AI组织责任:
治理、风险管理、合规与文化方面

AI Organizational Responsibility Working Group

CSA GCR cloud security alliance® GREATER CHINA REGION

CSA cloud security alliance®

AI 组织责任工作组的永久官方地址是

https://cloudsecurityalliance.org/research/working-groups/ai-organizational-sponsibility

云安全联盟

创立于2009年，作为世界领先的独立、权威国际产业组织，致力于定义和提高业界对云计算和下一代数字技术安全最佳实践的认识和全面发展。

云安全联盟大中华区

在香港注册并在上海登记备案的国际NGO组织，旨在 立足中国，连接全球，推动 中国数字安全技术标准与产业的发展及国际合作。

4 大区
大中华区、美洲区、欧非区、亚太区

180+ 分会
英国、法国、加拿大、旧金山、马来西亚等覆盖50多个国家和地区

2.5K+ 成员单位
世界500强科技公司、安全厂商、中小型企业 、研究机构

20W+ 社区专业人员
研究工作组专家、社区志愿者、从业人员、CSA认证学员

前沿研究

#云安全 #AI安全
#零信任 #数据安全
#5G安全 #区块链安全
#量子安全 #物联网安全
#金融安全 #医疗安全
#智能座舱安全
#关键基础设施安全
.....

培训与认证

CCSK 云安全知识认证
CDSP 数据安全认证专家
CAISP人工智能认证专家
CZTP 零信任认证专家
CCPTP 云渗透测试认证专家
CCAK 云计算审计知识认证
.....

会议活动

CSA summit@RSAC
CSA GCR Congress
CSA研讨会
AI For GOOD峰会
.....

评估与认证

AI STR AI安全、可信、负责任认证
STAR 云安全评估认证
CAST 云应用安全可信认证
CNST 云原生安全可信认证
.....

1000+研究成果 **10W+认证学员** **1000W+传播量**

成员单位（部分）

ZEEKR Google Microsoft amazon HUAWEI intel IBM Security cisco ORACLE Meta 阿里云
UCLOUD优刻得 京东 上海CA中心 移动云 天翼云 中国电子云 CECLOUD China unicom中国联通 中国移动 China Mobile 中国电信 CHINA TELECOM ICBC 中国工商银行 LEXIN乐信

企业合作微信号:csagcr　认证培训微信号:CSAlynn　邮箱:info@c-csa.cn

目 录

引言

这份白皮书是一个系列中的第二部分，致力于阐明围绕人工智能（AI）组织责任的细节。首份白皮书探讨了核心安全原则，而本部分则关注治理、风险和合规（GRC）方面的内容。未来的白皮书将继续解决组织在采用和实施 AI 应用时的其他挑战，如供应链完整性和滥用行为的缓解。

该系列的首份白皮书《AI 组织责任 - 核心安全责任》深入探讨了与 AI 相关的企业核心安全责任，包括数据安全、模型安全和漏洞管理。

本白皮书综合了专家推荐的 GRC 最佳实践、文化方面以及影子 AI 预防措施，通过这六个关键领域的建议来指导企业负责任且安全的 AI 开发与部署。

所有责任的六个跨领域关注点

我们通过以下六个维度分析每项责任：

1．评估标准：通过量化的指标，帮助利益相关者衡量法规合规性、风险暴露情况，并与组织政策对齐，以确保 AI 技术中的 GRC 实践。

2．RACI 模型：执行（Responsible）、负责（Accountable）、咨询（Consulted）和知情（Informed）RACI 模型为任务、里程碑和 GRC 相关过程的可交付成果定义了角色和责任的结构化框架。此模型确保在整个 AI 生命周期中角色和责任的透明性和问责制。

译者注：由于 Responsible 和 Accountable 在翻译成中文以后都有责任相关的意思，为了更加清晰地表达原作者的意图，文报告将 Responsible 翻译为执行，后文中 RACI 即为:执行、负责、咨询和知情，并与《AI 组织责任 - 核心安全责任》中的翻译保持一致。

3．高级实施策略：说明 GRC 责任如何在组织层面实施，以及为成功采用需要克服的障碍。

4．持续监控与报告：持续的监控与报告机制对于保持 AI 系统中 GRC 的完整性至关重要。实时跟踪、合规性问题警报、审计轨迹等有助于识别安全事件，并为及时解决 GRC 相关问题提供支持。

5．访问控制：有效管理模型注册表、数据存储库和适当的访问权限有助于缓解与未经授权访问或滥用 AI 资源相关的风险。通过实施健壮的访问控制机制，组织可以保护敏感数据并确保遵守监管要求。

6．适用的框架与法规：遵守行业标准（如 ISO/IEC 27001、国家标准与技术研究所（NIST）指南及法规，如欧盟（EU）AI 法案）有助于确保 AI 项目与已建立的 GRC 实践对齐，维护组织价值观、责任和法规义务。

假设

本文件假设立场为行业中立，提供适用于不同部门的指南和建议，而不偏向特定行业。

目标受众

本白皮书旨在满足各种受众群体的需求，每个群体都有不同的目标和兴趣点：

1．首席信息安全官（CISOs）：该白皮书提供可操作的 AI 安全控制实施指导，帮助 CISOs 有效管理与 AI 相关的风险，确保符合行业标准，并将 AI 安全集成到其网络安全战略中。

2．AI 研究人员、工程师与开发人员：本白皮书为 AI 研究人员和工程师提供了全面的最佳实践指南，帮助他们开发出具有道德性和可信性的 AI 系统。

3．商业领袖与决策者：白皮书为 C-suite 高管提供了战略性指导，帮助他们做出 AI 采用相关的明智决策，优化 AI 驱动的业务价值，并确保与组织目标的对齐。

4．政策制定者与监管者：白皮书为政策制定者和监管者提供了关键的洞见，

帮助他们塑造 AI 治理方面的政策和监管框架。

5．投资者与股东：通过本白皮书，投资者和股东能够更好地理解组织对负责任 AI 实践的承诺，为投资决策提供重要的参考。

6．客户与公众：白皮书帮助组织向公众表明其对负责任 AI 开发的承诺，让个人了解其数据如何被保护，AI 系统如何为社会效力。

责任角色定义

以下表格提供了一个通用指南，展示了在整合或操作 AI 技术的组织中常见的各种角色。每个组织可能会根据其运营需求、文化和 AI 项目的特定需求，定义不同的角色及其相应的责任。因此，表格旨在作为参考，组织可根据需要调整这些角色，以确保其结构与战略目标和运营框架保持一致。

管理和策略

角色名称	角色描述
首席数据官（CDO）	负责监督企业数据管理、政策制定、数据质量及数据生命周期。
首席技术官（CTO）	领导技术战略并监督技术发展。
首席信息安全官（CISO）	监督完整的网络安全战略和运营。
业务部门领导	领导业务部门并确保 AI 项目与业务目标对齐。
首席 AI 官	负责 AI 技术在组织中的战略实施和管理。
首席产品官（CPO）	领导产品战略，确保 AI 项目和技术发展与业务目标一致。
管理层	监督并指导整体战略，确保与组织目标一致，包括 CEO、CTO、CISO 等。
首席云计算官	领导云战略，确保云资源与业务和技术目标保持一致。

治理、风险与合规

角色名称	角色描述	类别名称
数据保护官	负责管理数据保护战略和 GDPR 合规。	治理与合规
首席隐私官	确保符合隐私法律和法规要求。	治理与合规
法律与合规部门	就 AI 部署和使用相关的法律/监管义务提供建议。	治理与合规
法律团队	提供有关 AI 部署和使用的法律指导，并与供应商商定 AI 特定条款。	治理与合规
数据治理委员会	为数据治理和使用制定政策和标准。	治理与合规
合规团队	验证法律和法规的合规性，以及组织政策的合规性。	治理与合规
数据治理官	负责组织内部的数据治理，确保符合政策、数据隐私法和法规合规要求。	治理与合规
GRC 审计员	确保组织符合监管要求，管理风险并维持强健的治理实践。	治理与合规

技术与安全

角色名称	角色描述
IT 安全团队	实施并监控安全协议以保护数据和系统。
网络安全团队	保护网络免受威胁和漏洞攻击。
云安全团队	确保基于云的资源和服务的安全性。
网络安全团队	保护网络免受网络威胁、漏洞和未经授权的访问。
IT 团队	支持并维护 IT 基础设施，确保其操作性和安全性。

角色名称	角色描述
网络安全官	监督网络的安全性，确保数据保护和威胁缓解。
硬件安全团队	保护物理硬件免受篡改和未经授权的访问。
系统管理员	管理和配置 IT 系统和服务器，确保性能和安全性。
应用安全团队	识别、缓解并防止整个应用生命周期中的安全漏洞。

运营与开发

角色名称	角色描述
AI 开发团队	开发并实施 AI 模型和解决方案。
开发与运营（DevOps）团队	自动化和简化软件交付与基础设施管理流程，促进开发与运营团队之间的协作。
质量保证团队	测试并确保 AI 应用程序和系统的质量。
AI 运营团队	管理 AI 系统的性能和可靠性。
应用开发团队	开发应用程序，集成 AI 功能。
AI/ML 测试团队	专门测试人工智能/机器学习（AI/ML）模型的准确性、性能和可靠性。
应用安全与测试团队	确保应用程序的安全性，并能够抵御各种威胁。
AI 维护团队	维护 AI 系统和模型，确保在更新和优化后正常工作。
项目管理团队	从项目启动到完成全程监督 AI 项目，确保其目标和时间表达成。
开发团队	从事 AI 模型和系统的创建和改进工作。
数据科学团队	收集并准备用于 AI 模型训练和分析的数据。
容器管理团队	管理容器化应用程序，促进部署和扩展。
AI 开发经理	领导 AI 开发项目，指导团队实现成功实施。
AI 运营主管	监督 AI 相关的运营，确保 AI 解决方案的效率和有效性。

规范性引用

下列文件对于应用和理解本文档至关重要。

- 生成式 AI 安全：理论与实践
- OpenAI 应急准备框架
- 将 CCM 的 AIS 域应用于生成式 AI
- Google 安全 AI 框架 (SAIF)
- 欧盟 AI 法案
- 拜登关于安全、可信的人工智能的行政命令
- OWASP LLM 应用程序 Top 10
- CSA 云控制矩阵 (CCM v4)
- MITRE ATLAS™ (人工智能系统的对抗性威胁环境)
- NIST 安全软件开发框架 (SSDF)
- NIST 人工智能可信度和风险管理框架
- 通用数据保护条例 (GDPR)
- OWASP LLM AI 网络安全与治理检查清单
- OWASP 机器学习 Top 10
- OWASP 攻击面分析备忘单
- 世界经济论坛简报-人工智能治理联盟系列

术语表

云安全术语表链接

https://cloudsecurityalliance.org/cloud-security-glossary

（正文内容如下）

1. 风险管理

有效的风险管理是稳健 AI 治理的基础，涵盖了一系列方法来识别、评估和缓解 AI 系统及其输出的潜在威胁。在当前快速变化的 AI 环境中，灵活的风险管理对于确保 AI 技术的可靠性、安全性和谨慎操作至关重要。本节探讨了风险管理的多个方面，包括威胁建模、全面的风险评估、攻击模拟、事件响应规划、灾难恢复策略、审计日志、活动监控和数据漂移监控。AI 风险管理是一个持续的过程，应该嵌入到 AI 解决方案的开发生命周期和运营中。这包括 AI 解决方案的初步设计、开发、测试、实施以及持续的监控。为 AI 模型应用确定的每个业务用例都应经过以下关键的 AI 风险管理组件，无论是构建内部 AI 模型还是引入第三方 AI 技术/解决方案。

通过整合这些实践，组织可以主动解决漏洞，增强其抵御 AI 相关威胁的能力，并维护 AI 系统的完整性和可信度。以下小节提供了深入分析这些 AI 风险管理的关键组件。

1.1 威胁建模

AI 威胁建模指的是组织系统性地评估和理解其 AI 系统潜在漏洞和风险的义务。此责任包括识别和分析 AI 系统的各种入口点、接口和组件，这些组件可能被恶意

行为者利用，或导致意外后果。

具体来说，AI 威胁建模包括：

- **数据流图（DFDs）分析：**DFDs 为理解系统潜在的攻击面提供了重要洞见。通过研究 DFDs，AI 安全评估人员可以识别易受攻击的入口和出口点。这些图表直观地展示了数据流，暴露了接口、API、数据库和其他可能被利用的组件。此外，DFDs 有助于说明信任边界，清晰划定可信与不可信域之间的过渡点，这对于实施有效的安全控制至关重要。

- **数据输入和输出分析：**这涉及检查 AI 系统生成的数据输入来源和输出结果，以了解与数据质量、完整性和隐私相关的潜在安全风险。

- **系统依赖性理解：**识别 AI 系统与组织基础设施中其他组件之间的依赖关系和交互，包括 API、数据库和外部服务。

- **潜在攻击向量识别：**分析攻击者可能如何针对 AI 系统，例如通过数据中毒、模型操纵或推断攻击。

- **安全控制评估：**这涉及评估 AI 系统中现有的安全控制和机制的有效性，以减轻潜在的威胁和漏洞（参见 CSA 大型语言模型威胁分类）。

与威胁建模相关的跨领域责任包括：

1．评估标准：组织应建立量化指标来评估其 AI 威胁建模的有效性。指标可能包括已识别的威胁数量、漏洞严重程度以及成功缓解威胁的比率。

2．RACI 模型：RACI 模型帮助澄清 AI/ML 威胁建模的组织角色和责任。关键人员应被指定为执行、负责、咨询或知情，确保整个威胁建模过程中的明显监督和问责。

3．高级实施策略：为 AI/ML 威胁建模实施 GRC 责任涉及制定和执行高级策略，概述组织的威胁建模方法。应解决的障碍包括资源约束和变革阻力。

4．持续监控和报告：持续监控工具和报告机制对于保持 AI/ML 威胁建模的完整性至关重要。实时警报、审计记录和定期报告使组织能够及时识别和处理安全事件或合规违规行为。

5．访问控制：访问控制机制保护 AI/ML 威胁建模过程。组织必须实施健全的控制措施，以管理对敏感数据、模型注册表和其他关键资产的访问。

6．适用框架与法规：如 NIST AI RMF、NIST SSDF、NIST 800-53 等框架是一些顶级的威胁建模框架，例如 STRIDE（微软）、MITRE ATT&CK（MITRE）和 OCTAVE（卡内基梅隆大学）。

1.2 风险评估

风险评估在 AI 项目中至关重要，因为它们识别并分析贯穿整个 AI 生命周期的潜在风险。风险评估的步骤如下：

1．识别风险：在进行 AI 项目的风险评估时，至关重要的是有条不紊地识别所有来自 AI 技术及其使用的潜在风险。这些风险可能源于各种不同的来源，如数据质量问题（参见 AI 组织责任 - 核心安全责任）、算法偏见、网络安全威胁、合规性问题以及伦理问题。

AI 项目适用的一些风险分类包括：

- **数据风险：**与 AI 系统中使用的数据质量、完整性、隐私和安全相关的风险。
- **模型风险：**与 AI 模型的开发、验证和部署相关的风险，包括偏见、公平性、准确性和可解释性。
- **运营风险：**由 AI 系统的日常运行引发的风险，如性能下降、系统故障和监控不足。

