 到 L5 的晋升，是从技术专家到治理领导者的能力转变。**每一次晋升，都要以实际能力为准**，AI 安全属于快速演进领域，前期可以将传统安全、AI 工程、数据治理、风险合规等经验折算为相关能力和工作经验，最终以可验证的工作成果、责任范围和决策影响作为主要证据，**只要能力完成升级，则可以晋升下一个等级**。

4.3 各等级详细描述

L1 AI 安全助理工程师

L1 为 AI 安全助理工程师，是 AI 安全从业基础执行的入门阶段。这个阶段的从业者，应理解提示词注入、越狱逃逸、敏感信息泄露、数据投毒、第三方模型/API 风险、Shadow AI 以及 Agent 过度授权等基本概念，并能够在高等级从业人员的指导下或既定流程下，使用检查表、评测工具和安全配置，完成基础核查、测试、记录与问题升级等工作。

L1 并不要求从业者具备独立设计复杂方案的能力，但要求其能够把基本知识转化为可复核的工作结果。对企业而言，这一级别既适用于初级 AI 安全岗位，也可以作为 AI 开发、安全工程师、产品经理、数据人员等相关角色，作为承担 AI 安全职责时的基础能力门槛。

关键里程碑： 完成 AI 安全基础学习；参与并能够独立完成若干标准化检查或测试任务；能够输出可复核的记录或报告；能够识别超出自身能力边界的问题并正确升级。

L2 AI 安全工程师

L2 对应 AI 安全工程师，是 AI 安全从业者具备**按流程执行能力**进入**针对场景设计方案能力**的分水岭。从业者应能够针对具体业务系统完成 AI 应用威胁建模、数据/RAG 安全方案、对抗防御、Agent 身份与权限设计、红队方案、AI 安全合规项推进、安全监控或事件响应中的一个或多个完整工作流，并能够解释设计取舍和残余风险。

这个等级的核心不在于掌握多少工具，而在于能否把风险转化为控制要求，并推动控制真正落地。L2 通常已经是团队中的技术骨干，需要能够独立处理复杂技术问题，并与模型、研发、数据、运维和合规人员协作完成项目。

关键里程碑： 主导至少 1 个重要 AI 安全项目；能够独立设计并实施 AI 安全方案；能够将风险转化为控制要求并形成实施证据；在能力域 1 或能力域 2 至少一项能力上达到明显技术深度。

L3 AI 安全架构师

L3 对应 AI 安全架构师，负责跨系统、跨团队的 AI 安全架构与工程标准。其工作重点包括企业 AI 安全参考架构、AI-SDL/Agent-SDL 规划设计、安全评测平台、AI 网关/AI 安全护栏、Agent 运行时隔离、监管合规方案、供应链治理、可观测与审计标准，以及与现有 IAM、SOC、DevSecOps 和数据治理体系的集成。

L3 要求从业者从单个项目走向企业级技术体系，作为架构师，不仅需要设计正确，还要让控制可以规模化实施、持续维护和被多个业务复用，因此技术标准、平台化能力和跨团队推动能力都是这一等级的重要组成部分。

关键里程碑： 企业级架构或标准被多个系统采用；推动形成自动化或平台化安全能力；能够主持复杂架构评审和重大技术风险决策；能够建立 AI 与 Agent 安全的统一技术基线；能够全面主持算法备案/大模型备案工作。

L4 AI 安全总监

L4 对应 AI 安全总监，是从技术架构者向企业治理领导者的关键晋升。这个阶段的从业者，已经不仅负责技术体系，还要领导 AI 安全团队，建立治理机制，将 AI 资产、技术风险、合规要求和业务影响统一纳入组织决策，并对风险接受、资源投入、重大项目和事件处置承担管理责任。

L4 需要具备足够的技术判断力，虽然这个阶段从业人员的技术工作不一定是亲自完成的，但能力上必须具备让企业形成持续良好运行的 AI 安全治理能力。其典型职责包括：建立治理委员会或责任机制、制定年度计划和预算、推动高风险 AI 上线评审、组织跨部门整改，以及向管理层说明风险与投入产出。

关键里程碑： 建立或领导企业 AI 安全治理体系；组建或领导 AI 安全团队；对重大 AI 风险、资源投入和高风险上线承担管理决策责任；能够推动技术、业务、法务和合规形成闭环。

L5 AI 安全官

L5 对应 AI 安全官，是 AI 安全官能力框架定义的最高个人能力等级。

这个阶段的从业者需要制定企业级 AI 安全战略，参与董事会或最高管理层的重大决策，并在重大事件、监管检查、外部审计、行业协作等场景中，代表企业说明 AI 安全状况。其影响范围通常超越单一技术线，或业务单元，通常覆盖集团级 AI 战略、重大风险治理和外部利益相关方。

关键里程碑： 制定并推动企业级 AI 安全战略；参与董事会或最高管理层 AI 风险治理；能够处理监管、审计和重大外部沟通；能够组织重大 AI 安全事件或集团级高风险 AI 决策；能够推动行业或生态层面的安全协作。

第 5 章 AI 安全官能力评估

5.1 能力评估方法

4-12-5 能力模型定义了 12 项能力在 5 个职级上的要求，但如何判断一个从业者是否达到了某个等级，仍需要具体的评估方法。

本白皮书推荐采用**多方法的组合评估**，根据不同等级的侧重点，选择不同的评估方式。对于 L1 到 L3 等级，知识测评和实操考核的权重较高；而对于 L4 到 L5 等级，项目评审、案例答辩和同行评议的权重则更大。这样的设计，是从过去的经验来看，我们认为高级别从业人员的能力，很难通过标准化测试来衡量。

如下表 5-1 所示，给出不同等级的评估方法和权重建议。

表 5-1 不同等级 AI 安全官的评估方法和权重建议

评估方法	适用场景	权重建议	说明
知识测评	L1-L3	20-30%	标准化知识考试，覆盖 AI 安全理论和法规
实操考核	L2-L4	25-35%	实际环境中的技术操作和方案设计
项目评审	L3-L5	20-30%	对过往 AI 安全项目的经验评审
案例答辩	L4-L5	15-25%	针对复杂场景的方案阐述和答辩
同行评议	L5	10-15%	行业专家或企业高管的综合评价

具体到 12 项能力在 L1~L5 不同等级上的水平判断，5.2 章节会给出详细的能力评估细则。企业在制定 AI 安全人员的职级职等标准时，可以参考此白皮书的评估方法，通过组合评估，力求公平、公正地评估每一个从业者的真实水平。

5.2 能力评估细则

C01 算法与数据安全

此项能力关注的是 AI 系统全生命周期中数据和算法的保护，从训练数据的采集、清洗、标注到存储的每一个环节，都有别于传统数据安全的特殊要求，尤其是在涉及用户隐私和商业机密的场景中，数据安全的落地措施更是不可或缺。

AI 语境下的数据分级分类，主要是训练数据、推理数据、模型权重数据，各有不同的敏感度和保护需求。数据脱敏与匿名化技术、数据保留与销毁策略，需要针对 AI 系统的特殊性进行安全设计。差分隐私、联邦学习、同态加密和安全多方计算等技术，从业者需要理解其原理并能在 AI 应用场景中做出合理的技术选型。

数据篡改与污染检测、数据来源验证与溯源、数据质量监控，是构成防御数据投毒的三道防线，从业者需要掌握此类防护技术。在模型安全方面，模型窃取、模型逆向攻击等常见威胁，要求从业者掌握模型水印技术和权重保护等防御手段，确保核心模型资产不被窃取或逆向工程。

最新的 Agent 技术，引入长期记忆、外部知识库和持续上下文，导致数据安全边界也从训练数据和推理输入输出，进一步扩展到记忆数据、RAG 语料、向量索引和 Agent 运行过程中的中间状态，这要求从业者掌握数据安全边界扩展后的安全保护技术。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-2 所示：

表 5-2 算法与数据安全 L1~L5 等级划分评估细则

等级	能力要求
L1	了解数据生命周期各阶段的安全风险；知道基本的隐私保护概念；能执行数据脱敏操作；能识别常见数据投毒模式；能使用基础隐私保护工具
L2	能设计算法和数据安全技术方案；能实施差分隐私或联邦学习方案；能建立数据质量监控体系
L3	能设计企业级算法和数据安全架构；能评估和选型隐私保护技术栈；能建立模型安全防护体系
L4	能制定企业 AI 算法和数据安全策略；能推动跨部门算法和数据安全治理；能评估算法和数据安全投资 ROI
L5	能制定企业级 AI 算法和数据安全战略；能参与算法和行业数据安全标准制定；能在监管机构前代表企业陈述算法和数据安全实践

C02 AI 对抗防御

AI 对抗防御是当前 AI 安全领域最活跃、也是攻击手法更新最快的能力方向，它要求从业者能够识别、防御和响应针对 AI 系统的各类对抗性攻击，并针对这些攻击方式，建立系统化的防御体系。

提示词注入是目前 AI 系统最普遍的威胁之一，攻击者通过精心构造的提示词，来操纵 AI 系统的行为。这种攻击又分为直接注入、间接注入、跨上下文注入多种形式。直接注入由用户在输入中直接嵌入恶意指令，间接注入则通过外部数据源（如网页、文档）间接传递恶意提示，跨上下文注入则利用多轮对话或外部工具调用的上下文切换来实施攻击。防御提示词注入，需要从输入清洗、意图验证到多层过滤的完整策略。

越狱攻击是另一类棘手的攻击方式，攻击者通过角色扮演、编码绕过、多轮攻击等技术手段，试图绕过模型内置的安全对齐机制，比如宪法式 AI、RLHF 安全对齐等，防御者需要持续跟踪最新的攻击模式，并更新检测方法。与此同时，模型滥用问题也不容忽视，有害内容生成、恶意代码生成、社会工程辅助等滥用场景，要求防御者建立完善的滥用检测与审计机制。

面对上述这些威胁，安全护栏框架提供了系统化的解决方案。从输入过滤到输出过滤，再到对话策略控制，NeMo Guardrails、LLM Guard 等开源框架，为企业提供了可落地的防护手段。

随着 AI Agent 的兴起，Agent 安全，包括权限控制、工具调用安全、Agent 行为审计和 Agent 逃逸防御，也成为了这一能力域不可忽视的组成部分。在 Agent 场景下，对抗防御要覆盖目标劫持、间接提示词注入、过度自治、工具滥用和权限升级等风险，防御重点从过滤文本扩展到约束行为。与此同时，企业还需要将系统提示词泄露、输出处理不当、过度自主权以及无边界资源消耗纳入统一的防御视角。除了 AI 安全护栏外，需要更加全面的 AI IAM 之类集用户、角色、权限、工具、行为、审计等于一体的安全防护手段。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-3 所示：

表 5-3 AI 对抗防御 L1~L5 等级划分评估细则

等级	能力要求
L1	了解提示词注入和越狱攻击的基本概念；知道安全护栏的作用；能识别常见注入和越狱模式；能配置基础输入/输出过滤规则；能使用安全护栏工具
L2	能设计多层防御方案；能针对特定业务场景定制安全策略；能分析和拆解新型攻击手法
L3	能设计企业级 AI 安全护栏架构；能建立攻击分类和响应体系；能评估模型安全对齐效果
L4	能制定企业 AI 对抗防御策略；能组建和管理 AI 安全防御团队；能在安全、用户体验与业务效率之间

	取得平衡；能推动智能体对抗防御常态化
L5	能制定企业级 AI 对抗安全战略；能推动行业对抗防御标准；能参与重大智能体安全风险治理与外部协作

C03 AI 红队测试

红队测试体现了现代安全治理中以攻促防的理念，只有真正理解攻击者是如何思考、如何行动，才能构建出有效的防御体系。在 AI 安全领域，红队测试通过系统化的攻击测试，来发现 AI 系统的安全弱点，是构建有效防御体系的前提和基础。

在组织层面，AI 红队需要明确红队的职责边界、测试范围界定和测试用例设计方法。Garak、PyRIT、promptfoo 等自动化红队工具，可以极大提升测试效率，但手动红队与自动化红队的结合才是最佳实践，使用自动化工具覆盖广度，采用手动测试控制挖掘深度，相互结合，提高红队测试效果。

安全评测体系是红队测试的一个重要维度，业界通用 LLM 安全评测基准（如 HELM Safety、TruthfulQA、BBQ 等）为模型安全性提供了标准化的衡量尺度，结合多模态模型和 AI Agent 的普及，评测体系也要与时俱进，需要进一步覆盖工具调用、权限链、长期记忆、RAG 来源、跨 Agent 协作和高风险动作确认机制，测试对象也从模型输出扩展为完整执行轨迹。从业者需要具备设计自定义评测框架的能力，以应对特定业务场景的安全评估需求。

红队测试的价值不仅在于发现问题，更在于推动持续改进。漏洞分级与量化、风险发现后的评估与跟踪、安全评测报告的编写，以及与开发团队的协作修复，构成了完整的闭环流程。蓝红对抗设计、演练场景构建、演练复盘与改进，则是让红队测试找到持续存在的高价值场景。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-4 所示：

表 5-4 AI 红队测试 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 红队测试的概念和基本流程；知道常用红队工具名称；能在指导下执行基础红队测试用例；能使用 Garak 等工具进行扫描；能记录和报告发现

L2	能独立设计红队测试方案；能编写定制化测试用例；能分析测试结果并提出修复建议
L3	能设计企业级 AI 安全评测体系；能领导红队团队；能建立自动化安全评测流水线
L4	能制定企业 AI 红队策略和预算；能推动红蓝对抗常态化；能将模型、RAG、Agent 和多模态红队成果转化为防御改进
L5	能推动行业级 AI 安全评测标准；能参与国际 AI 安全评估合作；能建立跨组织联防联测和智能体安全评测机制

C04 AI 基础设施安全

C04 作为能力域 2 的第一项能力，涉及的是支撑 AI 系统运行的底层物理和虚拟环境的安全保障。

GPU 集群安全是这一领域区别传统安全最具 AI 特色的部分，GPU/TPU 资源隔离、显存安全、CUDA/ROCm 安全配置等，传统安全工具无法覆盖的领域，需要专门的技术能力。分布式训练中的通信安全、GPU 侧信道攻击防御等新兴威胁，更是传统 IT 安全团队鲜有涉猎的盲区。

云原生环境为 AI 工作负载提供了弹性部署的能力，但也引入了新的安全挑战。AI 云平台的安全配置、弹性推理集群的安全管理、Serverless AI 服务的安全保障，以及多云和混合云 AI 部署场景下的数据安全，都需要专门的安全策略。容器化部署则带来了镜像安全、运行时安全、k8s AI 工作负载保护等问题，镜像供应链安全尤其值得关注。

推理 API 是 AI 系统与外部世界交互的主要接口，其安全保障至关重要。认证与授权、速率限制与防滥用、输入验证与输出安全检查、版本管理与废弃策略，构成了推理 API 安全防护的基本要素。

随着企业从单模型调用走向多模型路由和 Agent 运行时，基础设施安全还需要覆盖 AI 网关、模型路由策略、多租户推理隔离、上下文与缓存数据保护、凭据管理以及 TEE/机密计算等能力。尤其在共享 GPU、共享推理集群或跨云部署场景中，应明确租户边界、缓存边界和模型上下文边界，避免传统网络隔离无法覆盖的推理侧数据泄露。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-5 所示：