- **伦理风险：**与 AI 的伦理影响相关的风险，包括意外后果、社会影响以及对个人或群体的潜在伤害。
- **法规风险：**因不遵守 AI 使用、数据保护和隐私等相关法律、法规和行业标准引发的风险。
- **法律风险：**与 AI 相关活动引发的潜在法律责任、诉讼和纠纷，包括知识产权侵权和合同义务。
- **声誉风险：**由于 AI 相关事件或争议导致负面宣传、公众反对或失去信任，给组织的声誉和品牌形象带来的风险。
- **战略风险：**与将 AI 项目与组织目标、长期战略和利益相关者期望保持一致相关的风险。
- **财务风险：**与 AI 项目的财务影响相关的风险，包括预算超支、成本不确定性以及未能实现预期回报。
- **供应链风险：**由于依赖第三方供应商、供应商或服务提供商，参与 AI 系统开发、部署或维护的风险。

另一份 CSA AI 文档中记录了一些可能的 AI 威胁类别。

2．分析风险：识别风险后，必须对其进行分析，以评估其潜在影响及其发生的可能性。通过严谨的分析，组织可以根据风险的关键性和组织的风险容忍度水平优先处理这些风险。通过这样的分析，组织能够有效优先分配资源应对最重要的风险。

具体步骤包括：

- **严重性评估：**根据风险的严重性对风险进行评估，涵盖其对组织、利益相关者及更广泛生态系统可能带来的潜在后果。评估内容包括财务损失、声誉损害、监管处罚和运营中断等因素。

- **后果评估：**进一步基于风险的潜在后果对其进行评估，包括对组织及其利益相关者的直接和间接影响。评估内容还包括风险可能对业务运营、客户信任、市场竞争力和法律合规性产生的影响程度。

- **发生概率评估：**基于历史数据、行业趋势、内部控制和外部威胁等因素对风险发生的可能性进行评估。发生概率的评估帮助组织衡量风险成真的概率，并为风险管理优先级和资源分配决策提供依据。

- **优先级标准：**根据风险的关键性及组织的风险容忍度确定其优先级。这包括建立风险优先处理的标准，如潜在影响的大小、发生的可能性、响应的紧迫性以及组织的战略目标。对组织目标和运营构成最大威胁的风险优先进行缓解。

- **严格分析：**风险分析过程涉及对每个识别出的风险进行严格审查，采用定量和定性方法。这可能包括统计建模、场景分析、敏感性测试、专家判断以及利益相关者的磋商，以收集多方视角和洞见。

通过对已识别的风险进行深入分析，组织可以更好地理解这些风险的潜在影响及其发生的可能性。这使得组织能够有效地优先处理风险管理工作，并根据风险的关键性优先级分配资源。该信息驱动的方法帮助组织提升其应对能力，并提高主动管理风险的准备度。

3．技术控制： 实施技术控制涉及利用安全机制、协议和工具来保护 AI 系统免受潜在威胁和漏洞的侵害。这可能包括加密技术，以保护数据的完整性和机密性；基于角色的访问控制，以确保适当的数据访问、最小权限访问；以及检测系统，以缓解和应对恶意活动。

- **数据治理实践：**增强数据治理实践需要建立稳健的政策、程序和标准，以管理和保护数据的生命周期。这包括数据质量保证措施，以确保训练数据的

准确性和可靠性，数据沿袭跟踪以保持透明度和问责制，数据访问控制以执行隐私和安全要求。

- **安全评估与缓解：**开发安全评估对 AI 系统的安全性和功能至关重要。应在软件开发生命周期的各个阶段减少幻觉、过度依赖、偏差和有害输出。
- **网络安全措施：**建立健全的网络安全措施需要实施全面的安全协议和实践，以防止网络威胁和攻击。这包括网络安全措施，以保护 AI 系统免受未经授权的访问和数据泄露，端点安全措施以保护连接到 AI 系统的设备和端点，以及威胁情报计划以主动识别和缓解新兴威胁。
- **风险管理目标对齐：**确保缓解措施与组织的整体风险管理目标对齐，涉及将风险缓解策略整合到更广泛的风险管理框架和流程中。这包括将缓解工作与组织的优先级、资源分配和风险容忍度对齐，以有效应对已识别的风险和漏洞。

评估标准：

- 全面识别 AI 生命周期各阶段的风险
- 深入分析风险的准确性（影响和可能性评估）
- 风险缓解策略的有效性
- 风险监控和审查流程的及时性和规律性
- 在风险评估中使用的数据的质量和相关性
- 风险评估结果与组织风险容忍度的对齐
- 风险评估结果在决策过程中的整合
- 适应新兴 AI 相关风险的风险评估方法的灵活性

责任矩阵（RACI 模型）：

- **执行：** IT 安全团队，AI 开发团队，数据科学团队
- **负责：** 首席信息安全官（CISO），首席 AI 官员
- **咨询：** 法律和合规部门，业务部门领导，首席隐私官，云服务提供商，第三方 AI/ML 模型提供商
- **知情：** 管理层，首席技术官，首席数据官

高层次实施策略：

1. 建立全面的 AI 风险评估框架。
2. 通过利用多种来源和视角，开发风险识别流程。
3. 实施针对 AI 特定风险的稳健风险分析方法。
4. 创建与组织目标对齐的风险缓解策略库。
5. 建立持续的风险监控机制和定期审查周期。

持续监控与报告：

- 实施 AI 系统关键风险指标（KRIs）的实时监控。
- 建立针对风险指标临界值突破的自动化警报系统。
- 定期（如每季度）和重大变化时进行风险评估审查。
- 开发适用于不同利益相关者群体的标准化风险报告模板。
- 实施风险仪表板，以可视化和跟踪随时间变化的 AI 相关风险。
- 建立反馈循环，以不断改善风险评估流程。

访问控制映射：

- **IT 安全团队：** 完全访问风险评估工具和数据。
- **AI 开发团队：** 访问与其项目相关的风险评估结果。

- **数据科学团队：** 访问与数据相关的风险评估和缓解策略。
- **CISO 与首席 AI 官员：** 无限制访问所有风险评估信息。
- **法律与合规部门：** 访问与合规相关的风险评估。
- **业务部门领导：** 访问高级风险评估摘要。
- **管理层：** 访问执行摘要和战略风险洞察。

适用框架和法规：

- 遵守行业风险管理标准（例如，ISO 31000，NIST RMF）。

1.3 攻击模拟

模拟攻击可以对 AI 系统进行压力测试，使其在部署后更加稳健。这些模拟应尽可能在与真实世界环境接近的条件下进行。以下是基于上述威胁的一些攻击模拟示例。

1. 场景：数据投毒攻击

- **威胁：** 恶意行为者将虚假或操纵的数据注入训练数据集，以开发 AI 模型。
- **影响：** AI 模型从受污染的数据中学习，导致部署期间的预测或决策不准确。
- **可能性：** 中等到高，特别是在训练数据源没有得到充分保护或审查的情况下。
- **模拟：** 模拟一个攻击场景，攻击者未经授权访问训练数据存储库，并注入设计用于歪曲 AI 模型学习过程（完整性）的虚假数据实例，或阻止访问部分新数据或旧数据（可用性）。示例包括标签投毒（改变数据标签）、有针对性的投毒（引入少量新数据干扰训练过程）、以及后门投毒（以某种方式改变原始数据，例如翻转像素）以干扰训练。

- **缓解措施：** 实施数据验证和异常检测机制，以识别和缓解训练期间的受污染数据实例。此外，应使用访问控制和加密来保护训练数据集的完整性。通过采取主动措施来训练对抗性样本，模型可以标记并阻止某些数据投毒，减少数据投毒的影响。

2. 场景：对抗性样本攻击

- **威胁：** 攻击者构造输入（例如，图像、文本）以欺骗 AI 模型并产生错误输出。
- **影响：** AI 模型错误分类或误解对抗性输入，导致实际应用中的错误结果。
- **可能性：** 中等，可以通过使用专门技术生成对抗性样本，这些技术利用 AI 模型架构中的漏洞。
- **模拟：** 生成针对已部署 AI 模型（例如，图像识别系统）的对抗性样本，并评估其抵御此类攻击的稳健性，通过衡量对抗性输入的预测准确性来进行评估。
- **缓解措施：** 在模型开发阶段使用对抗性训练技术，增强模型对对抗性样本的抵抗力。通过使用多样化的数据集定期更新和重新训练 AI 模型，以提高模型的泛化性和稳健性。

3. 场景：模型反转攻击

- **威胁：** 攻击者利用 AI 模型的输出推断训练数据或个人数据主体的敏感信息。
- **影响：** 未经授权泄露机密信息，例如从 AI 模型输出推断的个人属性或专有知识。
- **可能性：** 低到中等，取决于数据的敏感性和模型输出的透明度。
- **模拟：** 通过利用已部署 AI 模型（例如，面部识别系统）的输出进行模型反转攻击，以重建敏感训练数据或推断个人的私人属性。

- **缓解措施：** 实施隐私保护技术，如数据最小化、数据加密、差分隐私、联邦学习或输入/输出扰动，以减轻通过模型输出泄露信息的风险。此外，应限制对敏感模型输出的访问，实施访问控制以防止未经授权的披露。定期更新和重新训练模型以适应最新威胁也可以有助于缓解此类风险。

4. 场景：模型规避攻击

- **威胁：** 攻击者操纵输入数据，以规避基于 AI 的安全系统（如入侵检测系统和恶意软件检测器）的检测或分类。
- **影响：** 成功规避基于 AI 的安全防御，导致恶意活动或攻击者利用的漏洞未被检测到。
- **可能性：** 中等到高，攻击者不断进化规避技术以绕过 AI 的安全措施。
- **模拟：** 设计并执行规避攻击，使用设计用于规避检测或触发误报的对抗性输入对基于 AI 的安全系统进行攻击。
- **缓解措施：** 通过整合多种检测机制并使用集成学习技术来检测和缓解规避尝试，增强基于 AI 的安全系统的弹性。通过使用真实世界的攻击数据定期更新和重新训练安全模型，以适应不断演变的威胁和规避策略。此外，应实施异常检测和评分系统以识别规避攻击的可疑模式，输入监控和清理也可以减少规避攻击。

以下将讨论此职责项下的六个跨领域关键概念。

评估标准：

- 覆盖攻击场景的全面性
- 模拟攻击的现实性和准确性
- 攻击检测机制的有效性
- 缓解响应的速度和效率

- 不同 AI 模型类型和应用的覆盖范围
- 与当前威胁形势和新兴攻击矢量的对齐
- 将模拟结果整合到安全改进过程中的程度
- 攻击模拟的频率和规律性

责任矩阵（RACI 模型）：

- **执行：** IT 安全团队，网络安全团队
- **负责：** 首席信息安全官（CISO）
- **咨询：** AI 开发团队，数据科学团队，AI 运营团队
- **知情：** 首席技术官，首席 AI 官，业务部门领导

高层次实施策略：

1. 开发 *AI* 特定攻击场景的全面目录。
2. 为每个场景设计并实施现实的攻击模拟。
3. 建立专用环境以进行攻击模拟。
4. 创建定期攻击模拟时间表，涵盖不同的 *AI* 系统。
5. 制定指标和评估标准，用于评估模拟的有效性。
6. 实施反馈回路，将模拟结果整合到安全改进中。
7. 对相关利益相关者进行模拟后的分析和报告。
8. 根据新兴威胁定期更新攻击模拟技术。

持续监控与报告：

1. 在攻击模拟期间实施实时监控。
2. 开发用于模拟结果的自动化报告机制。

3. 定期审查模拟结果和趋势。
4. 实施漏洞跟踪和优先级确定系统。
5. 建立模拟技术的持续改进流程。

访问控制映射：

- **IT 安全团队和网络安全团队：** 完全访问模拟工具和结果。
- **首席信息安全官（CISO）：** 无限制访问所有模拟数据和报告。
- **AI 开发团队：** 访问与其项目相关的模拟结果。
- **数据科学团队：** 访问与数据相关的模拟结果。
- **AI 运营团队：** 访问模拟的运营影响评估。
- **首席技术官与首席 AI 官：** 访问高级模拟报告。
- **业务部门领导：** 访问模拟结果的业务影响摘要。

适用框架和法规：

- 遵循 NIST AI 风险管理框架
- 欧盟 AI 法案
- NIST AI 100-2 E2023
- OWASP LLM Top-10

1.4 事件响应计划

为 AI 制定事件响应计划涉及多个关键步骤，以确保组织能够有效地检测、响应并从与 AI 相关的事件中恢复。以下是该过程的概述。

1. 准备：

a. 建立由具备 AI、网络安全、法律和沟通专业知识的人员组成的事件响应团队。

b. 明确事件响应团队中的角色和职责，包括事件协调员、技术分析师、法律顾问和沟通联络员。

c. 针对 AI 系统进行风险评估，以识别潜在的威胁、漏洞和影响场景。

d. 制定针对 AI 相关事件的事件响应政策、程序和操作手册，包括检测、遏制、消除、恢复和事件后分析。

2. 检测：

a. 实施针对 AI 的监控和日志功能，以检测异常行为、偏离预期模式的情况或可能的妥协迹象。

b. 部署 AI 驱动的安全解决方案，用于威胁检测，例如异常检测算法、行为分析和模式识别技术。

c. 为 AI 模型和系统建立基准性能指标，以便检测可能表明安全事件的偏差或异常。

3. 遏制与消除：

a. 在检测到与 AI 相关的事件时，立即采取遏制措施，防止进一步扩散或影响。

b. 隔离受影响的系统、网络或数据存储库，以最小化事件的范围，防止未经授权的访问或利用。

c. 部署补救措施，消除恶意组件，恢复受影响的系统至已知的良好状态，并消除持久威胁或后门。

4. 恢复：

a. 从备份存储库或干净快照中恢复受影响的 AI 系统、模型或数据集，以确保操作的连续性。

b. 通过全面的测试和验证程序，验证恢复系统的完整性和功能。

c. 实施额外的安全控制、补丁或更新，以增强 AI 系统抵御未来事件的能力。

5. 事件后分析：

a. 进行深入的事件后分析，以识别事件的根本原因、攻击向量和从中学到的经验教训。

b. 记录发现、观察和改进建议，用于提升事件响应程序、安全控制和风险管理实践。

c. 根据从事件后分析中获得的见解，更新事件响应手册、政策和培训材料，以提高未来事件的应对能力。

6. 培训与意识：

a. 为事件响应团队成员及相关利益相关者提供定期的培训和意识提升计划，确保他们熟悉与 AI 相关的威胁、攻击向量和响应程序。

b. 进行桌面演练、模拟或红队演练，以测试事件响应计划的有效性，并识别需要改进的领域。

以下内容涵盖了该责任项下六个跨领域关键关注点：

评估标准：

- 事件响应计划涵盖所有 AI 系统的全面性
- 事件检测和响应的速度与效率
- 遏制与消除措施的有效性
- 恢复程序的稳健性
- 事件后分析的质量和深度
- 培训与意识提升计划的频率与有效性