表 5-5 AI 基础设施安全 L1~L5 等级划分评估细则

等级	能力要求

L1	了解 AI 基础设施的基本组成；知道 GPU 集群、容器、API 的基本安全概念；能执行 AI 基础设施安全基线检查；能配置基础 API 安全策略；能识别常见配置错误
L2	能设计 AI 基础设施安全方案；能实施容器和 Kubernetes 安全策略；能建立 API 安全防护体系
L3	能设计企业级 AI 基础设施安全架构；能制定 GPU 集群安全策略；能建立 AI 云安全标准
L4	能统筹企业 AI 基础设施安全；能推动 DevSecOps/AI-SDL 落地；能评估多租户推理、AI 网关、模型路由和机密计算等基础设施风险
L5	能制定企业 AI 基础设施安全战略；能参与 AI 基础设施安全标准制定；能推动行业级安全实践与关键基础设施风险治理

C05 AI 供应链安全

AI 供应链安全在 AI 时代变得尤为复杂，其复杂程度远超传统软件供应链。

开源模型是供应链中最受关注的环节，模型文件完整性验证、模型许可证合规检查、预留后门检测、模型来源可信度验证，构成了开源模型安全评估的基本要求。

第三方大模型 API 的安全评估、供应商安全审计、SLA 与安全条款的审视、数据传输安全保障，以及供应商集中度风险的管理，是供应链安全的重要组成部分。Agent、MCP 和 Skills 生态的兴起，也带来了新的挑战，比如 MCP Server 的安全评估、工具调用权限控制、工具链组合风险分析、动态工具注册安全等，这些在传统软件供应链中不存在的问题，需要纳入安全管控中。

从过去两年的安全事件来看，在更广泛的意义上，AI 框架和类库（如 PyTorch、TensorFlow）的安全管理、Python 包安全、依赖关系分析、AI SBOM 的应用、漏洞追踪与修复流程，也是 AI 软件供应链治理的完整图景的一部分。如果企业安全能力充足，建议进一步把 SBOM 扩展为 AI BOM。既要包括 Model BOM（基础模型、微调模型、Adapter/LoRA、Embedding 模型及版本来源），也包括 Data BOM（训练、微调、评测和 RAG 数据源），还应记录系统提示词、Agent 定义、工具/MCP Server、安全护栏等运行依赖。

添加 AI BOM 的目的是让企业在漏洞、许可证变化、模型替换、数据污染或第三方服务事件发生时，能够快速判断受影响范围。基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-6 所示：

表 5-6 AI 供应链安全 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 供应链的基本概念；知道开源模型、第三方 API、MCP 的基本安全风险；能执行开源模型安全检查；能评估第三方 AI 服务基础安全；能使用基础供应链安全工具
L2	能设计 AI 供应链安全评估方案；能建立模型可信度验证流程；能评估 MCP 工具链风险
L3	能设计企业级 AI 供应链安全架构；能建立供应商安全管理体系；能制定 AI SBOM 标准
L4	能制定企业 AI 供应链安全策略；能推动 AI BOM、供应商和 Agent/MCP/工具链治理常态化；能评估供应链风险对业务的影响
L5	能制定企业级 AI 供应链安全战略；能参与行业供应链安全标准制定；能推动模型、数据、Agent 和工具生态伙伴安全协作

C06 AI 安全运营

安全运营是 AI 系统持续安全的底线保障，无论架构设计多么完善，缺乏持续的监控、审计和响应能力，安全防线随时可能失守。

对于现代 AI 与 Agent 系统，运营数据除了包含传统主机和 API 日志，还需覆盖模型 / 服务提供商版本、提示词与策略版本、检索增强生成溯源信息、智能体执行轨迹、工具调用链、授权及策略判定、关键人工确认点以及高风险动作结果，使安全事件能够沿完整执行链进行还原。

C06 要求从业者能够建立和运营 AI 系统的安全监控、日志审计、告警响应和事件处理机制，确保 AI 系统在上线运行后能够得到持续的安全保障。

AI 系统的日志审计，比如模型调用日志、输入输出日志、权限变更日志的采集，需要覆盖 AI 特有的操作维度；同时，日志的存储与加密、审计追踪与取证需要满足合规要求。AI 安全运营中心要适配上述 AI 场景的日志产物，AI 特有的安全事件分类、工具选型、AI 事件集成、AI 安全态势感知、自动化响应剧本，是 AI 安全运营的核心能力。

监控告警指标需要覆盖模型行为异常检测、输出质量监控、资源使用异常告警等多个维度。Agent 安全运营需要能够关联用户、模型、Agent 身份、会话、工具调用、权限决策、RAG 来源和执行结果，形成可回放、可审计的完整运行轨迹。最终，要结合 AI 安全事件的分类与分级、响应流程设计、模型紧急回滚、服务降级与熔断、事后复盘与改进，构成了完整的事件管理闭环。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-7 所示：

表 5-7 AI 安全运营 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 安全运营的基本流程；知道 AI SOC 的基本概念；能处理基础告警；能配置 AI 系统日志采集；能执行基础安全监控操作；能按流程处理安全事件
L2	能设计 AI 安全运营方案；能建立 AI 特有的监控指标体系；能主导安全事件响应
L3	能设计企业级 AI SOC 架构；能建立自动化安全运营流水线；能制定 AI 事件响应体系
L4	能统筹企业 AI 安全运营；能建立覆盖模型、RAG 和 Agent 的运营度量体系；能推动执行轨迹、工具调用和策略决策等可观测能力持续优化。
L5	能制定企业 AI 安全运营战略；能推动行业安全运营最佳实践；能建立跨企业安全联防和重大 AI 事件协作机制

C07 AI 治理体系建设

从技术层面转向制度层面，C07 是其中的关键一步。C07 关注的是企业 AI 治理的顶层设计，即如何在组织内部，建立起一套能够自我运转、持续改进的 AI 安全治理体系。

AI 治理委员会的组建是体系建设的起点，它需要明确委员会的组成与职责、决策机制与议事规则、跨部门协同模型以及汇报路线。一个有效的 AI 治理委员会，是真正能够推动 AI 安全决策落地的权力中心。

管理制度建设是治理落地的载体，比如数据训练安全规范、数据标注安全规范、AI 系统使用管理制度、AI 开发安全规范、AI 部署审批流程、AI 安全责任矩阵等。通过制度文件，将抽象的安全要求转化为具体的行为规范，在企业内部落地。同时，开展制度的宣贯与执行监督，防止制度文件变成一纸空文。

AI 资产管理是 AI 治理的基础，AI 资产分类分级、台账管理、退役管理、模型版本管理，是 AI 资产管理的基本框架。和其他的软件系统一样，企业需要建立 AI 资产的发现与盘点机制，加强 Shadow AI 的发现和管控，将 AI 系统纳入生命周期管理，从需求定义、开发、测试、部署到运行、退役的每一个阶段嵌入安全，以提供前置保障。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-8 所示：

表 5-8 AI 治理体系建设 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 治理的基本概念；知道 AI 治理委员会的作用；了解 AI 资产管理的重要性；

	能参与 AI 治理制度的执行；能协助进行 AI 资产盘点；能记录和报告治理执行情况
L2	能设计 AI 治理流程和制度；能建立 AI 资产管理体系；能推动治理措施在团队落地
L3	能设计企业级 AI 治理架构；能建立 AI 生命周期安全治理体系；能制定治理度量指标
L4	能领导企业 AI 治理体系建设；能推动治理文化的形成；能在管理层获得治理资源支持
L5	能制定企业级 AI 治理战略；能影响行业治理标准；能在监管机构前代表企业陈述治理实践

C08 AI 法规与合规

在当前快速演进的监管环境中，AI 法规与合规能力的重要性怎么强调都不为过。目前，中国大陆已经建立起较为完善的 AI 专项监管法规体系，AI 安全从业者需要深入理解并确保企业 AI 应用的合规运行。

中国大陆的 AI 监管法规涵盖了多个层面，《生成式人工智能服务管理暂行办法》对面向公众的生成式 AI 服务提出了算法备案、安全评估、内容安全和数据合规等要求；《互联网信息服务算法推荐管理规定》要求具有舆论属性或社会动员能力的算法推荐服务提供者履行备案手续，并确保算法透明度和用户权益保护；《互联网信息服务深度合成管理规定》则对深度合成技术提供者提出了内容安全和数据保护的要求。此外，《人工智能生成合成内容标识办法》要求对 AI 生成内容进行强制标识，《数据安全法》和《个人信息保护法》中的 AI 相关条款同样需要纳入合规考量。

在国际标准方面，《欧盟人工智能法案》的风险分级体系、高风险 AI 系统要求、通用 AI 模型要求等，为出海企业提供了合规参照。ISO/IEC 42001（AI 管理体系）、ISO/IEC 23894（AI 风险管理）、NIST AI 风险管理框架、MITRE ATLAS 等国际标准/框架，则为 AI 安全治理提供了方法论支撑。不同行业，比如金融、医疗、自动驾驶、教育等，还有行业专项监管要求，从业者需要根据企业所在行业进行针对性的合规管理。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-9 所示：

表 5-9 AI 法规和合规 L1~L5 等级划分评估细则

等级	**能力要求**

L1	了解主要 AI 法规的基本要求；知道算法备案和大模型备案的存在；能执行基础合规检查；能整理合规文档；能识别明显的合规风险
L2	能设计合规检查方案；能编写合规评估报告；能跟踪法规变化并评估影响
L3	能设计企业 AI 合规管理体系；能主导重要合规项目；能建立持续合规监控机制
L4	能制定企业 AI 合规战略；能与监管机构有效沟通；能在合规要求间找到最优解
L5	能参与 AI 法规的讨论和制定；能在国际场合代表企业发言；能建立行业合规协作网络

C09 AI 伦理与知识产权

AI 伦理与知识产权虽然不像技术漏洞那样，会直接导致系统崩溃，但一旦处理不当，可能给企业带来声誉损害、法律纠纷乃至社会信任危机。

公平性是 AI 伦理中最受关注的核心议题之一，AI 偏见可能来自训练数据的不均衡、特征工程的选择偏差，或模型设计的固有缺陷。AI 安全从业者需要掌握偏见的来源和类型、公平性度量指标、偏见检测和缓解方法，以及公平性审计的完整流程。多样性和包容性考量则要求，在 AI 系统设计之初就将公平性纳入考量。

可解释性是连接技术与信任的桥梁，LIME、SHAP 等可解释 AI 方法能够帮助理解模型决策的逻辑，但解释性与准确性之间往往存在张力。AI 安全从业者需要理解面向不同受众，比如技术人员、管理层、终端用户、监管机构的解释策略，掌握可解释性在合规场景中的关键作用。

版权问题是当前 AI 领域争议最大的议题之一，AI 生成内容的版权归属、训练数据的版权合规、AI 辅助创作的法律界定、开源模型许可证合规，这些法律灰色地带正在通过各国的司法实践和立法进程逐步明晰，但目前仍存在大量不确定性。企业需要在创新与合规之间寻找平衡点，建立有效的版权风险管理机制。

AI 责任归属是当前 AI 领域急需明确的议题之一，当 AI 系统做出错误决策或造成损害时，责任应该由模型开发者、部署者还是使用者承担？AI 事故的法律责任如何界定？保险和赔偿机制如何设计？这些问题需要在合同条款、内部责任矩阵和企业治理层面得到系统性的回答。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-10 所示：

表 5-10 AI 伦理与知识产权 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 伦理基本概念；知道 AI 偏见和版权的基本问题；能识别基础的 AI 伦理风险；能执行简单的偏见检测；能整理版权合规文档
L2	能设计公平性评估方案；能选择和应用可解释性工具；能评估 AI 版权风险
L3	能设计企业 AI 伦理框架；能建立 AI 伦理审查机制；能处理复杂版权问题
L4	能领导企业 AI 伦理治理；能推动 AI 伦理文化的形成；能在伦理争议中做出判断
L5	能参与 AI 伦理标准制定；能在国际 AI 伦理讨论中代表企业；能推动行业伦理实践

C10 AI 业务风险评估

将技术风险翻译为业务语言，是获得管理层的重视和资源支持的重要途径。C10 要求从业者能够识别和评估 AI 应用场景中的业务风险，将技术风险与业务影响关联，为管理层提供风险决策依据，这是连接技术团队与管理层的关键桥梁能力。

场景识别是 AI 业务风险评估的起点。AI 应用场景千差万别，从智能客服到风控模型，从内容推荐到自动驾驶，每个场景的风险特征各不相同。从业者需要建立场景风险画像，识别行业特有的 AI 场景风险，并对新兴场景的风险进行前瞻性预判。

场景识别之后，需要开展风险分析。业界有系统化的评估框架，比如 AI RMF。AI 安全从业者需要将框架与企业实际相结合，采用风险量化方法、攻击面分析、业务影响分析和风险关联性分析等工具或方法，从定性和定量两个维度理解 AI 风险，帮助决策者判断如何做出下一步的风险接受与风险控制。

通用的风险管理框架应包含风险容忍度的定义、风险接受审批流程、残余风险管理、风险沟通与报告，这些都是风险决策的制度性保障。具体到执行层面，应采用技术控制、管理控制、控制措施有效性评估、风险转移策略以及安全投入产出分析，有助于帮助企业在安全投入与业务收益之间找到最优平衡。

而对于具备自主执行能力的 Agent，业务风险评估还须判断其自治等级、可调用工具、数据权限、单次动作影响范围和是否需要人工确认。实际分级时，可以同时使用几组简单且可执行的维度，帮助从业者区分风险等级。比如，服务对象区分内部使用与面向公众；决策性质区分辅助决策与自主决策；Agent 权限区分只读、受限写入与高影响执行；操作后果区分可逆与不可逆；数据范围区分公开、内部、敏感和重要数据。多维组合后的最高风险等级，应作为上线审批、人工确认、红队深度和持续监控强度的依据。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-11 所示：

表 5-11 AI 业务风险评估 L1~L5 等级划分评估细则

等级	能力要求
L1	了解 AI 业务风险评估的基本概念；能识别简单的 AI 场景风险；能执行标准化风险评估；能使用风险评估工具；能编写风险评估报告
L2	能设计风险评估方案；能量化 AI 风险对业务的影响；能提出风险控制建议
L3	能设计企业级 AI 风险管理框架；能建立风险评估标准体系；能为管理层提供风险决策支持
L4	能统筹企业 AI 风险管理；能推动用户场景与 Agent 风险分级融入业务流程；能在风险与创新间取得平衡
L5	能制定企业级 AI 风险管理战略；能决定高风险 AI Agent 的风险接受原则；能参与行业风险标准和跨企业风险情报协作

C11 AI 安全事件管理

当 AI 安全事件真正发生时，响应速度和处置质量对事件损失的影响至关重要。C11 要求从业者能够建立 AI 安全事件的预防、检测、响应和恢复机制，确保在事件发生时能够快速有效处置，将业务影响降到最低。

应急预案是事件管理的基石。AI 安全事件的分类，数据泄露、模型滥用、输出失控、对抗攻击等攻击方式的差异，决定了响应策略的不同。应急预案的设计需要覆盖从团队组建、响应流程到演练执行的完整环节，并建立定期更新机制，以应对不断演变的威胁态势。

与传统软件不同，模型回滚作为安全事件管理的组成部分，需要考虑模型版本管理策略、紧急回滚决策流程、回滚后的验证检查、灰度发布与快速回退，以及模型降级策略。一套成熟的模型回滚机制，能够在关键时刻将损失控制在最小范围。

舆情管理在 AI 安全事件中尤为重要。由于 AI 直接面向用户交互，AI 安全事件往往更容易引发公众关注。舆情监测、对外沟通策略、媒体应对指南、用户通知机制和品牌声誉保护，这些都需要在事件发生前就做好预案。危机响应指挥体系、跨部门协调、法律顾问介入、监管报告义务，以及事后复盘与改进，共同构成了完整的事件管理闭环。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-12 所示：

表 5-12 AI 安全事件管理 L1~L5 等级划分评估细则

等级	能力要求

L1	了解 AI 安全事件的基本分类；知道应急预案的存在和重要性；能按预案执行基础事件响应；能协助进行事件记录和报告；能参与应急演练
L2	能主导 AI 安全事件响应；能设计事件响应流程；能编写事件分析报告和改进建议
L3	能设计企业级 AI 事件管理体系；能建立模型回滚机制；能制定舆情管理策略
L4	能领导企业 AI 安全事件管理；能建立智能体紧急停止开关、凭据吊销、工具隔离、记忆隔离等处置机制；能在危机中做出关键决策
L5	能制定企业级 AI 安全事件战略；能推动行业事件信息共享；能在重大模型或 Agent 安全事件中代表企业对外和监管沟通

C12 安全沟通与培训

在所有十二项核心能力中，C12 可能是最容易被忽视但影响最深远的一项。再完善的安全体系，如果无法在组织内部有效传播和落地，就只是一套纸面制度。安全沟通与培训能力，决定了 AI 安全理念能否真正渗透到从技术团队到管理层的每一个角落。

向管理层汇报是 AI 安全官必须掌握的核心技能。AI 安全仪表盘设计、风险量化与可视化、用业务语言翻译技术风险、汇报频率与内容设计、投资回报分析表达，这些能力有助于帮助 AI 安全官在有限的汇报时间内传递最关键的信息，争取管理层的理解和支持。

跨部门协作是 AI 安全落地的关键路径。与 AI 研发团队的协作，需要理解开发流程，并在安全与效率之间找到平衡；与产品团队的安全左移，要求在产品设计阶段就介入安全考量；与法务团队的合规协同，确保安全措施与法规要求对齐；与管理层的沟通策略，需要因人而异、因势而变。与外部安全社区的连接，企业需要有获取前沿威胁情报和技术交流的渠道。

安全文化建设是 C12 的终极目标。AI 安全文化评估、安全激励机制、安全事件分享机制、安全竞赛、安全文化度量指标，有助于帮助企业从被动合规走向主动安全。逐步构建全员 AI 安全培训体系，需要根据受众（AI 开发者、产品经理、管理层、全员）设计分层培训课程，建立培训效果评估和持续更新机制，确保安全意识在文化建设中的可持续性。

基于上述要求，对 L1~L5 等级的能力要求划分如下表 5-13 所示：

表 5-13 安全沟通与培训 L1~L5 等级划分评估细则

等级	能力要求
L1	能清晰表达 AI 安全基本概念；能参与安全培训的辅助工作；能编写基础 AI

	安全文档；能在团队内分享 AI 安全知识；能参与培训课程的辅助教学
L2	能设计 AI 安全培训课程；能向技术团队有效传达安全要求；能推动安全实践在团队落地
L3	能设计企业级 AI 安全培训体系；能向管理层有效汇报 AI 安全；能推动跨部门安全协作
L4	能领导企业 AI 安全文化建设；能在董事会层面沟通 AI 安全；能推动全员安全意识提升
L5	能在行业层面传播 AI 安全理念；能影响行业安全文化；能代表企业参与公共安全讨论

5.3 企业 AI 采纳深度与能力适配

在前面的章节中，完成了 AI 安全官个人需要具备什么能力的框架构建。但在实际落地时，不同企业应用 AI 技术的深度和方式，对 AI 安全官的能力需求和核心要点往往存在巨大差异。

一家从基础设施到应用层全栈自研的 AI 巨头企业，和一家仅通过调用大模型 API 为产品添加智能功能的传统企业，它们面临的 AI 安全风险在类型、规模和优先级上都截然不同。如果用同一套标准要求所有企业，结果要么是标准过高导致中小企业望而却步，要么是标准过低导致全栈企业防护不足。

为解决这一问题，本节引入**企业 AI 采纳深度**的概念，将企业按 AI 技术栈的覆盖范围划分为三种典型类型，并为每种类型明确能力适配的重点和团队配置的建议，帮助每家企业，找到最适合自身当前阶段的 AI 安全治理起点。

5.3.1 三种企业 AI 采纳类型

类型 A：AI 全栈型企业

这类企业覆盖 AI 技术栈的完整层次，从 GPU 集群管理、模型训练与开发，到模型部署与推理服务，再到面向用户或业务的 AI 应用。它们拥有自己的 AI 研发团队，可能训练基础模型或进行大规模微调，管理着相当规模的计算资源。典型的代表包括大型科技公司的 AI 部门、AI 基础模型创业公司，以及拥有自建 AI 平台的大型集团，比如国内的 DeepSeek、智谱、阿里千问等，国外的 OpenAI、Anthropic、xAI 等。

全栈企业的 AI 安全风险是全方位的，它们既面临 GPU 集群安全、分布式训练通信安全等基础设施层面的威胁，又面临模型窃取、数据投毒、后门植入等模型层面的风险，同时还必须应对提示词注入、越狱攻击等应用层面的攻击，任何一个环节的失守都可能产生连锁反应。因此，本白皮书建议这类企业，需要的是一支覆盖所有能力域的专业团队，通常由 L5 级以上的 AI 安全官统一领导。

类型 B：私有化部署型企业

这类企业不训练自己的基础模型，但在自己的数据中心或云端基础设施上部署、运行 AI 模型。它们通常基于开源模型（如 LLaMA、ChatGLM 等）进行领域微调，或者将第三方模型部署在自己的推理环境中。典型的代表包括在私有化环境中部署大模型的企业、对开源模型进行定制化的 SaaS 公司，以及在自有服务器上运行 AI 质检、AI 客服等应用的传统企业。比如，国内的硅基流动/问小白、国外的 OpenRouter。

私有化部署型企业的风险主要集中在模型定制和部署环节，需要评估开源模型的安全性，是否存在预留后门、许可证是否合规、模型完整性是否可靠。推理环境的容器安全、API 安全、模型版本管理也是 AI 私有化部署型企业关键风险点。相比全栈型企业，它们不需要关注基础模型训练阶段的风险，但比纯消费型企业承担了更多的基础设施和模型安全责任。因此，本白皮书建议这类企业，通常需要 L4 级以上的 AI 安全负责人员，配合 3 到 5 人的专职团队即可。

类型 C：AI 消费型企业

这类企业不开发基础模型，也不部署开源 AI 模型，而是通过调用第三方 AI 服务的 API 来集成 AI 能力。它们的典型场景包括：使用大模型 API 构建智能客服、利用 AI API 进行内容生成、通过第三方推荐 API 实现个性化推荐等。绝大多数传统企业的 AI 转型，都从这一类型起步。

AI 消费型企业的风险面相对聚焦，它们不需要担心 GPU 集群安全或模型训练阶段的数据投毒，但面临着两类核心风险：一是应用层面的对抗攻击，比如提示词注入、越狱攻击、输出滥用等问题，这些在面向用户的场景中尤为突出；二是供应链层面的 API 依赖风险，AI 服务提供商的安全能力、SLA 保障、数据传输安全、供应商集中程度等，都直接影响企业自身的 AI 安全。合规管理（算法备案、内容安全等）同样是这类企业的刚性需求。因此，本白皮书建议这类企业，通常由 L3 级以上的安全人员兼职或专职负责，团队规模 1 到 3 人即可覆盖核心需求。

5.3.2 不同类型 AI 企业能力适配

上述三种企业类型对十二项核心能力的需求存在显著差异，故本白皮书针对这些企业特点，整理出下表，以重要性等级（核心 / 重要 / 一般）来标注了每项能力对不同企业类型的重要性，详情如下表 5-14 所示：

表 5-14 不同类型 AI 企业 12 项核心能力重要性等级

核心能力	类型 A 全栈型企业	类型 B 私有化部署型企业	类型 C 消费型企业
C01 算法与数据安全	核心	重要	一般
C02 AI 对抗防御	核心	核心	核心
C03 AI 红队测试	核心	重要	一般
C04 AI 基础设施安全	核心	重要	一般
C05 AI 供应链安全	核心	核心	核心
C06 AI 安全运营	核心	核心	重要
C07 AI 治理体系建设	核心	核心	核心
C08 AI 法规与合规	核心	核心	核心
C09 AI 伦理与知识产权	核心	重要	重要
C10 AI 业务风险评估	核心	核心	核心
C11 AI 安全事件管理	核心	核心	重要
C12 安全沟通与培训	核心	核心	核心

在这张表中，无论企业属于哪种类型，C02（AI 对抗防御）、C05（AI 供应链安全）、C07（AI 治理体系建设）、C08（AI 法规与合规）、C10（AI 业务风险评估）和 C12（安全沟通与培训）这六项能力都是必须具备的，它们构成了 AI 安全的**最小可行能力集**，适用于所有采纳 AI 技术的企业。

5.3.3 风险优先级差异

虽然三种类型都面临全方位的 AI 安全风险，但因企业采用 AI 的深度不同，风险的优先级排序截然不同。

对于类型 A 全栈企业，风险优先级从高到低大致为：模型安全（窃取、投毒、后门）> 基础设施安全（GPU 集群、分布式训练）> 应用安全（提示词注入、越狱攻击）> 供应链安全。这是因为，基础大模型是全栈企业的核心资产，一旦模型被窃取或投毒，影响是决定性的。

对于类型 B 部署企业，风险优先级为：开源模型供应链安全 > 应用层对抗攻击 > 推理环境安全 > 合规管理。开源模型是这类企业最独特也最脆弱的环节，未经充分安全评估的开源模型，最容易通过已知漏洞将整个 AI 系统暴露在风险之下。

对于类型 C 消费企业，风险优先级为：API 供应链安全 > 应用层对抗攻击 > 合规管理 > 事件响应。API 是这类企业与 AI 能力之间的唯一连接点，其安全性直接决定了整个 AI 应用的安全水平，因此理应受到重视。

5.3.4 团队配置建议

基于三种企业类型的差异化需求，团队配置也应有所不同。下表 5-15 为本白皮书建议的人员配备，详细如下：

表 5-15 不同类型 AI 安全团队人员配置

维度	类型 A 全栈型企业	类型 B 私有化部署型企业	类型 C 消费型企业
AI 安全官等级	L5	L4	L3
团队规模	10+人	3-5 人	1-3 人（可兼职）
D1 配置	专职攻防团队（3-4 人）	1-2 人负责模型安全与红队	可由安全工程师兼管
D2 配置	专职基础设施安全团队（3-4 人）	1-2 人负责部署安全与运营	可由安全工程师兼管
D3 配置	专职合规与治理团队（2-3 人）	1-2 人负责合规与治理	1 人负责合规
D4 配置	专职风险与沟通团队（2-3 人）	1 人兼管风险与沟通	由 AI 安全官或管理层兼顾

需要特别强调的是，上述三种企业类型并非一成不变的。一家类型 C 企业，可能在积累经验后开始微调开源模型，向类型 B 过渡；一家类型 B 企业，可能在业务发展到一定规模后，开始自研模型，向类型 A 演进；而一家类型 A 企业，可能因为自研模型竞争力不足，放弃自研，选择微调开源模型，向类型 B 过渡。

作为企业管理者，AI 安全官定义的 AI 安全治理框架需要具备足够的弹性来适应这种变化或演进，在企业每个年度的安全规划中，重新评估自身的 AI 采纳类型，相应调整能力优先级和团队配置。同时，在向更高采纳深度跃迁时，应提前 6 个月开始补齐相应的能力短板，而不能等 AI 系统已投入使用后再临时补救。

第 6 章 AI 安全考核指标

6.1 考核指标概述

前面几章讨论的是**一个人需要具备什么能力**，那么这一章要回答的是**企业的 AI 安全到底做到了什么水平**，通常使用 AI 安全治理成熟度，这是衡量组织 AI 安全治理水平的通用框架。但当前阶段，谈及企业的 AI 安全治理成熟度为时尚早，谈及 AI 安全建设过程中的考核指标更贴合实际情况。

而 AI 安全的度量因 AI 系统的概率性特征、当前快速演进的威胁态势、以及缺乏行业基准，目前市场上可参考的方法论支撑案例也是比较少，本白皮书采用通用指标（KPI）编制方法，适配 AI 安全实践案例，汇总如下。

6.2 指标设计原则

我们在构建度量指标时，参考业界通用做法，遵循 SMART 原则：

- Specific（具体）——每个指标必须明确衡量什么；
- Measurable（可衡量）——指标的数据必须可以通过技术手段或流程记录获取；
- Attainable（可实现）——指标通过合理努力可以达成；
- Relevant（相关性）——指标必须与 AI 安全目标直接相关；
- Time-bound（时限性）——指标必须有明确的完成时间周期。

AI 安全官在指标构建时，应结合企业当前的实际现状和当前业务战略，拟定不同的度量指标。一般来说，通常包含 AI 安全技术相关的核心度量指标、管理层关注的业务价值度量指标以及安全内部成本优化的相关指标。这些指标的具体案例，可参考 6.3 章节的内容。

6.3 核心考核指标参考案例

本白皮书建议，可以从下面 5 个维度构建 AI 安全度量体系，每个维度包含关键指标和计算方法，参考案例如下：

维度一：安全防御类

指标名称	计算方法	目标值样例	评估周期
攻击阻断率	成功阻断的攻击数 / 发现的总攻击数 × 100%	≥95%	月度
安全护栏拦截有效率	安全护栏正确拦截的恶意输入 / 总恶意输入数 × 100%	≥90%	月度
红队发现率	红队发现的高危漏洞数 / 红队测试用例总数 × 100%	持续追踪趋势	季度
新型攻击检测时间	从攻击首次出现到被检测到的平均时间	≤24 小时	月度

维度二：事件响应类

指标名称	计算方法	目标值样例	评估周期
AI 安全事件平均响应时间（MTTR）	事件发现到开始处置的平均时间	≤2 小时	月度
AI 安全事件平均修复时间	事件开始处置到完全解决的平均时间	≤8 小时	月度
事件根因分析完成率	完成根因分析的事件数 / 总事件数 × 100%	100%	月度
事件复发率	同类事件再次发生的次数 / 总事件数 × 100%	≤5%	季度

维度三：合规与治理类

指标名称	计算方法	目标值样例	评估周期
AI 安全合规审计通过率	通过审计的 AI 系统数 / 总 AI 系统数 × 100%	≥90%	半年度
AI 资产覆盖率	已纳入台账管理的 AI 资产数 / 总 AI 资产数 × 100%	≥95%	季度
AI 安全培训覆盖率	参加 AI 安全培训的人员数 / 总相关人员数 × 100%	≥80%	年度
制度更新及时率	在法规变更后规定时间内完成制度更新的比例	100%	按需

维度四：风险管控类

指标名称	计算方法	目标值样例	评估周期

AI 风险评估完成率	已完成风险评估的 AI 场景数 / 总 AI 场景数 × 100%	≥90%	季度
高风险控制有效率	高风险项控制措施的有效执行数 / 总高风险项数 × 100%	≥95%	月度
风险处置及时率	在规定时间内完成风险处置的比例	≥90%	月度
安全投入 ROI	安全投入避免的潜在损失 / 安全投入成本	≥3:1	年度

维度五：持续改进类

指标名称	计算方法	目标值样例	评估周期
安全能力成熟度提升	ASGMM 评估等级变化	每年提升≥0.5 级	年度
安全改进项完成率	按时完成的安全改进项数 / 总安全改进项数 × 100%	≥80%	季度
安全事件复盘率	完成复盘的事件数 / 总事件数 × 100%	100%	月度
行业对标差距	与行业基准的差距变化趋势，量化后用雷达图对比	持续缩小	年度