- 与行业最佳实践和监管要求的一致性
- 计划适应新兴 AI 特定威胁的能力

责任矩阵（RACI 模型）：

- **执行：** IT 安全团队、网络安全团队、AI 运营团队
- **负责：** 首席信息安全官（CISO）
- **咨询：** AI 开发团队、数据科学团队、法律与合规部门、沟通团队、产品管理
- **知情：** 首席技术官、首席 AI 官、业务部门领导、管理层

高层次实施策略：

1. 建立具备 AI 专业知识的跨职能事件响应团队。
2. 制定 AI 特定的事件响应政策、程序和操作手册。
3. 实施针对 AI 的监控与检测能力。
4. 为 AI 相关事件制定遏制与消除程序。
5. 为 AI 系统、模型和数据集建立恢复流程。
6. 开发事件后分析与报告框架。
7. 实施定期的培训和意识提升计划，如桌面演练和红队演练。
8. 定期测试和优化事件响应计划。
9. 不断改进并从事件响应经验中学习。

持续监控与报告：

1. 实施 AI 系统的实时监控，识别异常和潜在事件。
2. 建立关键绩效指标（KPI），评估事件响应的有效性。
3. 开发自动警报系统，用于检测到的事件。

4. 定期审查事件响应的性能和结果。
5. 实施漏洞跟踪和优先级确定系统。
6. 建立持续改进事件响应能力的流程。

访问控制映射：

- **事件响应团队：** 完全访问事件响应工具和受影响的系统。
- **首席信息安全官（CISO）：** 无限制访问所有与事件相关的信息和报告。
- **AI 运营团队：** 访问事件期间的运营数据和系统日志。
- **AI 开发团队：** 访问与其项目相关的事件数据。
- **数据科学团队：** 访问与数据相关的事件信息。
- **法律与合规部门：** 访问事件报告，以进行合规评估。
- **沟通团队：** 访问批准的外部沟通信息。
- **管理层：** 访问高级事件摘要和影响评估。

适用框架与法规：

- 遵守关于事件报告和数据保护的监管要求，例如 HIPAA、PCI-DSS、GDPR。

1.5 运营弹性

对于 AI 应用程序而言，业务连续性计划（BCP）和灾难恢复（DR）至关重要，考虑到重大事件可能中断运营，而 AI 在各个行业中的关键角色。BCP 和 DR 涉及积极的规划和强大的响应策略，以最大限度减少破坏性事件的影响，并快速恢复功能。与 AI 应用程序灾难恢复相关的几个关键风险值得关注。

数据丢失：AI 模型高度依赖数据，灾难造成的关键数据丢失可能会损害模型性能和完整性。

- 可能性： 高。考虑到自然灾害、硬件故障和网络攻击等普遍威胁，数据丢失是任何 AI 驱动操作的重大风险。
- 影响：严重。关键数据的丢失会削弱 AI 模型，影响性能和决策能力，并可能导致监管处罚。
- 模拟：进行涉及数据丢失场景的模拟演练，以评估恢复过程和时间。使用合成数据模拟关键数据集的丢失，并测试恢复能力。
- BCP/DR 建议： 实施强大的数据备份和服务恢复策略，并定期执行。使用异地和云存储解决方案以确保冗余。使用多可用区和跨区域复制的云服务。数据加密和安全备份存储至关重要。明确数据恢复和服务恢复流程中的职责分工。

模型损坏：灾难可能损坏或销毁 AI 模型，导致耗时且代价高昂的模型恢复或重新训练。

- 可能性：中等到高。人类错误、网络攻击和软件故障等因素大大增加了模型损坏的风险。
- 影响：显著。AI 模型损坏可能导致输出不准确、决策失误，并造成用户和利益相关者之间的信任丧失。
- 模拟：定期通过引入错误或故障来测试模型完整性，评估版本控制系统和回滚程序的有效性。
- BCP/DR 建议：为所有 AI 模型及其组件实施版本控制。定期备份并安全存储模型版本，以实现快速恢复。应建立自动化监控系统，检测并警告模型损坏的异常情况。

第三方依赖：依赖外部服务和 API 获取数据或计算资源使 AI 应用程序暴露于依赖关系故障可能带来的连锁反应中。

- 可能性：中等。对外部服务和 API 的依赖引入了重大风险，考虑到各供应商在安全性和操作标准上的差异。
- 影响：高。第三方服务中断或违规可能会干扰 AI 运营，导致服务停机和数据安全问题。
- 模拟：定期执行模拟第三方服务失败的演练，评估故障切换和替代流程的稳健性。
- BCP/DR 建议：开发多元化的供应商组合，并考虑多云策略以降低风险。与所有第三方供应商签订服务级别协议（SLA），包括正常运行时间保证和恢复支持。

安全漏洞：在恢复期间，跨环境的数据和模型复制引入了未授权访问或操纵的潜在漏洞。

- 可能性：高。随着网络威胁形势不断演变，安全漏洞成为重大关注点。
- 影响：严重。被利用的漏洞可能导致 AI 系统受损、数据泄露和严重的声誉损害。
- 模拟：定期进行渗透测试和红队演练，以识别和解决漏洞。模拟漏洞情景，以测试事件响应和恢复能力。
- BCP/DR 建议： 实施分层的安全架构，包括防火墙、入侵检测/防御系统和严格的访问控制。定期进行安全培训，并确保所有人员都熟悉事件响应计划。

可扩展性挑战：灾难引发的 AI 服务需求波动可能会压垮恢复计划，缺乏可扩展解决方案可能导致性能下降。

- 可能性：中等。AI 服务需求的快速变化可能导致可扩展性挑战，尤其是在未提前规划的情况下。

- 影响：中等到高。无法扩展可能导致性能下降、用户不满，并在高峰期导致潜在的收入损失。
- 模拟：进行压力和负载测试，以评估系统在极端条件下的性能，并识别瓶颈。
- BCP/DR 建议：实施可扩展的云服务，并考虑无服务器架构以应对波动的需求。自动扩展和资源优化策略应是系统设计中的重要组成部分。

法规合规： 灾难恢复策略必须符合数据保护、隐私和安全法规，以降低法律和合规风险。

- 可能性：高。AI 和数据隐私的监管环境是动态的，频繁引入新的和更新的法规。
- 影响：高。不合规可能导致巨额罚款、法律挑战和声誉损害。
- 模拟：定期进行合规审计和模拟监管检查，有助于为现实世界的合规评估做好准备。
- BCP/DR 建议：建立合规管理体系，包括定期培训、审计和根据法律法规变化对政策和程序进行更新。利用法律专业知识应对复杂的法规环境。

技术债务：如果灾难恢复计划未能随着 AI 系统架构和技术的演变进行更新，可能会导致其无效。

- 可能性：高。快速的技术进步和交付压力可能导致技术债务的累积。
- 影响：中到高。累积的技术债务可能会阻碍灾难恢复工作，导致停机时间延长和恢复成本增加。
- 模拟：定期对 AI 系统架构和代码库进行审查和审计，有助于识别可能影响灾难恢复的技术债务领域。

- BCP/DR 建议：通过定期重构和现代化措施，优先减少技术债务。建立明确的文档并更新灾难恢复计划，以反映当前的系统架构和技术。

人为错误：AI 系统的复杂性增加了在恢复过程中发生人为错误的风险，可能加剧灾难的影响。

- 可能性：高。AI 系统的复杂性和涉及的各种人员操作，使人为错误成为一个重大风险。
- 影响：中到高。人为错误可能导致数据丢失、系统故障以及 AI 模型结果的不正确。
- 模拟：进行桌面演练和灾难场景模拟，以训练员工正确的响应程序，并识别潜在的错误领域。
- BCP/DR 建议： 制定全面的培训计划和清晰的程序文件，以最小化人为错误。实施检查与制衡机制，如同伴审查和自动化的异常活动警报。

测试不足：不频繁或不现实的灾难恢复计划测试可能导致过时或无效的策略，在实际事件中无效。

- 可能性：中到高。AI 系统的动态特性和持续推出新功能的压力，可能导致对灾难恢复计划的测试不足。
- 影响：高。计划的失败可能导致系统长时间不可用，并在业务最需要数据时可能丢失数据，突显了严格测试的重要性。
- 模拟：定期对灾难恢复计划的各个方面进行全面测试，包括在真实世界条件下进行突击演练，以评估准备情况。
- BCP/DR 建议： 分配专用资源用于定期测试和更新灾难恢复计划。将从测试和实际事件中汲取的经验教训纳入持续改进过程中。

资源限制：为备份、复制和快速部署分配充足的资源对 AI 应用程序的有效灾难恢复至关重要。

- 可能性：中等。预算和资源限制是常见的，特别是在竞争激烈和快速发展的行业中。
- 影响：中到高。资源的可用性可能显著影响灾难恢复解决方案的有效性，并最终决定恢复的速度和组织的弹性。
- 模拟：进行容量规划演练和成本效益分析，以优化灾难恢复的资源分配。
- BCP/DR 建议：为了保护关键系统和数据，应在预算和资源分配中优先考虑灾难恢复。探索具有成本效益的解决方案，如云服务，以实现可扩展的按需资源。

认识到与 AI 系统相关的多方面风险，包括数据丢失、模型损坏、第三方依赖性、安全漏洞、可扩展性挑战、法规合规、技术债务、人为错误、测试不足以及资源限制，显然，积极且强大的灾难恢复策略不仅是必要的，而且是负责任地利用 AI 的基石。这样的策略不仅旨在最大限度地减少破坏性事件的影响，还确保快速恢复 AI 功能，从而保持运营的弹性，并符合不断变化的监管环境。

评估标准

- 客观测量：开发针对 AI 应用程序的恢复时间目标（RTO）和恢复点目标（RPO）的明确指标。
- 风险评估：定期评估数据丢失、模型损坏、第三方依赖性、安全漏洞、可扩展性挑战、法规合规、技术债务、人为错误、测试不足和资源限制的可能性和影响。

责任矩阵（RACI 模型）：

- 执行：AI 开发团队、IT 安全团队、数据保护官员，负责实施灾难恢复策略并确保数据保护和模型完整性。

- 负责：首席数据官（CDO）和首席技术官（CTO），负责确保整体治理和遵守合规标准。首席信息安全官（CISO）负责安全措施和漏洞评估。
- 咨询：业务部门领导、首席 AI 官、合规团队，提供业务影响分析和法规合规建议。法律和合规部门指导法规和合规要求。
- 知情：所有组织成员都被告知有关灾难恢复政策、程序和职责的信息，并在 RACI 框架中分配角色。

高层次实施策略：

- 数据管理：实施强大的数据备份、加密和安全存储解决方案。使用云服务实现冗余和可扩展性。
- 模型完整性：对 AI 模型及其组件使用版本控制和安全存储。自动化监控以便早期检测到损坏或故障。
- 安全架构：开发分层安全策略，包括防火墙、入侵检测系统和严格的访问控制。

持续监控与报告：应利用自动化系统持续监控 AI 系统的健康状况和安全性。为管理和战略角色生成定期报告，突出需要注意的任何问题或风险。

访问控制：实施严格的政策，确保只有授权人员可以访问关键数据和系统。使用自适应身份验证和基于角色的访问控制（RBAC）机制。

适用框架与法规

遵循 NIST AI 风险管理框架（RMF）和安全软件开发框架（SSDF），确保 AI 应用程序的安全开发和部署。

定期进行合规检查，并根据不断变化的法规和标准对政策和程序进行更新。

1.6 审计日志与活动监控

审计日志和 AI 系统的活动监控对于治理、风险和合规（GRC）实践至关重要。这些日志提供了在 AI 系统中执行的各种活动的详细记录，包括模型训练、推理、数据处理和系统配置更改。以下是 AI 如何实施审计日志和活动监控的方式。

1. 捕捉相关事件：

a. 审计日志应捕捉与 AI 系统相关的各种事件，包括模型训练迭代、数据预处理步骤、推理请求和模型性能指标。

b. 记录详细信息，例如负责执行操作的用户或服务账户、事件的时间戳、执行的具体操作以及与事件相关的任何元数据。

2. 细粒度日志记录：

a. 实施细粒度日志记录，捕捉每个事件的详细信息，例如用于模型训练的输入数据、配置的超参数、生成的输出预测以及处理过程中遇到的任何错误或异常。

b. 确保审计日志包含足够的上下文，以便跟踪和追溯每个在 AI 系统中执行的操作。

3. 集中日志存储：

a. 将审计日志存储在支持可扩展存储、高效检索和安全访问控制的集中式存储库或日志管理平台中。

b. 实施加密和访问控制，以保护审计日志中的敏感信息，并确保符合数据保护法规。

4. 实时监控：

a. 实时监控 AI 系统，以检测并响应可能表明安全漏洞、数据泄露或性能下降的异常或可疑活动。

b. 设置警报机制，以通知管理员或安全团队关于关键事件或偏离预期行为的情况，例如未经授权的访问尝试或模型预测中的异常模式。

5. 与 SIEM 解决方案的集成：

a. 将审计日志和活动监控与安全信息与事件管理（SIEM）解决方案集成，以将 AI 相关事件与更广泛的安全事件和威胁情报相关联。

b. 利用 SIEM 功能进行日志聚合、关联、分析和报告，以深入了解 AI 系统的行为和安全状态。

6. 合规报告：

a. 审计日志支持合规报告要求，例如遵守监管标准（如 GDPR、HIPAA）或行业最佳实践（如 ISO 27001、NIST SP 800-53）。

b. 生成基于记录事件的审计报告和合规仪表板，为利益相关者提供 AI 系统活动和安全控制的可见性。

7. 保留与归档：

a. 制定审计日志的保留策略，确保数据保留足够的时间，以满足法律、监管和操作要求。

b. 实施归档机制，将旧的日志数据卸载到长期存储中，同时保持审计、分析和报告的可访问性。

通过实施强大的审计日志记录和活动监控机制，组织可以提高其 AI 系统的可见性、问责制和安全监督，从而实现有效的风险管理，并确保符合监管要求。

评估标准：

- 全面性： 审计日志必须捕捉多样化的事件，包括模型训练、数据处理和配置更改。

- 细节和粒度： 日志应提供详细的、粒度化的见解，以便每个事件的准确可追溯性和问责制。
- 安全性与隐私： 日志必须安全存储和管理，遵守数据保护法规。在发送日志管理解决方案之前，日志中的敏感数据必须被模糊处理或删除。
- 实时监控与警报： 系统应启用实时监控，并针对可疑活动发送警报。
- 集成与合规： 与 SIEM 解决方案无缝集成，并支持合规报告要求。

责任矩阵（RACI 模型）：

实施 AI 系统的审计日志记录和监控涉及各种角色和责任，使用 RACI 模型定义如下：

- 首席数据官（CDO）： 负责监督数据治理和合规，确保适当的审计日志记录和监控实践。
- 首席技术官（CTO）： 负责技术战略和实施，确保将审计日志记录和监控能力集成到 AI 系统中。
- 首席信息安全官（CISO）： 负责总体安全策略和风险管理，确保审计日志记录和监控符合安全最佳实践。
- 业务部门领导： 咨询了解业务需求，并保持对审计日志记录和监控实施的知情。
- 首席 AI 官员： 负责 AI 战略和实施，确保将审计日志记录和监控能力集成到 AI 系统中。