6.3.1 业务价值考核指标拟定参考

技术指标固然重要，但管理层更关心的是 AI 安全投入这么多成本，究竟产生了多少商业价值？所以，从业务角度制定的度量指标非常重要。

下面是我们整理的常见业务类型的度量指标拟定方法，供参考，详细如下：

1. **损失规避价值：** 量化 AI 安全工作避免的潜在损失，来呈现安全工作的价值。典型指标包括：AI 安全事件预防数量×单次事件平均损失估算（含直接损失和声誉损失），合规审计发现问题在上线前拦截的比例×行业平均罚款金额，敏感数据泄露风险暴露次数×数据泄露平均每条记录成本（比如，根据 IBM 数据泄露成本报告，全球平均每条记录成本约 150 美元来换算）。这类指标，有助于帮助管理层理解**如果没有 AI 安全团队的工作，我们可能会损失多少**。
2. **效率提升价值：**量化 AI 安全带来的运营效率改善，来呈现安全工作的价值。典型指标包括：AI 安全事件平均响应时间（MTTR）的缩短幅度×事件处理人力成本，AI 红

队测试发现漏洞的提前量（在上线前还是上线后发现）×修复成本差异（上线后修复成本通常是上线前的 6 到 10 倍），自动化安全检查替代人工检查节省的人力工时等。

3. **信任资产价值：** 强调 AI 安全对客户信任和品牌声誉的正向影响力。典型指标包括：AI 安全认证（如 ISO 42001）对客户采购决策的影响，AI 安全事件发生率与客户留存率的关联分析，AI 安全实践在竞标中的差异化优势。虽然这类指标难以精确量化，但可以通过客户调研和竞标分析进行估算，很容易获得业务方和高层认可。

除了上述拟定 AI 安全指标的方法外，还可以从安全成本的角度，拟定改进和优化指标，比如，节约请求延迟成本，大模型产品和安全防护都增加请求延迟，如何减少这种额外耗时，是制定度量指标的一种方式；AI 安全团队人力投入、安全培训工时、安全评审工时，自动化安全流水线的耗时，安全控制点的减少，安全策略分级优化，等等这些都是可以拟定出度量指标的。

当然，对于不同规模、不同成熟度等级的企业，对度量指标的考核侧重点是有所区别的，这一点和传统安全并无多大区别。本白皮书给出一些通用的参考意见，供使用时参考：
Level 1-2 的企业，建议从事件响应类和合规治理类的核心指标开始，建立基础度量能力；
Level 3-4 的企业应覆盖 6.3 章节的五个维度指标，建立自动化数据采集和仪表盘；
Level 5 的企业则需要建立预测性度量能力，通过趋势分析，提前识别潜在风险。

第 7 章 AI 安全官岗位

7.1 AI 安全官岗位说明书

前面的章节回答了 **AI 安全官需要什么能力**，这一章要回答的是**如何将这些能力要求转化为实际的应聘岗位**。一份好的岗位说明书不仅是一张招聘清单，更是组织对 AI 安全官角色定位和期望的集中体现。

需要注意的是，不同 AI 采纳类型的企业（参见第 5.3 章节）对 AI 安全官的侧重方向有所不同：AI 全栈型企业需要技术深度与治理能力兼备的候选人，AI 私有化部署型企业侧重供应链安全和对抗防御经验，AI 消费型企业则更看重合规管理和应用安全能力。

除了企业特性带来的候选人要求差异外，不同岗位、不同职级的招聘需求也存在很大的差异。

比如，中低等级的安全人员招聘，招聘一名 AI 安全助理工程师或者 AI 安全工程师，主要看重的是候选人的执行能力、专业技术、发展潜力，只需要 D1 或者 D2 其中的一项能力突出，能上手干活即可。所以，这 L1~L2 两个层级的招聘，在市场上表现为多个方向，比如 AI 安全工程师-大模型安全方向、AI 安全工程师-红队测试方向、AI 安全工程师-大模型安全测评方向。而对于更高层级的 AI 安全招聘，比如 AI 安全总监或 AI 安全官，则比较统一，很少有区分方向的。这一点，本白皮书下面的 JD 模板会给出参考案例。

7.2 AI 安全官 JD 模板

这里准备了 4 份 JD 模板，分别为：

- **AI 安全工程师，通用大模型方向**，对人员能力要求偏攻防 + 防护 + 运营，综合执行岗。
- **AI 安全工程师，AI 模型对抗方向**，偏向模型底层、攻防研究、评测工具。
- **AI 安全工程师，AI 应用与数据合规安全方向**，偏向上层 AI 业务，重点关注 RAG/Agent 数据与合规落地。
- **AI 安全官 ，高层管理岗**，全面的 AI 治理、战略、团队、合规统筹。

4 个岗位详细 JD 模板如下，仅供企业参考。

模板 1：AI 安全工程师（大模型安全方向）

部门：信息安全部- AI 安全组 | 地点：合肥 | 经验：3- 5 年 | 学历：本科及以上 | 薪资：25 - 40K • 14 薪

岗位职责

1. 负责企业大模型、多模态模型、AI Agent、MCP 工具链的安全测试与攻防验证，开展提示注入、越狱对抗、数据投毒、模型窃取等红队测试，使用 Garak、PyRIT 等工具构建自动化评测能力，输出风险报告与修复建议。
2. 落地 AI 全生命周期安全防护，主要是训练数据安全管控、数据脱敏与投毒检测，部署安全护栏、输入输出过滤、模型水印等防护能力，缓解模型输出失控、敏感信息泄露风险。
3. 负责 AI 基础设施与供应链安全，重点是 GPU 集群、推理 API、容器/K8s 安全基线检查，开源大模型/第三方 API/AI 框架依赖风险评估，推进 AI SBOM 落地，管控 AI 供应链风险。
4. 建设 AI 安全运营能力，完成模型调用日志审计、异常行为监控、告警处置，承接 AI 安全事件应急响应，执行模型回滚、熔断处置，完成事件复盘与改进。
5. 参与编写 AI 安全规范、检查清单，推动安全左移，对接 AI 研发、数据、产品团队，把安全要求嵌入 AI 需求、开发、上线流程，跟踪 AI 安全攻防与监管动态。

任职要求

1. 本科及以上，网络安全、计算机相关专业，3 年以上安全相关工作经验，有大模型/AI 安全实战经验优先。
2. 熟悉 LLM 安全威胁，Prompt 注入、越狱、对抗样本、数据投毒、模型窃取等攻击原理与防御手段，有 AI 红队实战经历加分。
3. 了解 PyTorch/TensorFlow 基础概念，会使用至少一种 AI 安全测试工具，熟悉容器、K8s、API 安全、数据安全基础。
4. 具备漏洞挖掘、安全测试、应急响应能力，能独立输出风险评估报告。
5. 了解国内生成式 AI 相关监管要求，有算法/大模型备案相关经验优先。
6. 良好沟通协作，能够跨部门推动安全整改落地。

模板 2：AI 安全工程师（AI 对抗与模型安全方向）

部门：信息安全部- AI 安全组 | 地点：合肥 | 经验：3- 5 年 | 学历：本科及以上 | 薪资：28 - 42K • 14 薪

岗位职责

1. 聚焦模型底层安全研究，开展对抗样本、模型投毒、模型窃取、模型水印、模型逆向等攻防研究，针对 LLM、多模态图像/音视频模型设计攻击与防御方案。
2. 搭建模型安全评测底座，开发对抗样本生成、投毒检测、模型溯源检测工具，支撑内部自研模型、第三方接入模型的自动化安全评测。

3. 针对微调、蒸馏、提示调优场景识别模型安全缺陷，输出模型训练阶段安全约束方案，输出模型安全测试报告，推动算法团队修复模型固有风险。
4. 跟踪学术界、产业界 AI 对抗安全最新研究成果，复现前沿攻击方法，评估对公司业务模型的实际影响，输出风险预警与防御落地建议。
5. 参与 AI 安全专利、技术白皮书、内部技术沉淀输出，支撑红队演练、模型上线准入评审。

任职要求

1. 本科及以上，计算机、网络安全、人工智能相关专业，3 年及以上安全或 AI 研究相关经验。
2. 熟悉深度学习基础，熟练使用 PyTorch/TensorFlow，了解大模型、多模态模型基础原理。
3. 掌握对抗样本、数据投毒、模型窃取等模型层攻击原理，有相关实验、工具开发实战经验。
4. 具备 Python 开发能力，能够独立开发安全评测脚本与工具，有开源工具贡献、AI 安全 CTF 获奖、论文复现、AI 安全研究产出优先。
5. 了解模型水印、模型溯源、隐私计算相关技术，有模型安全红队测试经验加分。

模板 3：AI 安全工程师（AI 应用与数据合规安全方向）

部门：信息安全部- AI 安全组 | 地点：合肥 | 经验：3- 5 年 | 学历：本科及以上 | 薪资：24 - 38K • 14 薪

岗位职责

1. 面向 AI 业务应用侧做安全管控，覆盖 AI Agent、RAG 知识库、业务 AI 插件、第三方大模型 API 接入场景，负责应用层风险识别与防护落地。
2. 落地 AI 业务数据安全，例如训练/知识库/用户输入输出数据分级分类、脱敏、去重、隐私泄露检测，防范用户隐私、企业敏感数据被模型记忆、提取泄露。
3. 承接生成式 AI 合规落地，支撑大模型/算法备案材料准备，落实输出内容风控、伦理风险、生成内容版权风险管控，制定 AI 应用上线安全准入检查清单。
4. 建设 AI 应用安全运营，比如敏感输出检测、违规提示监控，处置 AI 业务安全事件，输出合规风险报告，推动产品、算法、法务落地整改。
5. 解读生成式 AI 管理办法、个人信息保护等法规，把合规要求转化为可执行安全技术规则与产品约束。

任职要求

1. 本科及以上，计算机、网络安全、法学（信息方向），3 年安全、数据安全或业务风控经验，有 AI 业务落地经验优先。
2. 熟悉 RAG、Agent 业务架构，理解检索注入、提示注入、记忆泄露等 AI 应用层风险，熟悉数据安全、个人信息保护相关法规。
3. 会 Python 脚本开发，能够做内容检测、数据扫描类工具开发，了解大模型安全护栏、内容安全护栏相关方案。
4. 有数据安全治理、内容安全、合规备案实操经验，能够编写检查清单、风险评估报告。

模板 4：AI 安全官

部门：信息安全部 | 汇报：CEO/CTO | 地点：合肥 | 经验：8- 12 年 | 学历：本科及以上 | 薪资：50- 80K • 16 薪，可配团队管理期权

岗位职责

1. 作为企业 AI 安全负责人，搭建公司 AI 安全治理体系，制定 AI 安全整体战略、制度流程与责任矩阵，覆盖 AI 需求、开发、训练、部署、运行、退役全生命周期风险管控。
2. 统筹 AI 安全风险治理，统筹 AI 攻防、数据安全、供应链安全、安全运营、合规伦理工作，识别评估业务场景下 AI 业务风险，定义风险容忍度，平衡 AI 创新与安全约束。
3. 负责 AI 合规管理，跟进《生成式人工智能服务管理暂行办法》、欧盟人工智能法律法规、ISO42001 等国内外法规标准，主导算法备案、大模型备案、AI 合规评估，处理 AI 伦理、知识产权、可解释性相关问题。
4. 搭建并管理 AI 安全团队，规划团队能力建设、人员招聘、绩效考核。建立常态化 AI 红队、AI 安全评测、AI SOC 运营机制，主导重大 AI 安全事件危机处置、舆情与监管沟通。
5. 向管理层、董事会汇报 AI 安全风险、投入产出与治理进展，推动跨部门协同（AI 研发、产品、法务、业务），建设企业 AI 安全文化，开展全员与研发 AI 安全培训。
6. 参与行业 AI 安全交流，输出实践，跟进行业标准与威胁态势。

任职要求

1. 本科及以上，计算机/信息安全相关专业，8 年以上信息安全工作经验，至少 2 年 AI 安全/AI 治理相关负责人经验。
2. 深刻理解大模型、Agent、MCP 生态安全风险，熟悉提示注入、对抗攻击、数据投毒、AI 供应链等威胁，懂技术同时具备治理合规视角。
3. 具备大型企业安全体系建设经验，有治理制度、流程、指标体系搭建实战。熟悉国内外 AI 监管合规体系，有算法/大模型备案实操经验。
4. 团队管理能力，能够搭建 AI 安全团队，做人才梯队规划。优秀的向上汇报与跨部门博弈能力，能够在业务创新和安全之间取得平衡。
5. 经历过重大安全事件完整应急处置，具备危机、舆情、监管对接经验。
6. 有大模型企业/互联网/金融/央企 AI 安全负责人经历优先，有行业白皮书/标准参与编制经验优先。

第 8 章 组织架构

8.1 AI 安全治理组织架构参考

本白皮书推荐的 AI 安全治理组织架构，如下图 8-1 所示，供企业使用时参考。

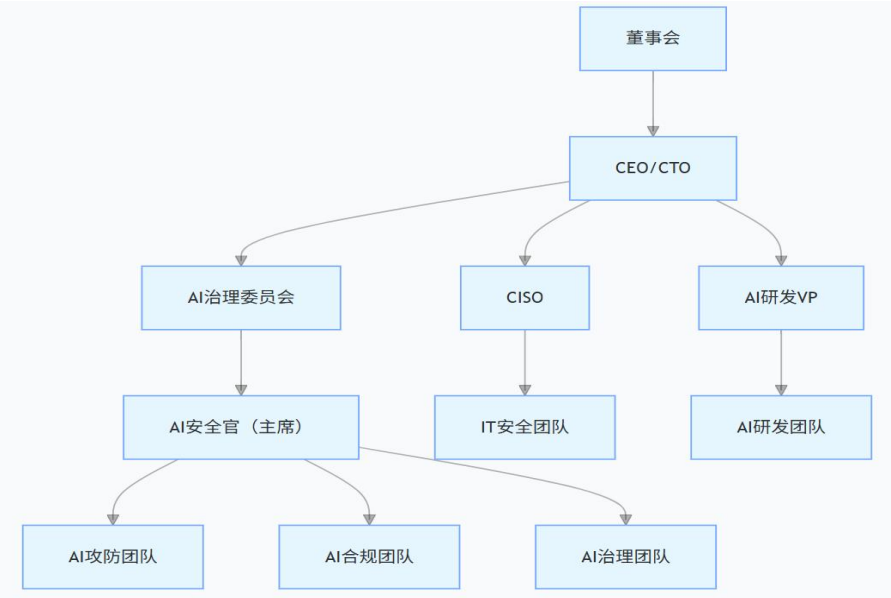


图 8-1 AI 安全与治理组织架构参考图

从图 8-1 中可以看出，AI 安全与治理组织架构采用自上而下的垂直治理模式，由**董事会**作为最高监督机构，向下对 CEO/CTO 进行监督问责。

CEO/CTO 统一统筹三条业务线：AI 治理委员会、CISO、AI 研发 VP。

1. **AI 治理委员会**下设 AI 安全官（主席），直接管理 AI 攻防团队、AI 合规团队、AI 治理团队，负责 AI 全生命周期的风险管控、合规审查与 AI 安全攻防能力建设。
2. **CISO** 分管 IT 安全团队，承担传统 IT 基础设施安全保障，与 AI 治理体系形成协同联动。
3. **AI 研发 VP** 管理 AI 研发团队，负责 AI 模型与产品的研发落地，接受上层 AI 治理体系的安全与合规约束。

整套架构实现 AI 研发、AI 安全治理、传统 IT 安全、审计监督的分层分离，既保障 AI 业务迭代效率，同时将 AI 风险、合规要求嵌入组织层级，实现自上而下的 AI 风险管控。

8.2 AI 治理委员会

AI 治理委员会是企业 AI 安全治理的最高决策机构，其成员构成主要跨领域分管领导。

AI 安全官作为主席，负责委员会日常运营和 AI 安全决策；CISO 统筹企业整体安全架构，确保 AI 安全与企业整体安全战略一致；AI 研发 VP 参与其中，确保安全要求在研发中落地；产品负责人负责产品侧的安全落地，在用户体验与安全之间寻找平衡；法务负责人从法律角度，确保合法合规；业务代表则从业务目标达成视角，确保安全措施不影响业务效率。

除上述常规人员外，必要时，还可以引入外部顾问，比如行业专家或合规顾问，以提供第三方独立视角。

委员会的运作机制需要制度化保障，建议每月召开一次全体会议，紧急事项可临时召集，重大安全决策需经委员会多数成员同意。委员会直接向 CEO/CTO 汇报，重大事项向董事会汇报，所有会议决议需形成书面记录，并跟踪执行情况，以保证决策得以落实。

8.3 各企业规模的团队配置建议

针对不同规模的企业，本白皮书建议 AI 安全团队做以下配置，如下表 8-1 所示：

表 8-1 不同企业规模的 AI 安全团队配置建议

企业规模	团队规模	最低能力要求	关键岗位
初创（50 人以下）	1 人（兼职）	L1	现有安全人员兼任
中型（50-500 人）	2-3 人	L2-L3（负责人）	AI 安全工程师、AI 合规专员
大型（500 人以上）	5-10 人	L4-L5（AI 安全官）	AI 安全官、AI 攻防团队、AI 合规团队、AI 治理团队

初创企业（1 人兼职模式）： 由于资源有限，初创企业不需要也不适合设置专职 AI 安全官岗位，但至少应有 1 人具备 L1 能力。这个人通常由现有安全人员兼任，负责 AI 使用规范的制定和执行、完成基础的 AI 安全检查，以及协助完成算法备案等合规事项。虽然企业规模最小，但有人管和没人管之间的安全水平差距是非常明显的。

中型企业（2-3 人专职模式）： 随着 AI 应用规模的增长，兼职模式已无法满足安全需求。这类中型企业至少需要 1 名 L2 或 L3 级别的 AI 安全专业人员作为技术骨干，配合 1 名 AI 合规专员负责法规合规，以及 1 名 AI 安全运营专员负责日常监控。三人虽少，但已基本能够覆盖技术安全、合规管理和运营监控三个核心职能。

大型企业（5-10 人团队模式）： 大型企业通常需要 L4 或 L5 级别的 AI 安全官统筹全局，下设三个能力域各 2-3 人的专业团队。其中 AI 攻防团队，负责红队测试和对抗防御；AI 合规团队，负责法规合规和治理体系建设；AI 治理团队，负责制度建设、培训和文化建设。三个团队在 AI 安全官的统一领导下协同运作，形成完整的企业级 AI 安全能力。

需要说明的是，上述配置主要基于企业规模维度考虑的。在实际落地时，还应结合第 5.3 章节提出的企业 AI 采纳类型进行调整，人员数量配置适合、适度即可。

第 9 章 企业案例与行业建设路线

明确了组织架构和团队配置之后，下面再来看看 AI 安全从业人员入职企业后，如何着手开展 AI 安全与治理工作。

这里，我们依然参考 8.3 章节，针对不同规模的企业，从初创企业、中型企业、大型企业，分别给出短中期的安全建设规划建议。

9.1 按企业规模的建设路线

9.1.1 初创企业（50 人以下）

初创企业的 AI 安全建设应以**轻量起步、风险优先**为原则，不建议一开始搭建复杂组织体系。前 6 个月重点完成基础风险评估、AI 使用规范和供应链安全基础能力，6－12 个月通过检查清单、轻量级资产管理和核心人员培训，建立可执行的 AI 安全基线，如图 9-1 所示。

人员配置在初期不需要专职 AI 安全官，但应确保至少有 1 名具备 L1 基础能力的人员承担 AI 安全职责。



图 9-1 初创企业 AI 安全工作规划

9.1.2 中型企业（50-500 人）

中型企业基本进入**制度建设 + 专人负责**的阶段，应该重点解决 AI 安全责任不清、资产底数不明、工具能力不足等问题。本白皮书建议在 0－6 个月内需要完成现状评估、制度建设、AI 资产盘点和专职人员配置，6－12 个月建立治理机制、工具链、事件响应和全员培训，12－24 个月进一步形成小团队作战能力，并评估设立 AI 安全官 岗位。这

样的规划，能更好地适配企业当前现状，重点解决合规问题，推动 AI 安全工作从单点合规走向持续治理。如下图 9-2 所示。

建议至少配备 1 名 L2 或 L3 级别的 AI 安全工程师。

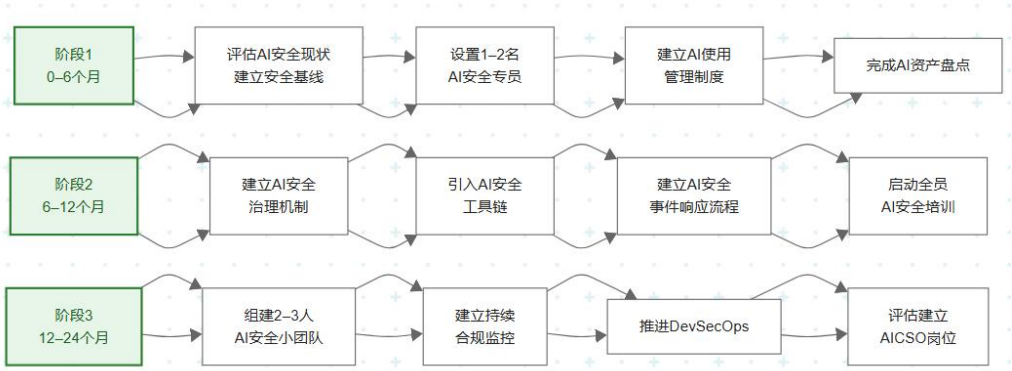


图 9-2 中型企业 AI 安全工作规划

9.1.3 大型企业（500 人以上）

大型企业已进入建立**战略治理 + 团队化运营 + 平台化支撑**的 AI 安全体系阶段。本白皮书建议在 0－6 个月重点完成顶层组织启动、全面评估和战略规划，6－12 个月建立 AI 治理委员会、专职团队、制度体系和工具平台,12－24 个月形成 AI SOC、红队演练、安全评测和合规认证能力，24－36 个月持续优化成熟度、安全文化和行业影响力。因为大型企业的业务规模、发展速度、资源投入、风险程度等因素要求 AI 安全工作开展，无法像中小企业那样慢步小跑，通常是针对重点问题，迅速集结资源完成团队搭建和问题处置，形成常态化运营机制，防范风险再次发生。所以，在 AI 安全的建设规划上，也要跟上业务发展的脚步。如下图 9-3 所示。

建议设置 L4 或 L5 级别的 AI 安全官，并在 AI 技术与攻防安全、AI 工程与架构安全、AI 治理与风险合规三大能力域分别配置 2－3 人。

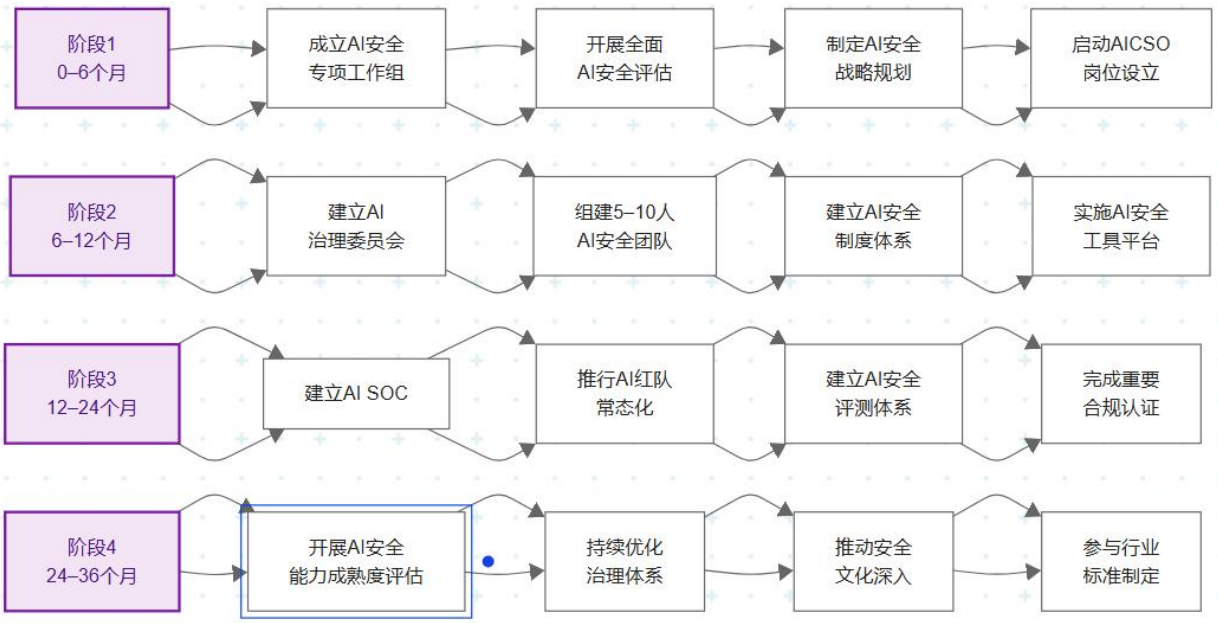


图 9-3 大型企业 AI 安全工作规划

9.2 按行业的建设重点

不同行业业务场景、监管要求不同，在实际工作开展中，因资源投入的有限，存在问题解决优先级的问题，所以 AI 安全建设侧重点也存在明显差异。

本章节梳理了各行业首要关注能力域、业务特殊关注点以及安全控制项建设优先级，供各位 AI 安全官参考，如下图 9-4 所示：

各行业AI安全建设重点			
行业	首要能力域	特殊关注点	建设优先级顺序
金融行业	D3治理合规	算法公平、可解释、监管合规	C07→C08→C01→C10
互联网行业	D1攻防安全	Prompt注入、用户数据、内容安全	C02→C03→C06→C05
医疗健康	D3治理合规	患者数据、诊断责任、伦理审查	C08→C09→C07→C01
自动驾驶	D1+D2技术架构	功能安全、实时性、传感器安全	C01→C04→C03→C11
制造业	D2工程架构	工控安全、OT/IT融合	C04→C05→C06→C10
教育行业	D3治理合规	学生数据、AI伦理、内容安全	C08→C09→C07→C12
政务/央企	D3治理合规	数据合规、国产化、等保	C08→C07→C05→C01

图 9-4 各行业 AI 安全建设重点项

9.3 典型企业案例

随着生成式人工智能从技术验证进入规模化生产应用，企业面对的 AI 安全问题已呈现多领域、多方向的特征。国内各大企业开展 AI 安全实践，构建 AI 安全能力是随着 AI 业务规模扩大逐步演进。下面分别选取大模型研发型企业 A、互联网 AI 应用企业 B 和金融企业 C 三家企业的案例，观察企业如何从 AI 应用问题出发，逐渐建立明确的 AI 安全责任主体、AI 安全团队和治理机制，数据来源均为公开案例。

案例一：大模型研发型企业 A

A 企业属于典型的基础大模型研发和服务企业，其 AI 业务已经覆盖基础模型、面向消费者的 AI 助手、大模型 API、企业级 MaaS 平台和 Agent 开发平台等多个层次。该公司 2025 年第四季度财报显示，截至 2025 年 12 月，AI 助手月活跃用户达到 2.02 亿。2026 年，其基模版本持续迭代、大模型开放平台以及 Agent 产品继续规模化发展，为大量消费者、开发者和企业客户提供的 AI 生产服务。

随着大模型从研发走向开放服务，该公司面对的安全问题也从传统互联网安全扩展到了 AI 全生命周期。训练和精调阶段存在原始语料、标注数据、个人信息和敏感数据泄露风险；模型部署和流转阶段存在模型文件窃取、模型篡改、知识产权泄露和对抗攻击风险；推理和运营阶段则进一步出现提示词注入、越狱、多轮诱导、隐私泄露、偏见歧视、有害内容生成、接口滥用、模型盗爬和投毒反馈等问题。**该公司公开的大模型安全方案，已经明确把这些风险划分到训练、精调、推理、部署和业务运营等不同阶段进行管理。**

面对这些问题，公司逐渐构建了区别于传统基础设施安全的专业 AI 安全力量。公开资料显示，除了原有负责集团业务安全、业务风控和大模型安全解决方案的传统安全团队外，该公司内部还存在负责大模型数据安全、模型保护等方向的专业人员。围绕大模型底座，该公司进一步形成了大模型内容安全、数据安全、模型保护、业务风控和蓝军评测等专业能力。

上述情况，已说明基模企业已经开始把 AI 安全从传统安全团队中的零散任务，演进为由明确负责人牵头、多个专业安全方向共同参与的体系化工作。

该公司在 AI 安全建设上，首先把安全能力前移到模型生产流程。训练数据进入模型之前，需要进行数据集质量评估、个人和敏感信息处理以及违规内容清洗；精调阶段通过安全指令微调和 RLHF 等技术增强模型安全对齐；模型流转和部署阶段则通过零信任、模型文件保护等机制降低核心模型与数据泄露风险。与此同时，构建输入输出安全护栏、

Prompt 攻击识别、多模态内容检测、红线知识库和安全代答等能力，部署到推理与业务运营阶段。

AI 安全测试工作在该公司是开始常态化开展的，公司构建 AIGC 内容安全蓝军评测能力，通过攻击指令、自动化评测数据集、人工复核等机制，持续验证模型在恶意诱导、内容安全和绕过攻击下的表现。其公开资料显示，大模型安全评测体系已经能够覆盖 100 余种风险分类和 20 余种高级攻击方法，支持模型研发完成后、正式上线前以及算法备案前的安全评测。

这样的 AI 安全组织架构演进，符合 AI 安全官体系中**大型 AI 原生企业的 AI 安全组织模式**。当模型本身就是企业核心产品时，仅依靠传统网络安全负责人已经难以覆盖数据安全、模型安全、内容安全、红队评测、AI 应用安全和监管合规等多方面、多领域风险面时，企业需要明确 AI 安全负责人，形成模型安全、AI 红队/蓝军、安全工程、数据安全和内容安全等多角色的安全团队，使安全工作真正融入模型设计、训练、评测、发布和运营全过程。

案例二：互联网 AI 应用企业 B

与基模企业不同，大多数互联网企业不会自行训练通用基础模型，大多数是通过外部模型 API 调用、私有化部署自有垂直模型和多模型平台组合，构建 AI 应用。互联网 AI 应用企业 B 公司，是这一模式的典型案例。

该公司借用外部 AI 能力，把大模型能力嵌入文字、表格、演示、知识库、智能搜索和办公 Agent 等大量场景，通过多模型调度机制，适配不同场景，选择不同的模型完成不同任务。2025 年末，该公司 AI 应用国内月活用户已经超过 8,013 万，同比增长 307%，AI 服务日均 Token 调用量超过 2,000 亿，同比增长超过 12 倍，大模型调用已经成为其产品中的核心基础能力。

当大模型进入其业务场景后，其典型的安全问题是：用户可能把合同、财务数据、客户资料、会议记录、内部制度和企业知识库交给 AI 处理，如果模型调用、权限继承或数据流转机制处理不当，就可能产生敏感数据外泄、跨用户越权访问、第三方模型获取企业数据等问题。同时，多模型接入还带来了不同模型安全能力不一致、模型供应商管理、Prompt 攻击、生成内容合规和模型调用审计等问题。