治理与合规：

- 数据保护官员： 负责确保符合数据保护法规，并就审计日志记录和监控要求提供咨询。
- 首席隐私官： 负责隐私合规，确保审计日志记录和监控符合隐私最佳实践。

- 法律与合规部门： 就法律和监管要求提供咨询，并保持对审计日志记录和监控实施的知情。
- 数据治理委员会： 就数据治理政策和标准提供咨询，并保持对审计日志记录和监控实施的知情。
- 合规团队： 负责监控和报告合规情况，并保持对审计日志记录和监控能力的知情。

技术与安全：

- IT 安全团队： 负责实施安全控制，并保持对审计日志记录和监控要求的知情。
- 网络安全团队： 负责网络安全措施，并保持对审计日志记录和监控能力的知情。
- 系统管理员： 负责系统管理和维护，并保持对审计日志记录和监控能力的知情。

运营与开发：

- AI 开发团队： 负责开发和实施 AI 系统，并保持对审计日志记录和监控要求的知情。
- DevOps 团队： 负责 DevOps 实践，并保持对审计日志记录和监控集成的知情。

管理与战略：

高层次实施策略：

- 集中日志存储： 使用可扩展、安全的平台进行日志存储，确保加密和适当的访问控制。

- 实时监控与警报： 实施复杂的监控工具，能够瞬时检测到异常，与 SIEM 集成以实现全面的安全监督。
- 合规报告与保留： 自动化合规报告，建立清晰的保留策略，并使用归档解决方案进行长期日志存储。
- 持续监控与报告： 建立持续的、实时的监控，使用自动化警报来识别潜在的安全威胁或操作问题。
- 访问控制： 实施严格的访问控制，确保只有授权人员可以查看或修改日志，保护敏感数据并维护合规。

适用框架和法规

使审计日志记录和监控实践与 NIST 指南保持一致，确保在 AI 系统中实现强大的治理、风险管理和合规性。

通过改进上述审计日志记录和监控实践，组织可以显著提升其 AI 系统的治理、风险管理和合规性，确保操作完整性、安全性和法规遵从性。这一全面的方法使组织能够保持高标准的问责制和透明度，在防范风险的同时，促进对 AI 应用的信任。

1.7 风险缓解

风险缓解是一种管理 AI 系统和操作中潜在威胁和不确定性的方法。它涵盖了处理风险的四种主要策略。首先是风险规避，涉及识别并完全消除高风险的 AI 应用或流程，从而防止风险的发生。其次，风险减少或缓解专注于实施控制和措施，以减少风险发生的可能性或如果发生其潜在影响。这可以包括技术保障、流程改进或增强的监控系统。第三，风险转移，涉及通过保险政策或合同协议将风险的潜在影响转移到其他方，从而保护组织免受风险带来的全部负面影响。最后，风险接受是一种有意的决策，经过仔细评估和成本效益分析后，选择承认并保留某些风险，通常是低影响风险。此策略通常在风险处理方法的成本高于风险本身的潜在影响时采用。

通过平衡和知情地使用这四种策略，组织可以有效管理与 AI 技术相关的复杂风险环境，确保强大的保护，同时仍然促进创新和进步。

1. 评估标准：

- 成功规避、缓解、转移或接受的识别风险百分比
- 与 AI 系统相关的事件数量和严重程度的减少
- 实施的风险缓解策略的成本效益
- 实施风险缓解措施所需的时间
- 风险重新评估和策略更新的频率
- 每种风险处理方法（规避、缓解、转移、接受）的有效性
- 风险管理程序的合规率

2. 责任矩阵（RACI 模型）：

- 执行： IT 安全团队、AI 操作团队
- 负责： 首席信息安全官（CISO）
- 咨询： 法律和合规部门、业务部门领导、AI 开发团队、首席技术官
- 知情： 管理层、首席 AI 官员、首席数据官

3. 高层次实施策略：

1. 开发全面的 AI 风险评估框架
2. 设立风险管理委员会，监督风险处理策略
3. 创建并维护风险登记册，按处理方法对风险进行分类

4. 为所有 AI 项目和系统实施定期风险评估周期
5. 为每种风险处理方法制定策略：
    - **规避：** 识别并消除高风险的 AI 应用或流程
    - **缓解：** 实施控制以减少风险的可能性或影响
    - **转移：** 为与 AI 相关的风险探索保险选项
    - **接受：** 定义接受低影响风险的标准

- 将风险处理的考虑因素整合到 AI 开发生命周期中
- 进行关于风险识别和处理方法的定期培训
- 建立决策制定协议，以选择适当的风险处理方法
- 实施一个系统，用于跟踪和报告风险处理工作的进展

4. 持续监控与报告：

1. 为关键的 AI 操作实施实时监控系统
2. 为每种风险处理方法建立关键风险指标（KRI）
3. 定期审计风险处理措施及其有效性
4. 开发一个仪表板，实现对风险状态和处理进展的实时可见性
5. 建立定期向管理层报告风险处理工作的系统
6. 实施反馈机制，以持续改进风险检测和处理策略
7. 建立一个流程，以便在新识别出的高影响风险发生时立即升级处理

5. 访问控制映射：

1. 限制对风险评估和处理计划的访问，仅限授权人员
2. 为风险管理系统实施基于角色的访问控制
3. 确保 IT 安全团队和 AI 操作团队具有适当的访问权限，能够监控并管理 AI 系统中的风险
4. 授权 CISO 和管理团队访问高级风险报告和仪表板
5. 为法律和合规部门提供对相关风险数据的访问，以进行法规合规目的
6. 允许 AI 开发团队有限访问与其项目相关的风险数据
7. 为包含敏感风险相关数据的系统实施严格的访问控制

6. 基础性指南：

- ISO 31000:2018 - 风险管理指南
- NIST SP 800-37 第 2 版 - 信息系统和组织风险管理框架
- COSO 企业风险管理框架
- 欧盟 AI 法案（拟议）- 包含基于风险的方法来监管 AI
- GDPR 第 35 条 - 高风险处理的数据保护影响评估
- NIST AI 风险管理框架 - 专用于 AI 系统的风险管理

1.8 数据漂移监控

数据漂移是指输入数据的统计特性随时间演变的过程。它发生在模型所训练的数据逐渐过时，不再适用于生产时。结果，模型的性能可能会下降。因此，主动的数据漂移监控在开发安全可靠的模型中至关重要。

数据投毒是由于对训练数据的故意污染而导致的一种数据漂移。

重要提醒： 模型性能会在没有明显信号的情况下逐渐衰退。这意味着模型输出必须定期检查，并在必要时重新训练。有效的机制也可用于检测与原始数据的偏差。

通常，有两种主要的数据漂移类型需要考虑：

- 协变量漂移： 这种情况发生在单个输入与输出的关系保持不变时，但输入数据的分布发生变化。协变量漂移可能由于用户行为、法规、数据收集因素等的变化而发生。
- 先验概率漂移： 当相对于训练数据，目标变量的分布随时间变化时发生。在这种情况下，输入特征与输出数据之间学到的关系被破坏。
- 模型性能还可能受其他类型的数据漂移影响，例如：
- 特征变化： 当特征发生变化时（如引入新特征或移除旧特征），会导致这种类型的数据漂移；
- 模型输出值范围的变化。

数据漂移监控 可以包括多种方法，推荐的方法包括：

- 相关领域的知识，有助于检测和调整模型性能以适应前沿趋势和特征重要性变化；
- 比较训练数据和新获得的数据中特征分布的统计测试（例如，Kolmogorov-Smirnov 测试、卡方检验、人口稳定性指数、Page-Hinkley 测试等）；
- 适用时的可视化分布比较，使用直方图、散点图等；
- 帮助检测数据漂移的专用算法；
- 除了上述方法外，还包括监控自动化管道、数据流图检查、数据来源的审查，以及定期检查数据质量和完整性的常规措施，以监控数据投毒攻击。

推荐的数据漂移监控实践包括：

- 确定一组需要监控的特征；
- 定义并描述参考数据。参考数据可以是训练数据，生产数据将与之进行比较；
- 为监控确定一个查找窗口；
- 为数据漂移监控设定和定义一组指标；
- 确定监控频率；
- 为指标设定阈值；
- 建立用于漂移检测的警报机制；
- 如果检测到显著的偏差，重新训练模型。

处理数据漂移的特定方法包括：

- **序列分析方法：** 实时监控数据流以检测变化。
    - **技术：**
        - **CUSUM（累积和控制图）：** CUSUM 通过累积偏差来监控流程均值的变化。
        - **漂移检测方法（DDM）：** DDM 监控模型性能指标（如错误率）的变化，并在检测到漂移时触发警报或更新。
        - **Page-Hinkley 测试：** 该测试检测数据流均值的变化，适用于实时监控。
- **基于模型的方法：** 使用模型处理漂移，通过适应或结合观察到的变化引入新策略。

- **技术：**
    - **集成方法：** 集成方法将多个模型的预测结合起来，并通过加权或替换模型来适应基于其在最近数据上的性能。
    - **自适应模型：** 这些模型随着新数据的到来逐渐更新，帮助处理漂移。
    - **概念漂移检测模型：** 这些模型设计用于检测概念漂移，例如 ADWIN，它通过调整窗口大小来维持性能。

基于时间分布的方法： 分析随时间推移的数据统计分布变化，以检测漂移。

- **技术：**
    - **Kolmogorov-Smirnov 测试：** 该测试比较两个数据集（当前与历史）的累积分布函数，以检测变化。
    - **基于直方图的方法：** 通过时间比较直方图，可以检测特征分布的变化。
    - **核密度估计（KDE）：** 该测试估计随机变量的概率密度函数，并帮助检测数据分布随时间的变化。

强烈建议将数据质量监控机制与数据漂移监控结合使用。数据漂移监控和数据质量监控需要与能够定义一致要求的数据科学家共同建立（关于责任分配的详细信息，请参见下面的 RACI 模型）。

1. 评估标准： 组织应制定一套可量化的指标，用于评估数据漂移监控。必须通过一致的警报机制监控输入数据分布和总体模型性能。

2. RACI 模型： 应该识别利益相关者、角色和责任。通过设置执行（Responsible）、负责（Accountable）、咨询（Consulted）和知情（Informed）的人员，避免职责重复或漏洞。

以下分配可能是有益的：

- **执行：** AI 运营主管、AI 维护团队、AI 操作团队、AI/ML 测试团队、质量保证团队、网络安全团队、IT 安全团队、硬件安全团队；
- **负责：** 首席数据官、首席 AI 官、首席信息安全官（在职责范围内）；
- **咨询：** 数据保护官、数据治理官、数据科学团队；
- **知情：** 知情的利益相关者列表必须与组织的 AI 相关流程保持一致。

3. 高层次实施策略： 高层次实施策略需要与公司的整体数据战略一致。

4. 持续监控与报告： 实施持续监控。警报、数据质量仪表板、模型性能监控和定期数据审核是持续监控活动的例子。必须定义负责持续监控数据漂移的角色，并根据 RACI 模型为利益相关者生成定期报告。

5. 访问控制： 必须为输入和输出数据及数据漂移监控活动建立访问控制机制，以避免潜在的对抗性方对数据进行投毒。

6. 适用框架和法规： NIST AI 风险管理框架（NIST AI RMF）、微软负责任 AI 标准。

2. 治理与合规

治理与合规构成了指导组织内 AI 系统负责任开发、部署和使用的结构框架。本部分深入探讨了建立和维护强大 AI 治理结构的多方面内容，同时确保遵守相关法规和标准。这包括制定全面的 AI 安全政策、实施严格的审计流程、建立明确的董事会报告机制以及应对复杂的监管要求。此外，它还探讨了创建可衡量和可审计控制的过程，新兴法规的影响（如欧盟《人工智能法案》和美国《人工智能行政命

令》），AI 使用政策的制定以及模型治理的实施。通过解决这些关键领域，组织可以在减少风险的同时，创造一个问责、透明、符合伦理的 AI 使用环境，并确保在不断变化的法律和监管环境中保持合规。

2.1 AI 安全政策、流程和程序

定义、发布和管理支持安全和负责任的 AI 实践的安全政策、流程和程序，应该与现有的网络安全政策和程序互补并互操作。这些流程和程序还应与公司层面的 AI 负责任政策一致，以确保与其他核心学科（如数据隐私、伦理、法律合规等）的一致性和互操作性。

一个组织可以选择将其网络安全相关的流程和程序与全公司的政策保持一致，比如应用于每个开发、评估或部署 AI 角色的全公司 AI 原则，或者是公司的 AI 负责任或 AI 伦理政策。因此，这些原则的应用将确保新的和正在发展的技术解决方案从企业标准和流程的角度受到保护。

政策应该从高层次上传达公司对使用这些新兴技术的立场。

A. 定义一个流程，使其与高层次的政策保持一致，并从 AI 项目启动到整个应用生命周期内的持续生产监控和使用案例更新中嵌入网络安全。

在为企业政策建立“顶层基调”要求后，应开发与政策相关的流程，描述为满足政策目标将采取的步骤。

流程不应过于严格，以免未来的用例超出范围，因缺乏充分的尽职调查而导致负面结果。AI 的社会技术用例迅速发展，可能会产生新漏洞，如果不及早评估，可能会产生广泛的负面影响。标准应以敏捷的方式定义，并能支持未来框架的迭代（如 NIST AI 风险管理框架）。

流程及其相关程序描述了网络安全团队如何通过其评估角色、工具和治理结构为负责任的 AI 做出贡献，以支持每个项目的审查。以下领域应被视为任何标准和相关程序的最佳实践：

1. 识别和评估风险。
2. 定义安全目标（可根据用例背景及其预期结果、数据来源、政策风险容忍度等进行调整）。
3. 建立安全控制。
4. 发布并定期审查和更新治理流程。
5. 为内部和外部角色提供评估、审查和审批流程的培训和教育。
6. 通过“反馈循环”持续监控和评估安全性，以解决内部或与外部利益相关者的潜在伦理或网络安全问题。
7. 定义并遵循事件响应和恢复计划与手册。

与政策一致的流程和程序还应考虑实施检查点和护栏，以进行评估和风险缓解（参考 NIST 测试、评估和红队测试）。

● 应用整个生命周期中的安全测试，使用测试、评估、验证和验证（TEVV）指导，考虑以下因素：

- 测试与评估（T&E）对于评估 AI 模型和系统的有效性和安全性至关重要，这些模型和系统是解决方案架构的一部分，涉及目标用例和应用限制。漏洞、弱点和潜在威胁应记录在漏洞评估、渗透测试和合规性测试中。
- 验证应包括模拟攻击和对抗性防御的红队测试。红队测试和威胁建模可以评估控制措施以防止数据泄露、获取未授权访问或利用数据或模型中的漏洞（包括时间上发生的任何建议更改）。

- 验证步骤还应考虑应用程序和用例背景中的偏差。从网络安全角度来看，评估数据源中的偏差可能会产生负面结果，必须在作出进一步决定之前进行严格的测试。该流程还必须评估模型是否足够具有弹性，以抵御或易受社会工程攻击的影响，这些攻击可能会导致输出错误或不准确，并且应预见到使用 AI 的风险超出文档用例的意图。

B. 流程和详细的程序必须与或定义治理结构和角色保持一致，评估、减轻或批准项目继续进行的风险，并防止项目在未应用额外控制措施之前继续进行。

流程和程序还应考虑风险指标和度量，以测量合规性和风险容忍度，并确保每季度/年度输入对网络安全计划的有效性做出贡献，作为对负责任 AI 的关键贡献。

1. 评估标准：

组织应建立量化指标以评估其 AI 计划的有效性，其中应包括网络安全的具体指标，但也可以包括其他学科（如法律和合规、数据隐私利益相关者、监管机构等）的指标；指标可能包括已识别威胁的数量、验证和验证阶段中漏洞的严重性，以及 AI 注册表中所有应用程序的风险级别。

2. RACI 模型：

RACI 模型有助于澄清关于安全应用负责任 AI 的流程和相关详细程序中的角色与责任。关键人员必须被指定为执行（Responsible）、负责（Accountable）、咨询（Consulted）或知情（Informed），以确保在网络安全评估和风险缓解的测试、评估、验证和验证（TEVV）阶段进行明确的监督和问责。

RACI 模型还应考虑治理结构，无论是直接概述在公司层面的政策中，还是通过网络安全标准进行链接。角色应定义为委员会的集中或分散责任，用于审查和批准任何项目内的政策和流程，谁有权质疑、谁知情等。

3. 高级实施策略：

负责任 AI 的治理策略实施应包括定期的网络安全培训和意识、事件响应手册以及确保持续监控和测试的一套反馈机制。该策略应在整个组织中始终如一地应用。该策略还应包括与内部和外部利益相关者的定期互动，根据需要，适当的行业信息共享协议，以便随时了解不断涌现的威胁、趋势和评估工具。

组织的政策、流程和程序更新结构应以优先更新为准，如果需要，至少有一个年度更新、审查和由政策委员会批准的时间表（包括网络特定或企业范围内涉及所有负责任 AI 学科的委员会）。

4. 持续监控和报告：

持续监控工具和报告机制对于维护应用程序及其用例上下文的完整性至关重要。度量应检测任何未经过重新评估、审查/批准流程和护栏的初始批准的任何偏离。

5. 访问控制：

访问控制机制对于保护数据和访问模型及应用程序至关重要。组织必须实施强有力的控制措施以管理对敏感数据、模型注册表及涉及威胁建模的其他关键资产的访问。还必须有事故响应手册，如果需要，实施适当的治理步骤以关闭应用程序，直到问题得到解决。

6. 适用框架与法规：

- NIST AI 风险管理框架
- NIST 安全软件开发框架
- 关于安全、可信及负责地开发和使用人工智能的行政指令
- 欧盟人工智能法案（最终草案 2024）

2.2 审计

AI 审计指的是系统地检查和评估人工智能系统、其底层算法及其部署。AI 审计的主要目标是确保合规性、促进透明度，并维护伦理使用。在 AI 审计期间，将检查多个方面，包括风险评估、数据治理、模型评估、伦理考虑和法律合规性。审计员会验证是否遵守相关标准、指南和法规，以维护 AI 系统的信任和问责制。

AI 审计的一些关键组成部分是：

- **风险评估**：评估 AI 系统的风险，包括偏见、隐私侵犯、安全漏洞等。
- **透明度和可解释性**：评估 AI 系统的透明度和可解释性。
- **数据治理**：检查数据质量、数据来源和数据预处理。
- **模型评估**：使用适当的指标评估 AI 模型的性能。
- **伦理考虑**：审查 AI 部署的伦理影响。
- **法律和法规合规性**：确保遵守相关法律（如 GDPR、CCPA 等）。

AI 审计是一个不断适应技术进步和不断变化的伦理规范的过程。组织和审计员在维护 AI 系统的信任和问责制方面至关重要。

1. 评估标准：使用特定的指标评估每个 AI 审计领域。以下是一些示例：

- **风险评估**：识别的技术风险的数量和严重性（例如，精度错误、模型漂移）。
- **透明度和可解释性**：具有可解释说明的 AI 模型的百分比。
- **数据治理**：数据质量分数基于完整性、准确性和一致性。
- **模型评估**：与 AI 系统目的相关的性能指标（如准确性、精度）。

- **伦理考虑：**AI 部署是否符合伦理准则和原则。
- **法律和法规合规性：**识别的法律和法规差距的数量。
- **审计评估的董事会指标：**董事会可以使用以下指标评估 AI 审计的有效性：
    - 审计提供的可行性见解和建议
    - 审计和报告的及时性
    - 管理层对审计发现的承诺和投入
    - 审计后 AI 治理实践的可衡量改进

通过使用这些指标，组织可以确保其 AI 审计的严谨性和信息性，董事会可以有效评估 AI 系统的可信度和伦理实施情况。

2. RACI 模型

以下表格概述了与审计 AI 系统相关的关键领域的 RACI 模型：

活动	执行 (R)	负责 (A)	咨询 (C)	知情 (I)
风险评估				
识别技术风险	AI 项目团队（负责人）	首席技术官（CTO）	数据科学与安全团队	董事会、业务单元管理
识别非技术风险	法务部门（负责人）	首席风险官（CRO）	伦理委员会	董事会、业务单元管理
透明度与可解释性				
评估模型的可解释性	数据科学团队（负责人）	AI 项目负责人	业务单元负责人	董事会、利益相关者
数据治理				
数据质量与来	数据治理团队	首席数据官	数据科学团	董事会、业务

活动	执行 (R)	负责 (A)	咨询 (C)	知情 (I)
源审查	（负责人）	（CDO）	队，法务	单元管理
培训数据偏差评估	数据科学团队（负责人）	AI 项目负责人	伦理委员会	董事会
数据隐私合规审查	法务部门（负责人）	首席隐私官（CPO）	数据治理团队	董事会
模型评估				
性能指标与分析	数据科学团队（负责人）	AI 项目负责人	业务单元负责人	董事会
公平性与偏差分析	数据科学团队（负责人）	首席数据官（CDO）	伦理委员会	董事会
对抗性鲁棒性测试	安全团队（负责人）	首席技术官（CTO）	数据科学团队	董事会
伦理考虑				
伦理影响评估	伦理委员会（负责人）	首席风险官（CRO）	法务，业务单元管理	董事会
符合伦理准则	法务部门（负责人）	首席执行官（CEO）	伦理委员会	董事会
法律与法规合规				
法律与法规审查	法务部门（负责人）	首席合规官（CCO）	业务单元负责人	董事会
整体审计				
进行内部审计	内部审计团队（负责人）	首席审计执行官（CAE）	根据需要的部门	董事会（审计委员会）
聘请外部审计员（可选）	管理层（负责人）	风险委员会董事会	内部审计团队	董事会

3. 高级实施策略： 有效的 AI 审计需要在组织结构内明确职责划分，并专注于可信 AI 使用的关键领域。以下是实施步骤：

1. **定义审计范围：** 确定哪些 AI 系统和流程将被审计。
2. **指定审计所有权：** 在确定完成审计目标所需的资格后，评估 IA（内部审计）团队。
3. **开发审计方法：** 定义具体程序和技术来评估审计范围内的 AI 特定领域。
4. **开发审计指标：** 识别关键指标，重点关注模型性能、公平性、偏差和伦理影响等关键方面。
5. **报告与跟进：**

   a. 建立清晰的报告结构，将审计发现和建议传达给相关方（如管理层、董事会）。

   B. 定义流程，解决已识别的问题，并实施纠正措施以提高 AI 系统的可信度.

4. 持续监控和报告：

虽然持续监控和报告对于维持组织内整体 GRC 至关重要，但这里的重点是 AI 审计。AI 审计是一个专门的系统化流程，用于评估 AI 系统、其算法和其部署。与持续监控不同，AI 审计提供了对特定方面（如风险评估、数据治理、伦理考虑）的深入检查。这种全面的评估确保合规性，促进透明度，并支持 AI 技术的伦理使用，最终培养对 AI 系统的信任和问责。

内部审计（IA）不会直接执行持续监控。IA 将审查 AI 系统生成的输出，以确保其按预期运行并识别潜在问题，相关内容包括：

- **数据质量：** IA 审查数据完整性，识别缺失的数据点或可能影响 AI 训练和决策的数据缺口。

- **模型性能：** IA 评估准确性、精度和召回率，以确保 AI 系统一致地执行并达到既定基准。
- **公平性与偏差：** IA 检查 AI 输出中潜在的偏差报告。
- **可解释性与透明性：** IA 审查和评估 AI 的解释的一致性和理解性，确保人类用户能够理解其决策依据。
- **安全漏洞：** IA 审查 AI 系统及其部署环境中潜在的安全漏洞报告。
- **控制有效性：** IA 评估现有控制措施的有效性，以减轻与 AI 系统相关的风险。
- **变更管理：** IA 审查组织的 AI 系统变更管理流程。

通过审查持续监控系统，IA 可以深入了解 AI 系统的整体健康和有效性。这使得 IA 能够评估组织的 GRC 要求合规性，并确保负责任且符合伦理地使用 AI 技术。

5. 访问控制：

AI 系统周围的安全措施包括模型注册表、数据存储库和特权访问点的访问控制。强有力的访问控制可以减轻与未经授权的访问或误用关键资源相关的风险。在 AI 审计期间，审计员将评估这些控制措施的有效性，以保护敏感数据并确保符合相关法规。

- **模型注册表：**
  - **用户访问控制：** 审查谁可以注册、修改或删除 AI 模型。
  - **认证方法：** 评估访问模型注册表的认证方法的强度。
  - **审计与日志记录：** 确认访问尝试和模型修改的日志记录，以确保问责制和异常检测。
- **数据存储库：**
  - **数据访问控制：** 审查谁可以访问用于训练和操作 AI 系统的数据。

- **数据安全控制：** 评估静态和传输中的数据加密，以保护敏感信息。
- **审计与日志记录：** 确认数据访问尝试和修改的日志记录，以跟踪目的和安全漏洞。

- **特权访问点：**
  - **用户访问控制：** 审查谁拥有管理或配置 AI 系统的特权访问。
  - **最小权限原则：** 确保特权用户只拥有完成任务所需的最低访问权限。
  - **多因素认证：** 确认特权访问点的强认证方法是否到位。
  - **审计与日志记录：** 验证特权用户活动的全面日志记录，以确保问责制和安全监控。

通过审查这些访问控制措施，IA 可以评估组织在减轻与未经授权的 AI 模型、数据和关键功能相关的风险方面的努力。这有助于确保符合相关数据保护法规，并促进负责任地使用 AI 技术。

6. 适用框架与法规：

- NIST AI RMF
- 美国总统《安全、可信和负责任的人工智能执行命令》
- 欧盟《人工智能法案》（2024 年最终草案）
- GDPR
- CCPA
- CPRA
- ISO/IEC 27701:2019（隐私信息管理系统）

- 内部审计师协会（IIA）AI 审计框架
- 经济合作与发展组织（OECD）AI 原则，审计人工智能
- ISO/IEC 42001:2023 人工智能管理系统
- ISO/IEC 23053:2022 使用机器学习 (ML) 的人工智能 (AI) 系统框架
- 联合国-《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》

2.3 董事会报告

董事会监督组织内外 AI 的伦理和有效使用。履行这一职责需要对 AI 实施的整个生命周期有全面的理解，从其目的和潜在风险到与整体业务战略的一致性。这意味着报告要求将重点放在治理和监督上，包括通过定期的绩效报告和利益相关者披露，建立负责任的 AI 框架以及透明度和问责机制。

治理与风险

理解 AI 的使用：

- 董事会应了解 AI 在公司内的使用情况。
- 这包括了解 AI 系统的目的、潜在风险以及与业务战略的一致性。

AI 政策与框架：

- 董事会应批准负责任的 AI 使用框架。
- 该框架应解决偏见、公平性、安全性、透明度、伦理和社会影响问题。
- 董事会应考虑 AI 对社会的潜在影响、公平性及与公司价值观的一致性。

风险管理与合规性：

- 董事会应确保有流程识别、评估和减轻与 AI 相关的风险。
- 这涉及将特定的监督职责分配给委员会，如审计委员会。

透明度与问责

AI 绩效报告：

- 董事会应定期收到 AI 系统绩效的报告。
- 这些报告可以包括关于准确性、效率和改进潜在领域的指标。

向利益相关者披露：

- 董事会可能需要考虑向利益相关者披露 AI 使用信息的适当程度。
- 这可能包括潜在的影响、伦理考虑和监管要求。

有效的董事会报告确保了透明度、问责性以及有关 AI 采用的知情决策。

1. 评估标准：有效的董事会对 AI 实施的监督需要全面报告治理、风险和合规 (GRC) 实践。评估重点在于详细报告 AI 系统的目的、潜在风险及其与整体业务战略的频率和清晰度。

- **治理与风险**
    - **AI 政策与框架：** 确保有记录的、负责任的和有效的 AI 框架存在。
    - 将伦理考虑、偏见缓解策略和潜在的社会影响评估纳入框架，并使其与公司价值观一致。
- **风险管理与合规性：**
    - 评估是否存在明确的流程来识别、评估和减轻 AI 风险。

- 明确分配监督责任（如分配给委员会）以及是否符合相关 AI 法规的证据。

透明度与问责

- **AI 绩效报告：**
  - 评估 AI 绩效报告的定期性和详细程度。
  - 这些报告应包括准确性、效率和改进领域的指标。
- **向利益相关者披露：**
  - 评估向利益相关者披露 AI 使用信息的适当性。
  - 这包括潜在的影响、伦理考虑和监管要求。

通过应用这些评估标准，组织可以评估董事会报告在 GRC 和 AI 方面的有效性。这可以促使负责任的 AI 实施和监督的持续改进。

2. RACI 模型：以下表格概述了与 AI 系统 GRC 相关的关键领域的 RACI 模型：

活动	执行 (R)	负责 (A)	咨询 (C)	知情 (I)
治理与风险				
了解 AI 的使用	AI 项目团队（负责人）	首席技术官（CTO）	业务单元负责人	董事会
AI 政策与框架开发	法务部门（负责人）	首席风险官（CRO）	伦理委员会	董事会
外部审计与独立评	首席审计执行官（CAO）	首席执行官（CEO）	董事会（审计委员会）	董事会

活动	执行 (R)	负责 (A)	咨询 (C)	知情 (I)
估				
风险管理与合规	首席风险官（CRO）	董事会（风险委员会）	法务部门，IT 安全团队	业务单元负责人，AI 项目团队
透明度与问责				
AI 绩效报告	AI 项目团队（负责人）	业务单元负责人	数据科学团队	董事会
向利益相关者披露	通讯部门（负责人）	首席执行官（CEO）	法务部门，业务单元负责人	董事会，利益相关者

该 RACI 模型阐明了有效 GRC（治理、风险和合规）在 AI 系统中的角色和责任。** 组织可以通过明确的所有权和沟通渠道，确保在 AI 治理中采取透明和问责的方式。