另一方面，随着产品能力从 AI 助手进入能够执行任务的智能体阶段，工具调用、权限控制、执行隔离和 Agent 行为可观测性也开始成为新的安全边界。该公司 2025 年年报

明确指出，安全沙箱、执行隔离、可观测性、数据安全和隐私计算是其 AI 产品的重要安全要求。

该公司的安全组织变化，符合传统安全团队的 AI 演进变化。公开资料显示，公司已经建立**董事会战略委员会—安全委员会/隐私保护委员会—安全中心和安全工作组**的三级安全管理架构，其中安全委员会负责网络安全和数据安全，安全中心和安全工作组承担安全规范、审计和具体安全能力建设。AI 安全仍没有形成独立的新部门，是在原有企业安全组织基础上，由安全负责人和安全技术团队逐步扩展职责范围。

公开资料显示，该公司针对大模型应用实施全生命周期数据安全管控，覆盖模型训练和上线后的推理服务。在工程层面，公司建立 AI 网关平台统一管理不同模型，引入实时审核机制进行内容安全控制，针对不同产品和业务场景建设 Prompt 管理平台，通过效果评测和参数调优降低不同模型在实际业务中的不确定性。

对于此类企业，本身不是通用基础模型厂商，通过大模型 API 调用正在成为大规模 AI 应用运营者。该类企业通常不需要像基模厂商那样，构建一定规模的 AI 安全团队，但必须明确 AI 安全责任人，由安全架构、应用安全、数据安全和 AI 产品研发人员共同承担 AI 风险管理工作。其典型建设路线是从传统安全团队支持 AI 项目开始，逐渐建立 AI 应用准入、多模型安全管理、AI 网关、数据出域控制、知识库权限、Prompt 安全、Agent 工具调用控制和持续评测等能力，最终形成专门的 AI 安全负责人或 AI 安全小组。

案例三：金融企业 C

企业 C 是 AI 技术深入核心生产业务的金融集团公司，其开展的金融业务属于流程强监管类型。公开数据显示，该公司已经把 AI 应用到客服、核保、理赔、反欺诈、风险管理、销售和内部工作流等大量金融业务。截止 2025 年度，已有超过 23 万名员工使用 AI Agent 平台，累计开发超过 7 万个 AI Agent 应用，全年模型调用量达到 36.5 亿次，AI 客服全年提供约 17.02 亿次服务，占集团客服总量约 80%。

该公司需解决的不仅是 AI 安全问题，还有 AI 治理和合规问题。例如，AI 在使用过程中，可能接触身份、账户、保单、健康、信贷和交易等高度敏感的数据。大模型和智能体生成错误信息，可能影响客服、核保、理赔和风险判断。大模型算法的偏差，可能造成不公平结果。未经批准的第三方模型调用，可能造成敏感数据泄露。内部 Agent 工具权限过大、错误调用、不可追溯的 Agent 自主行为等问题，对于金融机构而言，同时涉及信息安全、数据安全、模型风险、隐私保护、监管合规和 AI 伦理，因此很难由传统 IT 安全团队单独解决。

该公司采取的是典型的分层治理模式。集团成立 AI 伦理委员会，负责人工智能伦理政策方向以及 AI 开发和应用中的伦理治理，并进一步建立算法安全管理委员会、科技伦理审查委员会等组织，对人工智能服务实施持续监督。同时，集团原有的信息安全体系继续承担数据安全、个人信息保护、第三方安全和技术风险等职责，形成 AI 治理组织与专业安全团队协同工作的模式。

从职责划分来看，AI 治理层负责制定原则、政策和重大风险边界；信息安全和数据安全人员负责敏感数据保护、访问控制、漏洞和安全事件；算法及模型相关人员负责模型风险、安全评估、算法治理；业务、法务、合规及风险管理部门，则负责 AI 应用是否符合金融业务规则和监管要求。

在 AI 治理方面，其公开披露通过内部制度固化治理要求。例如，制定了《生成式人工智能服务安全管理办法》，并将 AI 治理从数据使用、算法研发和行业应用三个维度进行管理。集团同时建立以董事会为最高管理层的信息安全管理架构，并持续开展信息安全审计、演练和风险监测。

这种采用 AI 治理委员会 + AI 安全负责人 + 专业安全/模型/数据/合规团队的矩阵化组织架构，更适合**大型强监管企业 AI 安全治理模式的演进，**AI 安全官或 AI 安全负责人位于其中的连接位置，既不取代 CISO、数据安全、模型风险或合规负责人，也不只是技术专家，通过统一 AI 资产、风险分级、上线安全评估、跨部门协调、重大风险升级以及持续监督，从而把分散在多个部门的 AI 风险管理能力形成闭环。

上述三个案例共同说明，**AI 安全人员和团队不是企业为了新增一个安全职位而设置的，而是在 AI 业务规模达到一定程度之后，根据实际风险和治理复杂度推动形成的专业职能。** 对基模企业而言，这种职能会发展为 AI 安全专业团队；对普通 AI 应用企业而言，可能首先表现为 AI 安全负责人或小型跨职能团队；对金融等强监管企业而言，则更容易形成治理委员会、AI 安全责任人和传统风险、安全、合规体系共同参与的矩阵式治理架构。上述这些组织的真实 AI 安全演进，与本白皮书的观察情况、编制宗旨是一致的。

第 10 章 中国及欧美 AI 监管要求映射

特别说明：法规和标准具有时效性，本白皮书提供的是能力与控制框架，本章按 2026 年 8 月公开数据整理而来，考虑 AI 监管变化较快，企业在上线发布、备案、出海或重大风险决策前，应再次联系主管机关，以主管机关、标准发布机构和适用司法辖区的最新正式文本为准，本章内容不构成权威法律意见。

10.1 AI 监管合规框架概览

AI 监管合规目前在全球内主要为**中国、欧盟、美国**三种模式。

中国目前形成基础法律、网络与数据行政法规、算法/深度合成/生成式 AI 专项服务规则，以及国家标准并行的体系，整体比较完善，以监管治理促进 AI 技术发展为导向。欧盟以《欧盟人工智能法案》为统一横向立法，与 GDPR、产品安全、网络安全和行业规则衔接，更多是不断优化完善为导向。美国缺少与《欧盟人工智能法案》完全对应的统一联邦 AI 法，更多依赖联邦行政政策、部门监管、州级立法以及 NIST 等自愿风险管理框架，总体以管理控制防范风险为导向。

本章的内容，从各国监管要求出发，逐步建立**法律规范-AICSO 能力模型-AI 合规要求**的合规映射，帮助 AI 安全官建立合规能力。

10.2 中国 AI 监管体系

中国在 AI 监管领域已经走在了全球前列，它依托网络安全、数据安全、个人信息保护、算法治理、深度合成、生成式 AI 服务、内容标识等多领域制度规范。它们之间相互衔接、互补支撑，共同构筑全覆盖的 AI 监管网络。对于在国内开展 AI 业务的企业来说，合规不仅是法定义务，更是筑牢人工智能安全治理体系的核心基础。

作为企业，为实现合规工作标准化落地，需厘清国内 AI 监管的层级体系，将各级监管要求精准映射至内部 AI 安全治理能力，以实现外部监管规则与内部管控体系的深度对齐。

10.2.1 AI 安全监管相关法律法规

结合现行法律法规、行政法规、专项规范及国家标准，与企业 AI 安全合规高度相关的监管依据可划分为七大层级，覆盖基础法律底线、专项监管要求与技术落地标准。如下表 10-1 所示：

表 10-1 中国 AI 安全监管相关法律法规

层级	主要法律、规则或标准	核心合规意义
基础法律	《网络安全法》《数据安全法》《个人信息保护法》等	确立网络安全、数据安全、个人信息保护、风险监测的基础监管义务。2025 年修订的《网络安全法》新增人工智能发展、伦理规范、风险监测评估及专项安全监管内容，构成国内 AI 合规的底层法律框架。
行政法规	《网络数据安全管理条例》	统一规范网络数据全流程处理、委托与共同处理、个人信息保护、重要数据管控及平台主体责任，自 2025 年 1 月 1 日起施行，是数据类 AI 业务的核心合规依据。
算法与深度合成专项规则	《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》	确立算法治理核心规则，明确具备舆论属性或社会动员能力的算法服务备案要求，压实深度合成服务安全主体责任，规范内容标识、风险防控等核心工作。
生成式人工智能专项规则	《生成式人工智能服务管理暂行办法》	针对面向境内公众提供的生成式人工智能服务，明确内容安全、数据合规、用户权益保护、安全评估、备案管理等专项监管要求。截至 2026 年 8 月，该办法仍以暂行规章形式现行有效。
AI 内容标识专项规则	《人工智能生成合成内容标识办法》、GB 45438-2025	规范 AI 生成内容的显性标识、隐性标识、元数据留存、平台核验、日志留存及防篡改管理，自 2025 年 9 月 1 日起实施，形成管理办法＋强制性国家标准的闭环管控体系。
场景型 AI 专项规则	《人工智能拟人化互动服务管理暂行办法》	针对情感互动类拟人人工智能服务，建立全生命周期安全管控体系，明确安全评估、算法备案、用户权益保护等专项要求，自 2026 年 7 月 15 日起施行。
国家级技术标准	GB/T 45654-2025、GB/T 45652-2025 等	明确生成式人工智能服务安全、训练数据安全的技术基线，细化可落地的技术管控标准，为企业 AI 安全体系搭建、合规自查、第三方评估提供权威技术依据。

这些法规之间相互关联，共同构成了国内 AI 监管的完整网络。企业在合规实践中，需要将这些法规要求与 AI 安全官的能力框架进行映射，构建企业 AI 合规的外部约束体系，才能确保安全治理工作全面覆盖法规要求。

而作为企业的 AI 安全负责人，AI 安全官开展合规治理落地的前提是具备相应的专业能力。下表 10-2 所示，将 AI 安全官能力模型与上述监管要求做映射，形成可执行、可核查的安全管控能力，以保障个人能力与企业需求在安全合规上的深度融合。

表 10-2 AI 安全官能力模型与中国 AI 安全监管法律法规映射关系

AICSO-CF 能力	映射监管要求	合规要点
C08 AI 法规与合规	《生成式人工智能服务管理暂行办法》、AI 备案登记及内容标识相关制度	落实算法备案、生成式人工智能服务备案、应用功能登记、安全评估报告编制及 AI 生成内容标识，持续跟进监管规则更新迭代。需注意，大模型上线备案并非法定术语，企业内部文档应规范使用官方合规概念。
C01 算法与数据安全	《数据安全法》《个人信息保护法》《网络数据安全管理条例》、GB/T 45652-2025	落实数据分级分类、数据脱敏、跨境传输安全评估，规范 AI 训练数据集采集、处理、存储全流程安全管控，严控数据委托处理与共同处理合规风险。
C09 AI 伦理与知识产权	《深度合成管理规定》《人工智能生成合成内容标识办法》	规范 AI 生成内容标识管理，落实深度合成服务合规要求，管控 AI 训练素材知识产权风险，坚守 AI 应用伦理底线，杜绝违规生成与侵权使用。
C07 AI 治理体系建设	《算法推荐管理规定》及各类 AI 专项管理办法	保障算法运行透明度，完善用户权益保护机制，搭建覆盖全业务场景的 AI 全生命周期治理制度体系，明确各岗位安全责任矩阵，落实企业主体责任。
C10 AI 业务风险评估	《生成式人工智能服务管理暂行办法》《人工智能拟人化互动服务管理暂行办法》	常态化开展 AI 安全评估，建立风险分类分级管理机制，针对业务场景变更、技术迭代、用户规模扩容、功能新增等情形及时开展专项风险复评。
C11 AI 安全事件管理	《数据安全法》《个人信息保护法》及 AI 安全监管相关要求	建立 AI 安全事件应急处置机制，落实事件排查、上报、数据泄露通知、事后整改优化全流程工作，形成事件闭环管理。
C02 AI 对抗防御	《生成式人工智能服务管理暂行办法》、GB/T 45654-2025	强化 AI 内容安全管控，防范模型滥用、恶意攻击、对抗性干扰及违规生成等风险，保障模型运行稳定性、安全性与合规性。
C06 AI 安全运营	《数据安全法》《个人信息保护法》、GB 45438-2025	落实日志审计、全网安全监控、实时事件响应，常态化执行 AI 内容标识、溯源核验、防篡改管控等运营工作，保障合规治理常态化运行。

10.2.2 大模型备案/登记、安全评估与算法备案

当前国内市场上，从事 AI 合规企业服务的律所和公司很多，行业普遍将各类 AI 合规备案笼统称为大模型备案，但四类合规制度的适用场景、触发条件与监管主体各不相同，需独立判定、分别落实。故在本白皮书中，为了避免使用大模型备案这类泛化术语，根据服务形态和国家主管部门要求，分别使用法律法规定义的标准术语，让 AICSO 了解其内容和本质。

大模型备案和算法备案是中国 AI 监管的两项核心制度，也是企业合规落地的第一道门槛。

大模型备案的要求是，根据《生成式人工智能服务管理暂行办法》，面向公众提供生成式 AI 服务需要完成安全评估并取得评估报告、建立内容安全审核机制、实施用户协议和隐私政策、建立投诉举报机制，以及对生成内容进行标识。这些要求覆盖了从模型上线前的安全评估，到上线后的持续监管，形成了完整的合规闭环。

根据国家网信办公开公告显示，截至 2026 年 6 月 30 日，累计已有 988 款生成式人工智能服务完成备案。如下图 10-1 所示：



图 10-1 国家互联网信息办公室发布的生成式人工智能服务已备案信息公告

对于企业而言，理解备案的具体要求和流程，将备案工作纳入 AI 系统的上线流程，是确保合规的关键。

算法备案是指，根据《互联网信息服务算法推荐管理规定》，具有舆论属性或者社会动员能力的算法推荐服务提供者应当履行备案手续。备案内容包括算法基本信息（名称、类型、应用场景）、算法安全自评估报告、算法运行机制说明，以及用户权益保护措施。对于生成式 AI 和拟人化互动服务，也会触发相关算法备案义务。如下图 10-2 所示：



图 10-2 国家互联网信息办公室发布的互联网信息服务算法备案信息公告

生成式人工智能应用或功能登记，指通过 API 接口，或其他方式直接调用已备案模型能力的应用或功能模块，由地方网信部门开展登记。截至 2026 年 6 月 30 日，公开数据为累计 598 款应用或功能完成登记，已上线应用或功能还应按主管机关要求公示所使用的已备案或登记生成式 AI 服务信息。

AI 应用专项安全评估是指开展 AI 业务的企业，应根据具体服务类型、影响范围、相关专项规则及主管机关要求，开展安全评估工作。例如，《人工智能拟人化互动服务管理暂行办法》明确列举上线或增设拟人化互动功能、重大技术变化、达到特定用户规模，以及存在影响国家安全或公共利益风险等情形时，需要开展专项安全评估工作，具体执行时可以由企业开展自评估，也可以邀请第三方机构开展评估。

因此，企业 AI 业务上线流程，应独立判断是否触发算法备案、是否触发生成式 AI 服务备案、是否属于调用已备案模型后的应用/功能登记、是否触发专项安全评估。如果适用，则执行的过程和依据应留存证据，并由法务/合规与 AI 安全官共同复核，以存档备查。重要材料若须提交给国家部委、省、市级监管部门的，要及时提交。

10.3 ISO 4200 与中国法规的对应关系

ISO/IEC 42001 作为 AI 管理体系的国际标准，为中国企业提供了与国际接轨的治理框架。本白皮书中，将 ISO 42001 的条款要求与中国 AI 法规进行映射，帮助企业和 AI 安全官了解国际标准和法规要求，满足监管要点，避免重复建设。