3. 高级实施策略：董事会在监督组织负责任且有效地使用 AI 方面至关重要。这意味着将治理、风险和合规（GRC）原则整合到组织层面的 AI 实施中。

治理与风险

AI 政策与框架：

- 董事会应建立一个负责任的 AI 框架，概述伦理考虑、偏见缓解策略和风险管理实践。
- **风险管理与合规：**
    - 他们定期收到解释 AI 使用情况的报告，包括目的、风险及其与业务战略的一致性。
    - 他们实施流程来识别、评估和减轻 AI 风险。

- 这可能涉及将监督责任分配给特定的委员会。
- 董事会确保遵守与 AI 使用相关的法规和行业标准。

透明度与问责：

AI 绩效：

- 建立、审查和评估绩效报告及利益相关者披露。

利益相关者披露：

- 建立、审查和评估利益相关者披露报告。

有效的董事会监督对于负责任且成功的 AI 实施至关重要。通过解决这些障碍并将 GRC 原则整合到 AI 治理中，组织可以确保 AI 技术的伦理、安全和合规使用。

4. 持续监控与报告： 持续监控和全面报告对于董事会有效监督 AI 实施的伦理性和可持续性至关重要。下面是**董事会报告**的示例，概述了关键指标、职责分配和负责任的 AI 实施及风险管理的参考框架。 遵守这一框架至关重要，因为它使组织能够展示他们对组织内伦理和有效使用 AI 的承诺。通过定期报告这些指标，组织可以确保 AI 倡议的透明性和问责制，并识别改进领域及缓解潜在风险的措施。

类别	指标	描述	报告责任	参考框架
治理与监督	拥有批准的治理计划的 AI 项目百分比	遵守已建立的治理框架和政策	AI 治理委员会	欧盟 AI 法案
透明度与问责	AI 模型可解释性分数	衡量 AI 模型的可解释性和可理解性	AI 开发团队	Google 安全 AI 框架 (SAIF)

类别	指标	描述	报告责任	参考框架
安全与风险管理	AI 安全漏洞的数量	AI 系统中的安全漏洞或缺陷	安全分析师	MITRE ATLAS™（人工智能系统对抗性威胁）
数据隐私与保护	数据匿名化率（百分比）	AI 训练数据集中匿名的个人可识别信息 (PII) 的比例	数据隐私官	《通用数据保护条例》（GDPR）
数据隐私与保护	数据最小化率（百分比）	AI 训练数据集中最小化的个人可识别信息 (PII)	数据隐私官	《通用数据保护条例》（GDPR）
伦理使用与公平性	AI 模型公平性分数	评估不同维度上公平性和偏差的缺失	数据科学家	OWASP 机器学习 - 前十

5. 访问控制：董事会负责确保在组织内安全使用 AI，访问控制在降低风险和促进信任方面起着至关重要的作用。

治理与风险

理解 AI 的使用：

- 董事会应了解访问控制的设计如何支持每个 AI 系统的战略目标，并确保与整体商业战略保持一致。
- **AI 政策与框架：**
    - 董事会应批准 AI 框架，概述访问控制的原则和最佳实践。

- 董事会应讨论偏见和访问控制缓解策略的有效性。
- 董事会应讨论 AI 开发和部署中的伦理考量，并确保访问控制能够防止 AI 被滥用于不道德的目的。

风险管理与合规：

- 董事会最终负责识别和减轻整个组织内的 AI 风险。
- 访问控制是防止未经授权访问和数据泄露的关键风险缓解策略。

透明度与问责

AI 绩效：

- 董事会应了解访问控制如何影响性能。
- 过于严格的访问控制可能会阻碍协作并减慢优化流程。

利益相关方披露：

- 董事会应在为不同利益相关方定制的报告中解决访问控制问题。
- 这可能涉及突出访问控制如何保护敏感数据并促进负责任的 AI 使用。

6. 适用的框架与法规

- NIST 人工智能风险管理框架
- 《未来高管的治理：负责任的 AI 治理》
- 《人工智能：审计委员会的新兴监督责任？》
- 《第六部分：AI 系统的负责任企业治理》
- 《如何设计 AI 伦理委员会，德勤：董事会成员的 AI 治理》
- 《普华永道：AI 与生成式 AI 的力量——董事会需要了解的内容》

2.4 法律监管要求 - 法律

大规模使用 AI 技术对社会、经济以及伦理考量有着深远的法律影响。这些技术从增强医疗诊断到优化金融服务，具备显著的创新和效率潜力。然而，它们的快速发展也带来了重要的法律挑战。

以下是一些 AI 相关的法律要求类型及当前示例，以及不合规的后果：

1. **数据保护法规**：如欧洲的 GDPR 或美国的 CCPA 等法律对 AI 系统中个人数据的收集、处理和使用施加了限制。这些法规要求 AI 系统遵守数据保护、同意和透明度原则。例如，GDPR 明确规定了对 AI 驱动的数据处理的明确同意、安全措施和透明度要求，若不遵守规定，可能导致最高 2000 万欧元或全球营业额的 4%罚款。

2. **伦理指南**：某些地区引入了专门针对 AI 的伦理指南或原则，涵盖公平性、透明度、问责制和包容性等领域，用于 AI 的开发和部署。例如，电气与电子工程师学会（IEEE）全球倡议提供了 AI 设计的伦理原则，重点在于公平性、透明度和问责制。不合规可能面临伦理审查和声誉损害的风险。

3. **算法问责法**：这些法律要求组织解释和证明 AI 系统做出的决策，特别是当这些决策对个人有重大影响时。这些法规旨在确保自动决策过程中的透明性和公平性。例如，欧盟 GDPR 和美国算法问责法要求 AI 驱动的决策在透明性和问责制方面达标，否则可能面临法律诉讼和公众信任的损失。

4. **安全和保障标准**：法律要求 AI 系统达到某些安全标准，以最小化对用户或社会造成伤害的风险。法规可能适用于关键基础设施、医疗保健、交通或金融领域，确保 AI 的可靠性并防止事故或恶意使用。例如，美国的《AI 行政命令》要求进行标准化评估和风险缓解，不合规可能导致法律诉讼和声誉损害。

5. **责任与问责**：法律明确了在 AI 系统造成伤害或错误时的责任分配，开发人员、部署者或用户需对 AI 系统的行为负责。例如，加拿大的 PIPEDA 改革建议为 AI 建立全面的监管框架，确保法律合规并保护消费者权益。不合规将面临法律风险和声誉损害。

6. **监管审批**：在某些行业，AI 系统在部署前可能需要监管审批，以满足特定的安全、有效性或可靠性标准。例如，美国食品药品监督管理局（FDA）的预认证计划确保 AI 医疗设备的安全标准，不合规可能导致市场准入障碍和潜在处罚。

7. **反歧视措施**：法律可能禁止基于受保护特征（如种族、性别或年龄）的歧视，特别是在招聘、贷款或执法等敏感环境中。例如，《公平住房法》禁止歧视性 AI 算法，确保平等的住房机会，不合规将导致法律后果和声誉损害。

8. **国际协议**：与 AI 相关的法律要求可能通过促进跨国合作、标准化和法规协调的国际协议或条约来建立。例如，经济合作与发展组织（OECD）的《AI 原则》提供了国际伦理指南，促进 AI 政策的合作与对齐。不遵守协议可能带来外交压力和贸易壁垒。

评估标准

- 符合相关数据保护法规的 AI 系统百分比
- 成功的外部 AI 系统合规审计次数
- 内部合规审查的频率
- 解决已识别合规问题所需的时间
- AI 决策过程的透明度水平
- 报告的伦理违规或偏见事件数量

- 完成算法影响评估的 AI 系统百分比

责任矩阵 (RACI 模型)

角色	责任
执行 (Responsible)	法律和合规部门，数据保护官
负责 (Accountable)	首席隐私官
咨询 (Consulted)	AI 开发团队、业务部门领导、IT 安全团队
知情 (Informed)	管理层，运营人员

高层实施策略

1. 制定符合相关法规的全面 AI 合规框架。
2. 在所有 AI 系统中实施强有力的数据保护措施。
3. 为 AI 开发和部署建立伦理指南。
4. 创建算法问责与透明性流程。
5. 为关键领域的 AI 系统制定安全和保障标准。
6. 建立清晰的责任分配和事故处理协议。
7. 实施必要的监管审批流程。
8. 为 AI 系统开发反歧视措施。
9. 监控和适应国际 AI 协议和标准。

持续监控与报告

1. 定期进行 AI 系统的内部审计以确保合规性。
2. 监控影响 AI 的相关法律法规的变更。

3. 追踪并报告 AI 系统中的伦理问题和偏见事件。
4. 定期评估 AI 决策过程的透明度和可解释性。
5. 监控 AI 系统中的数据保护措施的有效性。
6. 为管理层和相关监管机构生成合规报告。
7. 定期审查 AI 系统的影响评估。

访问控制映射

1. 限制对 AI 系统中使用的敏感数据的访问，仅限授权人员。
2. 实施基于角色的访问控制，确保合规相关的文档和系统的安全。
3. 确保只有合格人员可以修改 AI 算法和模型。
4. 限制对 AI 系统审计日志和合规报告的访问，仅限授权个人。
5. 对处理个人数据的系统实施严格的访问控制，符合数据保护法规。

适用框架与法规

- **通用数据保护条例 (GDPR)**：欧盟的全面数据保护法律，全球范围内影响处理个人数据的 AI 系统。
- **加州消费者隐私法案 (CCPA)**：监管加州企业收集和使用消费者数据，包括 AI 应用。
- **AI 法案（欧盟）**：建立了基于风险级别的 AI 系统综合监管框架。
- **算法问责法案（美国提议）**：要求公司评估并减轻 AI 系统的风险。
- **FDA 法规**：监管 AI 在医疗设备和医疗应用中的使用。
- **公平住房法案**：禁止住房中的歧视性 AI 算法使用，确保平等的住房机会。

- **经合组织（OECD）AI 原则：**提供 AI 负责任开发的国际指南。

2.5 实施可测量/可审计的控制措施

实施审计控制的主要目标是验证 AI 系统在各个阶段都采取了必要的措施和步骤，确保其影响符合现有法律、信任和安全最佳实践以及社会期望。

实施可测量/可审计的控制措施包括定义风险及其相应的控制措施，随后执行控制的过程。控制的衡量可以体现在每个风险类别中的控制数量和控制是否符合相关政策或程序。

以下是该责任项的六个跨领域关注点：

评估标准

- 确定的 AI 风险具有相应控制措施的百分比
- 每个风险类别实施的控制数量
- 控制有效性评估的频率
- 控制符合或超过政策要求的百分比
- 为应对已识别风险实施新控制的时间
- 成功的内部和外部审计数量
- 完整审计记录的 AI 系统的百分比
- 控制监控和报告中的自动化水平

责任矩阵 (RACI 模型)

角色	责任

角色	责任
执行 (Responsible)	IT 安全团队、AI 开发团队
负责 (Accountable)	首席信息安全官（CISO）
咨询 (Consulted)	法律和合规部门，数据保护官
知情 (Informed)	管理层、业务部门领导、运营人员

高层实施策略

1. 对所有 AI 系统进行全面的风险评估。
2. 制定符合已识别风险和监管要求的控制框架。
3. 在 AI 开发和部署的各个阶段实施控制。
4. 为每个控制建立可测量的指标。
5. 创建自动化监控系统，持续评估控制措施。
6. 为 AI 系统活动开发审计记录和日志机制。
7. 实施常规的控制有效性审查和改进流程。
8. 建立应对新风险的快速控制实施流程。

持续监控与报告

1. 实施实时监控控制有效性。
2. 定期自动扫描 AI 系统，以确保符合控制要求。
3. 生成定期报告，显示控制的表现和差距。
4. 监控控制有效性随时间的变化趋势。
5. 跟踪并报告新控制措施的实施状态。

6. 定期对控制框架进行内部审计。

7. 建立控制失败或重大偏差的警报。

访问控制映射

1. 限制对控制实施和修改的访问，仅限授权人员。

2. 实施基于角色的访问控制，用于监控和报告系统。

3. 确保在控制实施和审计中的职责分离。

4. 限制对审计日志和控制有效性报告的访问，仅限授权个人。

5. 对涉及风险评估和控制管理的系统实施严格的访问控制。

适用框架与法规

- **ISO/IEC 27001**：提供信息安全管理的框架，包括风险评估和控制实施。
- **NIST 网络安全框架**：提供管理和减少网络安全风险的指南。
- **COBIT（信息和相关技术的控制目标）**：为 IT 治理和管理提供了全面的框架。
- **SOC 2（系统和组织控制）**：定义了基于五个"信任服务原则"的客户数据管理标准。
- **GDPR 第 25 条**：要求通过设计和默认数据保护，采用适当的技术和组织措施。

2.6 欧盟 AI 法案，美国行政命令：开发安全、可信的 AI 等

政策环境的变化引入了新的监管要求和最佳实践。欧盟 AI 法案和美国关于安全、可信开发和使用人工智能的行政命令均包含重要的监管要求、标准和最佳实践，组织应当注意。

1. **评估标准：**量化指标对于评估 AI 系统的治理、风险管理和合规性（GRC）至关重要。利益相关者需要衡量监管合规性、风险暴露情况与组织政策的对齐，以确保在 AI 技术中的强大 GRC 实践。

2. **RACI 模型：**角色和责任的明确对 AI 系统中有效的 GRC 至关重要。RACI 模型提供了一个结构化的框架，用于定义在 GRC 决策和监督中谁是执行（Responsible）、负责（Accountable）、咨询（Consulted）和知情（Informed）。这种划分确保了整个 AI 生命周期中的责任制和透明度。

3. **高层实施策略：**阐述 GRC 责任在组织层面的实施方式以及必须克服的障碍。

4. **持续监控与报告：**持续的监控和报告机制对于维护 AI 系统中的 GRC 完整性非常重要。实时监控、安全事件合规性违规警报、审计记录以及定期报告确保了透明度和问责制。这些实践使组织能够及时识别和解决与 GRC 相关的问题。

2.7 AI 使用政策

创建政策

所有组织都应实施 AI 使用政策。如果有助于与员工沟通，此政策可以作为现有文档的一部分实施，例如可接受使用政策 (AUP)。根据贵组织的产品和服务以及适用的法规，可能更适合制定单独的政策，详细说明开发 AI 驱动服务时的要求。在这种情况下，应创建一份专门针对员工的独立文档。

该政策及其影响应为组织的员工所知。当政策发布时以及当有重大更新时，员工应接受培训。

政策的制定过程应包括业务、法律/合规和技术人员，以确保所有要求和目标都得到考虑。还应包括组织的隐私专家，以确保 AI 政策不会与隐私政策重叠或相矛盾。

由于监管环境发展迅速，可能需要来自多个地区的法律专家参与，以确保所有监管要求都得到考虑。

政策内容

AI 使用政策应传达组织内允许和受限的 AI 使用案例。所有员工应理解使用的语言，文档中最好包含简单的示例，以确保信息被理解。

建议在政策中包含以下主题：

- **目的**：组织想要强调和推动的高级目标。
- **范围**：政策的预期范围及其可能参考的其他组织政策。
- **治理与责任**：
    - 谁负责组织内的 AI 使用，谁可以联系？是否有 AI 委员会或合规官提供指导？
    - 每个员工在考虑使用 AI 技术时的权利和责任是什么？员工不得做哪些事情？
    - AI 使用相关事件和违规应如何以及向谁报告？
- **负责任的 AI 使用**：
    - 组织在使用 AI 技术时遵循的伦理原则。
    - 其他原则和内部要求限制或指导 AI 技术的使用。
- **合规与合同**：适用于组织 AI 使用的关键法规和合同要求的列表和解释。
- **执行**：如认为合适，可以包括一条说明，违反此政策可能导致纪律处分。

评估标准

- 确认并完成 AI 使用政策培训的员工百分比
- 政策审查和更新的频率
- 报告的 AI 使用政策违规次数
- 解决政策违规问题所需的时间
- 员工对政策理解的水平（通过调查或评估测量）
- 展示政策合规性的成功审计次数
- 员工政策咨询的频率
- 经历过政策合规审查的 AI 项目数量

2.8 模型治理

模型治理是一项关键实践，确保 AI 模型在整个生命周期内得到负责任且有效的管理。它包括制定政策、程序和控制措施，以管理组织在 AI 模型开发、部署、监控和维护方面的活动。

AI 模型治理的生命周期通常包括以下阶段：

1. **发现与规划**：此阶段涉及识别新 AI 模型的需求或现有模型的改进。还涉及定义项目目标、范围、利益相关者和成功标准。
2. **数据收集与准备**：从各种来源收集和预处理数据，用于训练和评估 AI 模型。包括数据清洗、特征工程以及将数据集划分为训练集、验证集和测试集。
3. **模型开发**：此阶段涉及使用机器学习算法和技术构建和测试 AI 模型。还包括选择合适的模型架构、超参数和优化策略。

4. **评估与验证**：通过评估指标和验证技术评估 AI 模型的性能。包括交叉验证、留出验证以及与基准的对比。

5. **部署与集成**：将 AI 模型部署到生产环境中，并将其集成到现有系统和工作流中。涉及容器化、API 开发和部署自动化。

6. **监控与优化**：监控 AI 模型在生产中的性能，并进行优化，以保持或提高其性能。包括识别和解决漂移、偏差和性能下降问题。

7. **退役或替换**：退役不再有效的 AI 模型，或用新版本或替代解决方案进行替换。包括存档模型工件、更新文档并与利益相关者沟通变更。

模型治理涉及的各项考量因素包括：

1. **政策制定**：模型治理从全面的政策和指南开始，定义 AI 模型开发和部署的标准和最佳实践。政策涵盖数据隐私、安全、公平性、透明度和伦理等方面。它们为数据使用、模型训练方法、评估标准和部署协议提供明确的指示。

2. **风险管理**：模型治理的关键方面是风险管理，涉及识别、评估和缓解 AI 模型相关的风险。这包括评估训练数据中的潜在偏差，评估模型错误的影响，以及识别安全漏洞。风险管理策略旨在减少不利结果的可能性，并确保 AI 模型与组织目标和法规要求一致。

3. **合规性**：遵守相关法规、标准和行业指南对模型治理至关重要。组织必须确保 AI 模型遵守 GDPR、HIPAA、CCPA 以及特定行业的法律要求。这包括通过彻底的评估来验证合规性，实施必要的保障措施，并维护文档以证明遵守规则。

4. **文档记录**：有效的模型治理需要在 AI 模型生命周期中进行全面的文档记录。这包括记录数据源、预处理步骤、模型架构、超参数、训练方法、评估指标和部署配置。详细的文档记录有助于透明性、可重复性和审计性，促进数据科学家、工程师和合规人员之间的协作。

5. **版本控制**：版本控制对于管理 AI 模型及其相关工件的变更至关重要。版本控制系统（如 Git）可跟踪代码更改、模型迭代和实验结果。这使团队能够重复实验、比较模型性能，并在必要时恢复到以前的版本。版本控制确保 AI 模型开发中的一致性、可追溯性和责任性。

6. **监控与维护**：持续监控和维护对于确保 AI 模型在生产环境中的持续有效性和可靠性至关重要。监控工具和技术使组织能够跟踪模型性能，检测漂移或退化，并识别异常或错误。自动化的模型重新训练、更新和验证流程有助于保持模型的准确性和时效性。

7. **伦理考量**：模型治理还涵盖与 AI 技术相关的伦理考量。包括解决公平性、问责制、透明度和社会影响等问题。伦理 AI 框架和指南为 AI 开发和部署提供原则和指导。组织必须在模型开发过程中纳入公平性指标、偏差检测技术和可解释性方法，以确保 AI 实践的伦理性。

评估标准

模型治理的评估标准包括：

1. **法规合规性**：确保遵守相关法律、法规和行业标准。
2. **风险管理**：识别、评估和缓解与 AI 模型相关的风险。
3. **透明度**：提供清晰的文档和解释，涵盖模型开发和部署过程。
4. **公平性**：评估并减轻 AI 模型中的偏差，确保结果的公平性。
5. **安全性**：实施措施保护 AI 模型及数据免受未经授权的访问和网络威胁。
6. **伦理考量**：处理 AI 模型使用中的伦理影响及社会影响。
7. **性能**：评估模型在现实场景中的准确性、可靠性和效率。
8. **问责性**：定义模型开发、部署和监督中的角色和责任。

9. **持续改进：**建立监控、更新和优化 AI 模型的机制。
10. **利益相关者参与：**让相关利益相关者参与模型治理过程，确保与组织目标和价值观保持一致。

RACI 模型

任务	执行者	负责最终结果者	咨询对象	被告知者
制定 AI 模型政策	AI 治理委员会	首席 AI 官	法务团队、合规团队	高管领导
评估模型风险	数据科学家	AI 伦理官	合规团队、安全分析师	管理层、合规官员
确保法规合规	合规团队	首席合规官	法务团队、法规事务部门	高管领导、董事会
记录模型开发	数据科学家	AI 治理委员会	法务团队、数据治理官	IT 团队、合规团队
实施版本控制	数据工程师	AI 治理委员会	IT 安全团队、DevOps 团队	数据科学家、开发团队
监控模型性能	AI 运营团队	AI 治理委员会	IT 安全团队、数据科学家	管理层、合规官员
解决模型偏差	数据科学家	AI 伦理官	多元化与包容团队、法务团队	合规团队、管理层
增强模型安全性	IT 安全团队	首席信息安全官	数据工程师、合规团队	高管领导、董事会
更新模型文	数据科学	AI 治理委员	法务团队、合规团队	开发团队、IT 团

任务	执行者	负责最终结果者	咨询对象	被告知者
档	家	会		队
审查模型合规报告	合规团队	首席合规官	AI 治理委员会、法务团队	管理层、合规官员

高级实施策略

1. **建立治理框架**：制定政策和控制措施，以确保法规合规性、风险管理和伦理使用。
2. **定义角色**：为治理委员会、数据科学家和合规官员分配责任。
3. **整合治理**：将治理实践嵌入到开发过程中。
4. **实施版本控制**：跟踪更改并保持文档记录，以便于审计。
5. **部署监控**：持续监控生产中的模型性能和行为。
6. **进行审计**：定期审查模型和治理实践，确保合规。
7. **提供培训**：向利益相关者教授治理原则和责任。
8. **持续改进**：根据反馈和新兴趋势调整实践。

持续监控与报告

AI 模型治理的持续监控和报告包括：

1. **实时监控**：实施系统，以实时跟踪模型性能、数据质量和安全性。
2. **警报机制**：设置警报，通知利益相关者偏差、异常或安全违规行为。
3. **合规报告**：定期生成报告，证明符合法规要求和组织政策。

4. **性能指标**：监控关键绩效指标 (KPI)，如模型准确性、公平性和可靠性。
5. **异常检测**：采用技术检测模型的漂移、偏差或其他性能问题。
6. **反馈循环**：建立流程，将监控反馈纳入模型优化和治理实践中。

访问控制

1. **基于角色的访问控制**：实施基于角色的访问控制 (RBAC)，根据用户角色和权限限制对 AI 模型、数据和资源的访问。
2. **特权访问管理**：管理特权访问，确保敏感 AI 资源不被未经授权使用或修改。
3. **身份验证机制**：实施多因素身份验证 (MFA) 等身份验证机制，以验证访问 AI 模型的用户身份。
4. **加密**：加密数据和通信，保护敏感信息免遭未经授权的访问或拦截。
5. **审计跟踪**：维护审计跟踪记录，追踪对 AI 模型和数据的访问，实现可追溯性和问责性。
6. **定期审查**：定期审查访问控制，确保符合安全政策和法规要求。

适用框架和法规

符合 ISO/IEC 27001、NIST 指南以及 GDPR 等行业标准，确保 AI 计划符合既定的 GRC（治理、风险管理与合规）框架，维护组织价值和责任。

3. 安全文化与培训

培育以安全为导向的文化并提供全面的培训是 AI 系统负责任开发和使用的基础。本节探讨了在组织内部构建稳健的安全文化和实施有效的培训计划的多层次方法，涉及 AI 技术的部署。内容涵盖了特定角色的教育、提升组织各级员工安全意识的策略、负责任的 AI 实践的专业培训以及建立清晰的沟通和报告渠道。通过关注这

些关键领域，组织可以培养不仅掌握 AI 技术且深刻理解 AI 部署的伦理影响和安全考量的员工队伍。这种全面的方法确保安全和责任感被嵌入 AI 运营的各个方面，营造了一个创新与稳健的风险管理及伦理考量并存的环境。

3.1 基于角色的教育

概述：AI 正在改变各个行业，基于角色的教育已成为所有组织的需求。此教育策略对于确保组织各级别的所有员工（包括非员工）能够充分利用 AI 技术、在各自领域中创新并领导发展至关重要。理解 AI 的能力与局限性对于从高层管理人员到一线员工的所有角色都非常重要。

基于角色的 AI 教育计划可以通过将学习成果与特定岗位的需求对齐，显著提升个人和团队的表现。例如，理解预测分析的营销人员可以更好地根据 AI 预测的客户行为来调整营销活动。

课程设置：课程应具备灵活性，核心模块涉及 AI 基础知识，选修模块则根据具体角色量身定制。例如，AI 伦理学模块可能是所有人必修的，而 AI 在供应链管理中的应用模块则可能是为相关角色提供的选修课程。

交付方式：此基于角色的教育可以通过在线或线下研讨会和讲座进行。

评估标准：必须定义清晰的指标和标准，以评估每个角色类别的表现。例如，高管可能会根据他们作出基于 AI 的战略决策的能力进行评估，而数据科学家则可能根据其在训练 AI 模型方面的熟练程度进行评估。

责任矩阵（RACI 模型）

- **执行**：人力资源部和学习与发展团队，负责设计和提供培训。
- **负责**：首席信息官或首席 AI 官，负责整体 AI 教育战略。
- **咨询**：部门负责人，负责根据角色的具体需求定制内容。

- **知情**：所有员工，关于可用的 AI 培训及其重要性。

高级实施策略

实施基于角色的教育的策略包括将学习计划与具体岗位角色及更广泛的业务目标相对齐。为实现这一目标，组织需要培养 AI 学习文化并积极发展内部 AI 人才。最终目标是将获得的知识转化为切实的创新。关键领域包括：

- **进行全组织的 AI 素养评估**：这将为各部门的当前 AI 知识水平建立基线。
- **设计核心和特定角色的 AI 培训模块**：开发综合培训模块，涵盖基础 AI 概念及公司各角色的具体需求。
- **与关键部门启动试点计划**：在选定的部门进行初步推广，允许我们收集反馈并改进课程。
- **收集反馈并完善课程**：通过反馈循环实现持续改进，确保课程保持相关性和有效性。
- **全组织推广，季度更新**：在成功的试点之后，整个组织将实施该计划，并定期更新以跟上不断发展的 AI 形势。在组织内部培养 AI 学习文化，例如通过 AI 主题活动或黑客马拉松。

为确保该计划成功，还将考虑以下关键因素：

- **指标**：考虑包括一些具体指标来衡量战略的成功，如员工参与度的增加、工作表现的改善或创新方面的可量化影响。
- **时间线**：为每个实施阶段提供粗略时间表以提高清晰度。
- **沟通**：强调在整个过程中保持清晰和一致的沟通，确保各利益相关者的参与和支持。

衡量效果：可用于衡量效果的方法包括培训后的评估、员工反馈、生产力和创新变化的监控。

持续监控与报告：建立机制进行持续监督，如每季度技能审核和年度 AI 准备报告。设置警报系统，以标记 AI 采用滞后的部门。

访问控制映射：根据不同用户群体的具体需求对访问权限进行调整。例如，数据分析师需要访问大型数据集进行 AI 培训，而人力资源可能需要 AI 驱动的招聘工具。

适用的框架和法规

遵循 IEEE 的《符合伦理的 AI 系统设计》等行业标准。对于负责任的 AI 使用，请参考 NIST 的 AI 风险管理框架。

那些在全面、基于角色的 AI 教育上进行投入的组织，在利用 AI 实现战略优势和创新方面将处于更有利的地位。

3.2 意识建立

意识建立旨在赋予组织内部的个人做出明智决策的能力，并采取主动行动来保护敏感信息，防范社会工程攻击，并坚持良好治理和风险管理的原则。

通过培养一种意识文化，组织可以减少由人为错误、疏忽或恶意意图导致的安全事件的可能性，从而最大程度地降低相关的财务、声誉和法律后果。

意识建设计划的目标是为员工提供有效减轻风险并维护组织价值观所需的知识、技能和态度。意识建设的关键目标包括：

1. **创建清晰简明的文档、政策和程序：**这些帮助确定组织的基调，并与所有利益相关者沟通愿景、使命、目标和优先事项。政策文件应列出以下内容：
    - 组织内的关键联系人、角色和职责；

- 与人工智能相关的法律和监管要求；
- 定期审查人工智能系统的频率；
- 人工智能系统的接入、退役或逐步淘汰程序；
- 处理、响应和从事件及错误中恢复的流程。

- **意识建立策略与活动**：为所有员工及第三方合作伙伴（包括承包商和供应商）定期提供培训和意识计划，以便跟上组织不断变化的风险和期望。应实施责任结构和沟通渠道，以确保员工在现有政策、程序和协议范围内，能够最大限度地履行其职责。与人力资源和其他部门合作，将意识建立整合到入职流程、绩效评估和员工发展的持续计划中。
- **与组织文化的整合**：部署一致、可重复的流程，并进行定期培训以促进批判性思维。此外，组织团队必须传达风险及其更广泛影响。通过信息共享促进员工之间的协作，从而帮助建立透明和协作的文化。需要领导支持和积极参与意识计划，以展示安全性和风险管理承诺。培养公开性、透明性和持续改进的文化，使员工能够放心地报告安全问题并在需要时寻求帮助。
- **持续改进与适应**：记录员工对人工智能技术的每项正面和负面反馈，因为这有助于分析并识别任何特定情境中的潜在风险，同时评估人工智能系统的可信性。稍后，这些见解可以纳入系统设计，以增强人工智能决策过程。定期评估意识建设工作的影响，并调整策略以应对新兴风险和挑战。保持对行业趋势、最佳实践和新兴技术的了解，以增强意识建设工作随时间推移的有效性。