下表展示了 ISO/IEC 42001 主要条款与 AI 安全官能力模型、与中国法规要求的对应关系：

表 10-3 AI 安全官能力模型与 ISO/IEC 42001 映射关系

ISO 42001 条款	对应 AICSO-CF 能力	中国法规要求
4.1 理解组织及其环境	C10 业务风险评估	开展 AI 安全专项评估，建立风险分类分级管控体系
5.1 领导力和承诺	C07 治理体系建设	搭建 AI 治理专项委员会，明确各岗位安全责任矩阵
6.1 应对风险和机遇的措施	C10 业务风险评估	常态化开展风险排查评估，落实残余风险管控与整改
7.1 资源	C12 安全沟通与培训	完善 AI 安全团队建设，搭建常态化安全培训体系
8.1 运行规划和控制	C06 安全运营	落实全流程日志审计、常态化安全监控管控
9.1 监视、测量、分析和评价	C06 安全运营	持续开展合规效果监测、风险度量与体系优化
10.1 不符合和纠正措施	C11 安全事件管理	规范安全事件响应处置，落实事后整改与长效优化

10.4 欧美 AI 安全治理

欧盟与美国是全球另外两大 AI 监管的核心阵营，其中欧盟已建立统一、强制、风险导向的专门立法体系，监管约束性强、规则体系完备；美国未出台联邦统一的人工智能专项法律，合规约束由联邦行业法规、地方州级立法及自愿性治理框架共同构成，呈现碎片化、场景化的监管特征。

欧美各类 AI 监管规则和国内一样，同样相互交织、叠加适用，不存在孤立的单一合规条款。企业开展出海 AI 业务时，要精准区分欧盟与美国联邦、各州的规则适用边界，识别不同业务场景对应的合规义务，并将差异化的域外监管要求转化为内部标准化的安全管控措施。

10.4.1 欧美 AI 监管规则层级体系

结合欧美现行生效的法律法规、管理要求、技术标准与地方立法，企业开展跨境 AI 业务时，对与安全合规高度相关的监管依据，按照层级分类梳理，如下表 10-4 所示。

表 10-4 欧美 AI 监管法律法规层级关系

辖区	层级	主要法律、规则或标准	核心合规意义
欧盟	核心专项立法	《人工智能法案》	欧盟统一人工智能专项强制立法，以风险分级为核心逻辑，将 AI 系统划分为不可接受风险、高风险、有限风险、最低风险四个等级，具备域外效力，设置高额处罚机制，是企业布局欧盟市场的核心合规依据。
欧盟	基础配套立法	《通用数据保护条例》《数字	提供数据安全、个人信息保护、网络内容治理的基础合规底线，与

		服务法案》	《人工智能法案》叠加适用，全面约束 AI 训练、数据处理、内容生成全流程。
欧盟	技术配套标准	欧盟协调标准体系	为高风险 AI 系统提供合规评估、技术验证的标准依据，是企业完成符合性认证、证明合规有效性的核心技术支撑。
美国	联邦层面规则	联邦行业监管法律、《国家人工智能风险管理框架》	无联邦统一 AI 强制立法，《国家人工智能风险管理框架》为自愿性治理指导文件，无法律强制效力；合规约束主要依托现有网络安全、隐私保护、消费者权益、反歧视等行业法律落地。
美国	州级专项立法	科罗拉多州 AI 法案、纽约市 AI 招聘监管规则等	各州出台差异化强制 AI 合规规则，覆盖风险评估、透明度、反歧视、用户权益保护等要求，是美国市场最核心的强制性合规来源。
美国	行业专项监管	医疗 AI 监管、就业反歧视监管、金融行业风控规则	针对医疗、金融、招聘等特殊场景，对 AI 系统的模型使用、决策逻辑、风险防控、数据合规提出专项强制约束，场景适配性要求极高。

10.4.2 AI 安全官能力模型与欧美监管要求映射

因为法规法条的逐项罗列，无法直接应用于 AI 安全官工作，需要做进一步的归纳提炼，翻译为业务可理解的语言，所以本白皮书中，把欧盟、美国各层级监管要求与 AI 安全官能力模型的四大能力域做映射，让企业能够进一步形成责任人、控制措施、证据和持续监测机制。详细如下表 10-6 所示。

表 10-5 AI 安全官能力模型与欧美监管要求之间映射关系

AI 安全官能力模型	欧盟合规重点	美国合规重点
C01 算法与数据安全	落实 AI 训练数据合规治理、版权管控、个人信息保护，衔接《通用数据保护条例》要求，严控跨境数据传输风险，保障数据处理全流程合规。	依托行业数据安全规则与各州隐私立法，规范 AI 数据采集、使用、存储行为，防范数据泄露、隐私侵权及算法歧视风险。
C02 AI 对抗防御	强化高风险 AI 系统鲁棒性与网络安全防护，防范通用大模型系统性风险、恶意滥用与安全攻击，落实模型安全管控义务。	参照行业 AI 安全防护实践，搭建模型对抗攻击、防滥用、防违规生成的防护体系，适配行业安全监管要求。
C03 AI 红队测试	针对高风险 AI 系统、通用大模型开展常态化合规测试、风险验证与漏洞排查，留存完整测试报告用于合规核查。	依托国际评测标准与行业实践，常态化开展 AI 安全测试、风险评估与偏见检测，规避合规与伦理风险。
C05 AI 供应链安全	厘清 AI 产业链各主体合规责任，落实模型服务商、部署企业、下游使用者的责任划分，完善下游信息同步与风险管控机制。	强化第三方模型与接口服务的供应链管控，通过合同约束落实外部供应商安全责任，防范供应链合规风险。
C06 AI 安全运营	建立 AI 系统上市后持续监控、日志留存、安全事件上报、事后整改的全流程运营机制，满足欧盟常态化监管要求。	搭建全天候安全监控、事件响应、闭环整改的运营体系，适配美国行业监管与州级合规核查要求。
C07 AI 治理体系建	搭建适配欧盟风险分级的 AI 治理架构，完善风险管	依托国际治理框架，优化企业内部 AI 治理制

设	理、质量管理、全生命周期管控制度，明确岗位责任体系。	度，适配美国多区域、多行业的碎片化监管场景。
C08 AI 法规与合规	精准识别业务角色与 AI 风险等级，严格对标欧盟分阶段合规时间表，落实通用大模型与高风险系统专项合规义务。	持续跟踪联邦政策、各州 AI 立法、行业监管规则迭代，动态调整企业合规策略，区分强制规则与自愿指导框架。
C09 AI 伦理与知识产权	落实 AI 透明度、基本权利保障、版权管控要求，杜绝算法歧视、违规生成、知识产权侵权问题。	遵循消费者保护、反歧视、知识产权规则，完善企业 AI 伦理规范，规避商业与伦理合规风险。
C10 AI 业务风险评估	严格落实 AI 产品风险分级判定、业务角色识别，针对高风险场景开展专项风险评估与持续管控。	依托风险管理框架，常态化开展业务影响分析与风险评估，适配不同州、不同行业的风险管控要求。
C11 AI 安全事件管理	建立重大 AI 安全事件上报、复盘整改、长效管控机制，满足欧盟事后监管与持续合规要求。	遵循行业事件报告规范，完善应急响应、事件处置、闭环整改体系，适配属地监管核查。
C12 安全沟通与培训	强化全员 AI 合规认知与风险防控能力，落实欧盟组织层面的合规培训与能力建设要求。	搭建分层级 AI 安全培训体系，提升团队跨境合规、风险识别与应急处置能力。

企业在实际落地时，可以根据行业、服务形态和业务地域继续拆解，把每条法律义务进一步拆分到责任人、控制措施、证据、频率、例外与升级路径，以形成可审计的合规控制矩阵，以便跟踪、推进、回顾各项工作的完成情况，防止丢失或遗漏。

第 11 章 版本演进与迭代说明

11.1 版本管理

《AICSO AI 安全官白皮书》由 AICSO 人工智能安全官社区 aicso.ai 运营发布，其版本迭代采用主版本号.次版本号的版本管理策略，以确保其框架快速响应 AI 技术变化。

主版本号更新（Major Version） 涉及框架核心结构的调整，包括能力域划分、核心能力定义或等级体系的根本性变化。这类更新，预计每 3 到 5 年发生一次，需要经过完整的研究、征求意见和评审流程。主版本号更新时，会发布迁移指南，帮助从业者从旧版本平滑过渡到新版本。

次版本号更新（Minor Version） 涉及框架内容的补充和优化，包括新增能力项的细化描述、新兴威胁的纳入、工具推荐的更新、案例的补充等。这类更新预计每年更新 1 到 2 次，由 AICSO 人工智能安全官社区通过增量发布的方式向从业者推送。

本次发行的版本为 V1.0 版本，全称为《AICSO AI 安全官白皮书 V1.0》。

11.2 问题清单

本白皮书编写时间较长，因为 AI 技术的快速更新或前期框架设计考虑不足，在编写和专家评审过程中，陆续发现一些问题，我们统一记录在这里，作为 V2.0 版本考虑的内容，也方便反馈意见的专家、同行对照参考，避免重复反馈。

表 11-1 已知待解决问题清单

序号	问题描述
1	当前的 AI 安全中，对大模型应用带来的 AI 伦理安全、炸药制造、人身安全、智能驾驶安全等安全内容未覆盖，是否一并纳入 AI Safety 中去考虑？
2	AI 安全官的英文拼写，国内和国外都有英文简称，且不一致，具体使用哪个比较好？暂时不作考虑，统一使用中文版。

11.3 社区生态

本白皮书的内容，主要来源于国内互联网企业与人工智能企业的一线实践。为了维护白皮书内容的正确性和材料的可信度，人工智能安全官社区 aicso.ai 建立公开 issue/征求意见渠道、专家评审记录、版本变更日志、引用来源清单和企业实践反馈机制，官方联络邮箱为 hello@aicso.ai。

涉及行业数据时，应优先采用监管机构、标准组织、国内人工智能企业以及安全厂商的公开报告数据。无法验证来源的比例、排名和数据不进入正式白皮书内容。

基于此白皮书的 AI 安全官能力框架下涉及的 12 项核心能力的具体技术点，由使用者自行构建相关技术实践，来填充 12 项核心能力的参考案例。我们建议，细化案例时以真实项目、可成果化、紧跟 AI 技术前沿为核心选取依据，来构建能力学习图谱和知识路径。

第 12 章 附录

附录 A：AI 风险评估检查表

本检查表提供了 AI 系统安全评估的核心检查项，涵盖数据安全、模型安全、基础设施安全和治理合规四个维度。企业可根据自身 AI 应用场景和风险偏好，对检查项进行优先级排序和定制化调整。建议将本检查表纳入 AI 系统上线前的必检流程，并定期复查以确保持续合规。

A.1 数据安全检查项

编号	检查项	等级	状态
D-01	训练数据来源是否可追溯？	高	☐
D-02	训练数据是否经过投毒检测？	高	☐
D-03	敏感数据是否在训练前脱敏？	高	☐
D-04	是否实施差分隐私或联邦学习？	中	☐
D-05	数据保留策略是否明确？	中	☐
D-06	推理输入/输出数据是否加密传输？	高	☐

A.2 模型安全检查项

编号	检查项	等级	状态
M-01	模型文件完整性是否验证？	高	☐
M-02	是否进行模型安全评测？	高	☐
M-03	是否部署 Guardrails？	高	☐
M-04	模型回滚机制是否就绪？	中	☐
M-05	是否进行 Red Teaming 测试？	高	☐
M-06	模型版本管理是否完善？	中	☐

A.3 基础设施安全检查项

编号	检查项	等级	状态
I-01	GPU 集群访问控制是否完善？	高	☐
I-02	推理 API 是否认证授权？	高	☐

编号	检查项	等级	状态
I-03	容器镜像是否安全扫描？	中	☐
I-04	API 速率限制是否配置？	中	☐
I-05	日志审计是否开启？	高	☐
I-06	是否有异常行为监控？	中	☐

A.4 治理合规检查项

编号	检查项	等级	状态
G-01	是否建立 AI 治理制度？	高	☐
G-02	是否完成算法备案？	高	☐
G-03	AI 资产是否台账管理？	中	☐
G-04	是否开展 AI 安全培训？	中	☐
G-05	是否有 AI 安全事件应急预案？	高	☐
G-06	是否进行定期 AI 安全评估？	中	☐

A.5 智能体安全检查项

编号	检查项	等级	状态
AS-01	Agent 是否具有唯一身份标识，并使用短期、最小权限凭据进行鉴权	高	☐
AS-02	是否定义 Agent 可调用工具白名单及调用参数校验规则	高	☐
AS-03	高风险工具调用是否需要二次授权或人工确认	高	☐
AS-04	MCP/A2A/第三方工具是否经过来源、权限和更新安全评估	高	☐
AS-05	是否区分用户输入、外部检索内容和系统指令等不同信任级别，并采取差异化处理策略	高	☐
AS-06	长期记忆写入是否进行来源校验，是否设置敏感信息过滤与定期清理机制	中	☐
AS-07	是否完整记录 Agent 执行轨迹、工具调用日志与授权决策记录	高	☐
AS-08	是否设置最大任务步数、调用预算、请求速率与成本上限	中	☐
AS-09	Agent 执行代码/命令时是否使用隔离环境和最小网络权限	高	☐
AS-10	是否具备紧急终止、凭据吊销、工具隔离与记忆隔离等应急控制机制	中	☐
AS-11	多 Agent 间通信，是否进行身份认证、消息完整性校验和信任边界控制	中	☐
AS-12	是否针对 Agent 目标劫持、工具滥用、记忆投毒等攻击场景开展红队测试	高	☐

附录 B：AI 安全能力自测问卷

本问卷帮助从业者快速评估自身在 AI 安全领域的综合能力水平。使用方法是：根据实际能力水平为每项能力打分（1-5 分），然后对照等级要求判断自身所处阶段。自测结果应作为个人发展计划的起点，而非终点。

编号	核心能力	自评分(1-5)	能力域
C01	算法与数据安全	___	D1
C02	AI 对抗防御	___	D1
C03	AI 红队测试	___	D1
C04	AI 基础设施安全	___	D2
C05	AI 供应链安全	___	D2
C06	AI 安全运营	___	D2
C07	AI 治理体系建设	___	D3
C08	AI 法规与合规	___	D3
C09	AI 伦理与知识产权	___	D3
C10	AI 业务风险评估	___	D4
C11	AI 安全事件管理	___	D4
C12	安全沟通与培训	___	D4

结果解读：

你的等级	判断标准
L1	12 项能力均具备基础认知；至少 6 项能够完成标准化检查、测试或记录任务，并能正确升级复杂问题
L2	至少 6 项达到 L2，其余不低于 L1；D1 或 D2 至少 1 项达到 L2 以上，并有独立方案设计与真实项目证据
L3	至少 8 项达到 L2 以上，其中 D1 和 D2 各至少 1 项达到 L3；具有企业级架构、标准或平台化成果
L4	12 项均达到 L2 以上，至少 4 项达到 L3 以上，D3 和 D4 至少各 1 项达到 L4；具有企业治理、团队领导和重大风险决策证据
L5	12 项均达到 L3 以上，至少 6 项达到 L4，至少 2 项达到 L5；具有企业级战略、最高管理层/监管沟通及重大风险治理证据

结果应用：采用补短板方式，识别分数最低的 3 项能力，将其作为未来 12 个月的重点发展方向，不建议所有的不满足同时做提升方案。

附录 C：AICSO 术语表

本术语表收录了本白皮书中使用的核心术语及其定义，旨在为读者提供统一的理解基准。

类别	术语	定义

AICSO 体系	**AI 安全官**	负责组织 AI 安全战略、治理、风险、架构、合规、运营与事件响应等工作的管理与技术复合型角色，推动 AI 风险在组织治理和 AI 全生命周期中得到识别、评估、控制和持续监督。
AICSO 体系	**AICSO-CF**	AICSO AI 安全官能力框架（Competency Framework），用于定义 AI 安全官及相关 AI 安全岗位所需的能力域、能力项、知识、技能、实践任务和能力等级，为人才培养、课程设计、能力评价和职业发展提供统一基准。
基础概念	**AI 安全**	围绕 AI 系统及其全生命周期，通过治理、管理和技术措施识别、评估和控制由模型、数据、应用、Agent、工具、供应链及运行环境产生的安全风险，保护 AI 系统及相关资产免受攻击、滥用、操纵、泄露和其他非预期行为影响的安全实践。
基础概念	**AI 系统**	基于机器学习、生成式 AI 或其他人工智能技术，对输入进行处理并产生预测、内容、建议、决策或行动的技术系统，可包含模型、数据、应用、Agent、工具、基础设施及相关运行组件。
基础概念	**AI 应用**	将 AI 模型与业务逻辑、数据、知识库、工具或外部系统组合，用于实现特定业务功能的软件应用或服务。
基础概念	**AI Agent / 智能体**	能够基于目标和上下文进行推理或规划，并通过模型、记忆、工具及外部环境执行一个或多个动作以完成任务的 AI 系统或软件实体。
基础概念	**AI 资产**	组织在 AI 系统研发、训练、部署、运行和管理过程中使用或形成的模型、数据、Prompt、知识库、Agent、工具、接口、软件组件、算力资源及其他具有业务或安全价值的资源。
基础概念	**AI 全生命周期**	AI 系统从规划设计、数据准备、模型开发或选型、开发集成、测试评估、部署上线、运行维护到下线退役的完整生命周期。
治理与风险	**AI 治理**	组织用于指导、监督和控制 AI 研发、采购、部署、使用和退役的制度、职责、流程和控制体系，覆盖责任、风险、合规、安全、隐私、公平、透明等治理要求。
治理与风险	**AI 安全治理**	将安全目标、责任、制度、风险管理和控制措施嵌入 AI 全生命周期，对 AI 相关安全风险进行持续识别、评估、处置、监测和改进的组织治理活动。
治理与风险	**AI 风险**	AI 系统的设计、数据、模型、使用方式、外部依赖或运行行为对组织、用户、社会或其他利益相关方造成不利影响的可能性及其后果。
治理与风险	**AI 攻击面**	AI 系统中可能被攻击者利用的入口、接口、组件、数据流、信任边界和外部依赖的集合，包括模型接口、Prompt、知识库、Agent、工具、API、MCP Server 及供应链组件等。
治理与风险	**AI 安全控制**	为预防、检测、限制、响应或恢复 AI 安全风险而实施的管理、流程或技术措施，包括身份认证、授权、数据保护、安全护栏、监控审计、运行时控制和事件响应等。
治理与风险	**AI 安全评估**	通过资产识别、风险分析、架构审查、安全测试、控制验证和合规检查等活动，对 AI 系统的安全风险及安全控制有效性进行系统性评价的过程。
治理与风险	**AI 威胁建模**	对 AI 系统的资产、架构、数据流、信任边界、攻击者、威胁及攻击路径进行系统分析，并据此识别安全风险和确定安全控制措施的工程方法。
治理与风险	**AI RMF**	AI Risk Management Framework，通常指 NIST AI 风险管理框架，为组织识别、评估、管理和持续监测 AI 风险提供结构化方法。
架构与工程	**AI 安全架构**	将 AI 安全要求和控制措施系统性嵌入模型、数据、应用、Agent、工具、基础设施及其交互关系形成的整体安全设计，用于支撑 AI 系统的纵深防御和风险控制。
架构与工程	**模型安全**	围绕 AI 模型自身及其训练、微调、部署和推理过程实施的安全活动，关注模型完整性、机密性、可用性以及模型滥用、窃取、操纵和对抗攻击等风险。
架构与工程	**数据安全**	围绕 AI 系统训练、微调、评测、推理和知识检索过程中使用的数据实施分类分级、访问控制、隐私保护、完整性保护、泄露防护和生命周期管理等安全活动。
架构与工程	**Prompt 安全**	围绕系统提示词、用户提示词及其他上下文指令实施的保护措施，以降低提示注入、提示泄露、指令覆盖、敏感信息泄露和非预期模型行为等风险。
架构与工程	**RAG 安全**	针对检索增强生成系统的数据源、知识库、检索过程、权限控制、上下文构造及模型生成过程实施的安全保护，以降低知识库投毒、间接提示注入、越权检索和敏感信息泄露等风险。

架构与工程	**智能体安全**	围绕 AI Agent 的目标、身份、权限、模型、记忆、工具、编排、通信、运行时及供应链实施的全生命周期安全活动，以防止智能体被操纵、越权、滥用或产生非预期行为。
架构与工程	**智能体身份**	用于标识 AI Agent 或 Agent 工作负载，并支撑其认证、授权、凭证管理、访问控制和审计追踪的数字身份及相关机制。
架构与工程	**智能体执行轨迹**	Agent 执行任务过程中形成的规划、推理结果、模型调用、工具调用、状态变化、消息交互和关键决策等可记录行为序列，可用于安全监控、审计、取证和行为分析。
架构与工程	**AI 运行时安全**	在 AI 应用或 Agent 实际运行过程中，对模型交互、上下文、权限、工具调用、数据访问和执行行为进行持续检测、约束和防护的安全机制与实践。
架构与工程	**MCP**	Model Context Protocol，一种用于 AI 应用与外部工具、资源和服务进行标准化连接和上下文交互的开放协议；其安全关注点包括身份认证、授权、工具权限、数据访问、Server 信任和调用链安全。
攻击与测试	**AI 安全测试**	通过自动化或人工方式对 AI 模型、应用、RAG、Agent、工具和相关基础设施实施安全验证，以发现脆弱性、错误配置、不安全行为和安全控制缺陷的测试活动。
攻击与测试	**AI 红队**	通过模拟真实攻击者的目标、技术和攻击路径，对 AI 模型、应用、RAG、Agent、工具及其基础设施进行对抗性测试，以发现可被利用的安全弱点、验证防御有效性并评估实际风险的安全实践。
攻击与测试	**提示注入**	攻击者通过直接输入或外部内容间接向 AI 系统注入恶意或非预期指令，改变模型原有指令层级、上下文理解或执行行为，从而诱导系统执行非预期操作的攻击方式。
攻击与测试	**间接提示注入**	攻击者将恶意指令嵌入网页、文档、邮件、知识库、工具返回结果等外部内容，当 AI 系统读取这些内容时触发恶意指令，从而影响模型或 Agent 行为的提示注入方式。
攻击与测试	**越狱攻击**	通过构造输入、上下文或交互策略规避模型既定安全策略和行为限制，使模型产生原本被限制或禁止的输出或行为的攻击技术。
攻击与测试	**记忆投毒**	攻击者通过未经授权地写入、修改或影响 Agent 的短期或长期记忆，使受污染信息在后续任务中影响 Agent 的判断、规划、工具调用或其他行为的攻击方式。
攻击与测试	**数据投毒**	攻击者通过向训练、微调、评测或检索数据中注入恶意、错误或操纵性数据，影响 AI 模型训练结果、系统判断或后续行为的攻击方式。
攻击与测试	**模型投毒**	攻击者通过篡改模型参数、权重、微调过程或模型供应链，使模型产生攻击者预期的隐藏行为、错误行为或安全弱点的攻击方式。
攻击与测试	**知识库投毒**	攻击者通过向 RAG 知识库或其数据源植入恶意、虚假或操纵性内容，影响检索结果、模型回答或 Agent 后续行为的攻击方式。
攻击与测试	**工具投毒**	攻击者通过恶意或被篡改的工具描述、元数据、返回内容或工具实现影响 Agent 的工具选择、参数生成或执行行为的攻击方式。
攻击与测试	**工具滥用**	AI Agent 在受到攻击、错误指令、权限设计缺陷或自主决策偏差影响时，以非预期、越权或有害方式调用其可访问工具的安全风险。
攻击与测试	**对抗样本**	通过对输入进行特定构造或扰动，使 AI 模型产生错误分类、错误判断或其他攻击者预期行为的输入样本。
攻击与测试	**模型窃取**	攻击者通过模型接口查询、文件获取、供应链入侵或其他方式复制、推断或非法获取目标模型的参数、能力或功能的攻击行为。
攻击与测试	**成员推断**	攻击者通过分析模型输出或行为，推断某个特定数据样本是否被用于模型训练的隐私攻击技术。
攻击与测试	**幻觉**	AI 模型生成与事实、输入上下文或可靠数据来源不一致，但在表达形式上可能具有可信性的错误或虚构内容的现象。
防御与控制	**安全护栏**	部署在用户、模型、知识库、Agent 和工具等交互边界上的安全控制机制，对输入、输出、检索内容、上下文、模型响应以及 Agent 工具调用实施检测、校验、过滤、限制、重写或阻断。

防御与控制	**输入护栏**	对进入 AI 系统的用户输入、Prompt、文件或外部上下文实施安全检测和控制，以识别并限制提示注入、越狱、有害内容、敏感信息和其他风险输入。
防御与控制	**输出护栏**	对模型或 Agent 产生的输出实施安全检测和控制，以降低有害内容、敏感信息泄露、不合规内容或其他高风险结果被返回或执行的可能性。
防御与控制	**最小权限**	仅向用户、Agent、工具或工作负载授予完成其任务所必需的最小访问权限，并限制权限范围、持续时间和可访问资源的安全原则。
防御与控制	**人工介入**	在高风险 AI 决策或 Agent 执行过程中引入人工审核、批准、确认或接管机制，以限制完全自主执行可能产生的重大安全或业务风险。
资产与供应链	**AI 供应链安全**	对 AI 系统所依赖的模型、数据、软件组件、Agent、工具、插件、MCP Server、第三方服务和基础设施进行来源验证、风险评估、完整性保护和持续管理的安全实践。
资产与供应链	**AI BOM**	描述 AI 系统关键组成和依赖关系的结构化清单，可包含模型、数据集、Prompt、Agent、知识库、工具、软件组件、推理服务及关键策略等信息，用于资产管理、供应链风险分析、漏洞影响分析和审计追踪。
资产与供应链	**SBOM**	Software Bill of Materials，描述软件产品所包含组件、库、依赖项及其版本等信息的结构化软件物料清单，用于软件供应链透明度、漏洞管理和风险分析。
资产与供应链	**Model BOM**	描述 AI 系统所使用模型及其依赖关系的结构化清单，包括基础模型、微调模型、Adapter/LoRA、Embedding 模型以及相关版本、来源、许可证和完整性信息。
资产与供应链	**Data BOM**	描述 AI 系统训练、微调、评测和 RAG 等环节所使用数据及其依赖关系的结构化清单，包括数据来源、版本、用途、许可证、敏感性和处理方式等信息。
运营与响应	**AI 安全监控**	对 AI 模型、应用、Agent、Prompt、数据访问、工具调用及相关基础设施的安全状态和运行行为进行持续采集、分析、检测和告警的活动。
运营与响应	**AI SOC**	面向 AI 系统开展持续安全监测、检测、分析、响应和改进的安全运营能力或组织形态，其监测对象可覆盖模型、Prompt、Agent、RAG、工具调用、AI 资产和相关基础设施。
运营与响应	**AI 安全事件**	与 AI 系统相关并已经或可能对模型、数据、应用、业务、用户或组织产生安全影响的异常、攻击、滥用、失控、泄露或其他非预期事件。
运营与响应	**AI 事件响应**	针对 AI 安全事件开展准备、发现、分析、遏制、处置、恢复、取证和复盘改进的一系列组织与技术活动。
隐私与可信 AI	**差分隐私**	通过在统计查询、训练过程或输出中引入受控随机性，限制单个数据记录是否参与数据集对结果产生的可识别影响，从而降低个体信息泄露风险的隐私保护方法。
隐私与可信 AI	**联邦学习**	一种分布式机器学习方法，各参与方在本地数据上进行模型训练，并通过交换模型参数、梯度或其他受控更新信息协同训练模型，使原始训练数据原则上无需集中汇聚。
隐私与可信 AI	**可解释 AI**	使 AI 系统的输出、行为或决策依据能够被人理解、分析或说明的一组方法、技术和实践，用于支持透明性、风险分析、审计和责任追踪。

附录 D：AI 安全工具推荐

本附录收录截至 **2026 年 8 月中旬**仍具有代表性、公开源代码、可自行部署或独立运行的 AI 安全工具与框架。工具列表主要用于企业 AI 安全建设、技术验证和工具选型参考，不构成对任何项目、厂商或技术方案的背书，也不应直接作为采购、认证或合规依据。

企业需求	首选工具	备选工具	建议部署位置
LLM 漏洞扫描	garak	DeepTeam	测试环境

专业 AI 红队	PyRIT	garak	AI 安全实验室
AI 应用 CI/CD 安全测试	Promptfoo	Giskard OSS v3	DevSecOps 流水线
Agent 安全测试	PyRIT / Promptfoo	DeepTeam	上线前测试
输入输出 Guardrail	NeMo Guardrails	OpenAI Guardrails	AI Gateway / 应用层
Prompt Injection 检测	Prompt Guard	NeMo Guardrails	输入层
内容安全分类	Llama Guard	OpenAI Guardrails	输入/输出层
AI 生成代码安全	Code Shield	企业现有 SAST	vibe coding 代码生成与研发流程
Agent 运行时防护	NeMo Guardrails	LlamaFirewall	Agent Runtime
模型文件扫描	ModelScan	picklescan	Model Registry / CI
Pickle 深度分析	Fickling	picklescan	安全分析环境
模型签名与来源验证	Model Transparency	Sigstore	Model Registry
Agent / LLM Trace	Langfuse	Phoenix	Production
OTel 原生 AI 观测	OpenLIT	OpenLLMetry	Production
AI 渗透测试	CyberStrikeAI	Strix	AI 安全实验室
PII 识别与脱敏	Presidio	企业现有 DLP	AI Gateway
隐私计算	PySyft	—	数据平台
差分隐私	OpenDP	Google Differential Privacy	数据/模型训练平台

附录 E：参考资料

1. 国家互联网信息办公室：《生成式人工智能服务管理暂行办法》及生成式人工智能服务备案/登记公告。
2. 国家互联网信息办公室等：《人工智能生成合成内容标识办法》，2025 年 9 月 1 日起施行。
3. GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》。
4. GB/T 45654-2025《网络安全技术 生成式人工智能服务安全基本要求》。
5. GB/T 45652-2025《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》。
6. 国家互联网信息办公室等：《人工智能拟人化互动服务管理暂行办法》，2026 年 7 月 15 日起施行。
7. Regulation (EU) 2024/1689 (EU AI Act) 及 Regulation (EU) 2026/1744。
8. NIST AI Risk Management Framework 1.0、Generative AI Profile 及 2026 年更新工作。
9. OWASP Top 10 for LLM Applications 2025、OWASP Top 10 for Agentic Applications 2026 及 OWASP GenAI 2026 事件综述。
10. MITRE ATLAS（Adversarial Threat Landscape for AI Systems）。
11. ISO/IEC 42001、ISO/IEC 23894 等 AI 管理与风险管理标准。

© AICSO 2026. 本白皮书由 AICSO.AI 发布，遵循 CC BY-NC-SA 4.0 协议。