评估标准

- 涵盖人工智能治理和风险管理所有方面的意识计划的全面性
- 提高员工对人工智能相关风险和责任的理解的有效性

- 意识培训课程的频率和定期性
- 培训内容与组织内不同角色的相关性
- 安全实践和事件报告方面的可衡量改进
- 意识建立与组织文化的整合
- 意识计划适应新兴人工智能趋势和风险的能力
- 减少与人为错误相关的安全事件的影响

责任矩阵（RACI 模型）

- **执行**：IT 安全团队，人力资源
- **负责**：首席信息安全官（CISO）、首席 AI 官
- **咨询**：AI 开发团队，数据科学团队，法律和合规部门
- **知情**：管理层，业务单元领导，所有员工

高级实施策略

1. 制定全面的 AI 意识培训材料。
2. 为所有员工建立定期的培训计划。
3. 创建特定角色的意识计划（如开发人员、管理人员和最终用户）。
4. 实施机制以跟踪和衡量意识计划的有效性。
5. 将 AI 意识整合到新员工的入职流程中。
6. 制定沟通策略，定期强化意识信息。
7. 建立反馈循环，持续改进意识计划。

8. 与人力资源部门合作，将意识指标纳入绩效评估。

持续监控和报告

1. 实施培训前后评估，以衡量知识改进情况。
2. 跟踪各部门参与意识计划的参与率。
3. 监控报告的与 AI 相关的事件或问题的频率和性质。
4. 定期进行调查，评估员工对 AI 治理和风险的态度。
5. 建立 KPI 以衡量意识计划的有效性（如安全事件的减少）。
6. 每季度生成意识计划绩效和影响的报告。
7. 实施系统，让员工提供对意识计划的反馈。

访问控制映射

- **IT 安全团队和人力资源**：完全访问意识计划材料和指标
- **CISO 和首席 AI 官**：无限制访问所有与意识相关的数据和报告
- **AI 开发团队**：访问与其工作相关的技术意识材料
- **数据科学团队**：访问与数据相关的意识内容和最佳实践
- **法律和合规部门**：访问与合规相关的意识材料
- **管理层和业务单元领导**：访问高级别的意识计划报告
- **所有员工**：访问一般的 AI 意识培训材料和资源

适用框架和法规

组织可能不会为 AI 治理和风险管理中的意识建设制定特定的适用框架和法规，而是根据其独特需求量身定制他们的计划。然而，结合适用框架和法规可以为意识计划提供稳固的基础，并展示与行业最佳实践和标准的一致性。相关的适用框架和法规示例如下：

- 行业内认可的框架：NIST 网络安全框架、ISO 27001
- 法规：GDPR、CCPA
- 标准：IEEE 7010-2019 AI 治理标准

通过利用这些适用的框架和法规，组织可以确保其意识计划建立在坚实的基础上，并与行业基准保持一致。

3.3 负责任的 AI 培训

在组织环境中，负责任的 AI 是指以良好的意图设计、开发和部署 AI，赋能员工和企业，同时公平地影响客户和社会。它使公司能够在获得信任的同时，自信地扩展 AI。负责任的 AI 是一个新兴的 AI 治理领域，涵盖道德、法律和价值观在开发和部署有益 AI 时的应用。

总体而言，AI 系统面临的最大挑战来自训练数据的偏见。每个人在某些方向上都存在偏见，因此数据也会受到偏见的影响。在风险缓解的背景下，负责任的 AI 有助于减少与 AI 系统相关的风险，如偏见、数据所有权、隐私、准确性和网络安全。这有助于建立消费者信任、促进采用，并减少财务和法律风险。

负责任的 AI 实践包括：

1. **检查原始数据**：机器学习模型会反映其训练的数据，因此仔细分析原始数据以确保理解。
2. **减轻偏见**：应努力识别和减轻 AI 系统中的偏见。

3. **促进透明性和可解释性**：AI 系统应保持透明，其决策应是可解释的。
4. **纳入隐私考虑**：在设计和实施 AI 系统时，隐私应是一个关键考虑因素，使用差分隐私或安全多方计算等隐私保护技术。
5. **识别多种指标**：使用多个指标而非单一指标来理解不同类型的错误和经验之间的权衡。
6. **建模潜在的负面反馈**：在设计过程中提前建模潜在的负面反馈，并通过 A/B 测试或金丝雀发布等逐步进行测试和迭代，先在少量流量上部署。
7. **平衡 AI 能力与人类判断**：虽然 AI 可以提供有价值的见解和自动化，但在人类判断仍然至关重要的许多上下文中，平衡 AI 与人类判断的作用非常重要。
8. **适当的披露**：设计功能时应纳入适当的披露机制。清晰性和控制对于提供良好的用户体验至关重要。
9. **优先考虑教育**：对 AI 及其影响的教育应是所有利益相关者的优先事项。
10. **构建多样化和多学科团队**：多样化团队可以带来不同的视角和经验，有助于识别和减轻潜在偏见。
11. **与多样化用户群体互动**：与各种用户和场景进行互动，并在项目开发的整个过程中纳入反馈。
12. **以人为中心的设计方法**：系统实际用户的体验方式对于评估其预测、推荐和决策的真正影响至关重要。
13. **考虑增强和辅助**：有时，系统为用户提供几个选项可能是最优的。

有关六个值得关注的领域，请参见下文

评估标准

- 涵盖负责任 AI 所有方面的培训的全面性

- 提高员工对 AI 伦理和负责任实践理解的有效性
- 将负责任 AI 原则融入 AI 开发和部署流程的情况
- 在与 AI 相关的道德事件或偏见上的可衡量改进
- 负责任 AI 培训课程的频率和规律性
- 培训内容与组织内不同角色的相关性
- 影响培养道德 AI 开发和使用文化的效果
- 培训计划适应新兴 AI 伦理趋势和挑战的能力

责任矩阵 (RACI 模型)

- **执行**：AI 开发团队，数据科学团队，IT 安全团队
- **负责**：首席 AI 官，首席伦理官（如适用）
- **咨询**：法律与合规部门，人力资源
- **知情**：管理层，业务单元领导，所有与 AI 一起工作的员工

高级实施策略

1. 制定全面的负责任 AI 培训材料。
2. 为所有与 AI 相关的角色建立定期的培训计划。
3. 创建特定角色的培训计划（如开发人员、数据科学家和管理人员）。
4. 实施机制以跟踪和衡量培训的有效性。
5. 将负责任 AI 原则整合到 AI 开发工作流中。
6. 制定沟通策略，定期强化负责任 AI 实践。

7. 建立反馈循环，持续改进培训计划。

8. 与人力资源部门合作，将负责任 AI 指标纳入绩效评估。

持续监控与报告

1. 实施培训前后评估，以衡量知识改进情况。

2. 跟踪负责任 AI 实践在 AI 项目中的实施情况。

3. 监控报告的与 AI 伦理相关问题的频率和性质。

4. 定期审计 AI 系统，确保其符合负责任 AI 原则。

5. 为负责任 AI 实施设立 KPI（例如，减少偏见结果）。

6. 每季度生成负责任 AI 性能和影响的报告。

7. 实施系统，允许员工报告 AI 系统中的潜在伦理问题。

访问控制映射

- **AI 开发团队和数据科学团队**：完全访问负责任 AI 培训材料和工具
- **首席 AI 官和首席伦理官**：无限制访问所有与负责任 AI 相关的数据和报告
- **IT 安全团队**：访问负责任 AI 培训的安全相关方面
- **法律与合规部门**：访问与合规相关的负责任 AI 材料
- **人力资源部**：访问负责任 AI 培训记录和绩效指标
- **管理层和业务单元领导**：访问高级别的负责任 AI 实施报告
- **所有与 AI 工作的员工**：访问一般的负责任 AI 培训材料和资源

适用框架和法规

- 通用数据保护条例 (GDPR) - 欧盟
- 加利福尼亚州消费者隐私法 (CCPA) - 美国
- AI 法案（拟议） - 欧盟
- 算法问责法（拟议） - 美国
- IEEE 伦理对齐设计
- ISO/IEC JTC 1/SC 42 人工智能标准
- NIST AI 风险管理框架
- 蒙特利尔负责任 AI 开发宣言

3.4 沟通与报告

在 AI 背景下，沟通与报告指的是向内部和外部利益相关者系统性地传递关于组织的 AI 项目、其影响、风险及合规状态的信息的过程。该职责涵盖 AI 使用的透明披露，包括数据来源、算法和潜在偏见的详细信息；定期更新 AI 系统性能，包括准确性、公平性、可解释性的指标；以及风险评估、伦理考量和合规情况。

有效的沟通与报告通过与利益相关者建立信任，展示责任心，并在组织内部促进负责任 AI 使用的文化。它涉及创建明确的渠道来共享信息、建立报告机制，确保所有相关方了解公司在 AI 实践、挑战和成就方面的最新信息。

关键方面包括：

- 定期内部报告 AI 项目及其与企业价值和战略的对齐情况
- 向外部沟通 AI 使用、效益及潜在风险，包括客户、投资者及公众
- 透明披露与 AI 相关的事件或问题

- 定期更新 AI 治理措施和合规状态
- 清晰传达公司 AI 伦理原则及其实施方式

评估标准

- AI 相关报告的频率与质量（内部与外部）
- 利益相关者对 AI 相关沟通的满意度（通过调查衡量）
- AI 使用及其影响的披露数量
- 解决重大 AI 相关事件或变化所需的时间
- 有完整沟通计划的 AI 项目百分比
- AI 决策过程透明度（通过独立审计报告）
- AI 伦理声明和公开 AI 政策的更新频率

责任矩阵 (RACI 模型)

角色	职责
执行	通讯团队，AI 开发团队
负责	首席技术官
咨询	法律与合规部门，数据保护官，业务单元领导
知情	管理层，运营人员，外部利益相关者

高级实施策略

1. 制定符合公司价值观的全面 AI 沟通策略。
2. 建立 AI 项目和计划的内部报告机制。

3. 创建一致的 AI 相关沟通模板和指南。
4. 实施定期与利益相关者就 AI 主题进行互动的系统。
5. 制定 AI 相关事件的危机沟通计划。
6. 建立定期审查和更新公共 AI 政策及伦理声明的流程。
7. 为内部和外部使用创建 AI 透明度仪表板。
8. 培训关键人员，使其掌握有效的 AI 相关沟通技巧。

持续监控与报告

1. 跟踪 AI 相关沟通的频率和覆盖范围。
2. 监控利益相关者对 AI 沟通的反馈和情绪。
3. 定期评估内部 AI 报告机制的有效性。
4. 跟踪解决 AI 相关询问或问题所需的时间。
5. 监控媒体报道和公众对公司 AI 项目的看法。
6. 定期生成有关 AI 沟通透明度和清晰度的报告。
7. 定期审计 AI 项目文档和沟通记录。

访问控制映射

1. 限制正式 AI 沟通内容的编辑权限，仅限授权人员操作。
2. 实施基于角色的访问控制，用于 AI 报告系统和仪表板。
3. 确保不同利益相关者拥有适当的 AI 相关信息访问级别。
4. 控制 AI 项目敏感细节在公共沟通中的访问权限。

5. 实施外部 AI 相关沟通的审批流程。

适用框架与法规

- **证券交易委员会 (SEC) 披露要求**：要求对重要信息进行透明披露，其中可能包括重大 AI 项目或风险。
- **欧盟人工智能法案（拟议）**：一旦实施，该法案将要求对高风险 AI 系统进行透明度和报告。
- **OECD AI 原则**：强调透明性和负责任的 AI 系统披露。
- **ISO/IEC 38507:2022**：为 AI 治理提供指南，包括沟通和报告方面。
- **全球报告倡议 (GRI) 标准**：尽管不是专门针对 AI，这些标准为可持续性报告提供框架，可以适用于 AI 相关披露。

4. 影子 AI 防范

应对影子 AI 的挑战——即组织内未经授权或未记录的 AI 系统——对于维护 AI 操作中的控制、安全和合规性至关重要。本节深入探讨识别、管理和防范影子 AI 的策略和方法。内容包括创建全面的 AI 系统清单、进行彻底的差距分析以识别授权与实际 AI 使用之间的差异，并实施强有力的机制来识别未经授权的系统。此外，还将探讨建立严格的访问控制、部署先进的活动监控技术以及实施严谨的变更控制流程。通过关注这些关键领域，组织可以显著降低与影子 AI 相关的风险，确保所有 AI 系统符合组织政策、安全标准和监管要求。这种主动的方法提升了整体 AI 治理，并在 AI 部署和使用中培养了透明和问责的文化。

4.1 AI 系统清单

AI 资产管理系统是一种专门的框架或工具，旨在管理和分类组织内与人工智能相关的资产。该资产管理系统帮助组织跟踪他们已部署或正在开发的各种 AI 系统，并记录每个系统的相关细节。该系统超越了传统的资产管理，特别关注构成 AI 系统的各个组件，包括但不限于：

- **描述：**对 AI 系统的简要描述，包括其目的、功能和预期用途。每个 AI 系统在清单中应有唯一的标识。
- **AI 模型：**机器学习模型的详细记录，包括其版本、算法、训练数据集、参数、性能指标和部署状态。
- **数据集和数据源：**有关用于训练和测试 AI 模型的数据集的信息，包括其来源、大小、质量指标及任何预处理步骤。这包括训练数据和用于推断的实时数据流。
- **计算资源和环境：**AI 模型和算法开发、训练、部署以及计划部署的硬件和软件环境的详细信息，例如云资源、本地服务器、边缘设备以及专用硬件（如 GPU）。
- **开发和部署工具：**记录用于 AI 模型开发、部署、版本控制和监控的工具和平台。
- **文档和合规性：**涵盖 AI 资产生命周期的综合文档，包括伦理考虑、合规性要求以及对标准（如 NIST AI 风险管理框架（RMF）和 NIST 安全软件开发框架（SSDF））的遵循。
- **访问控制和安全：**保护 AI 资产（特别是敏感数据和专有模型）的访问控制、安全措施和协议的记录。

AI 资产管理系统的目的

维护 AI 系统清单对于确保组织内 AI 的透明度、问责制和有效治理至关重要。它使利益相关者能够了解 AI 部署的现状，评估风险，监控性能，并就与 AI 相关的举措做出明智的决策。

- **可见性：**提供组织内所有与 AI 相关资产的清晰概览，便于管理和决策。
- **合规与治理：**维护 AI 系统及其组件的详细记录，以帮助确保 AI 部署符合法律、伦理和监管标准。
- **风险管理：**识别和减轻与 AI 系统相关的风险，包括数据隐私问题、模型偏差和安全漏洞。
- **资源优化：**实现计算资源和数据集的高效分配和利用。用于评估 AI 系统性能的指标包括准确率、精确率、召回率、延迟、吞吐量等。
- **生命周期管理：**支持 AI 模型从开发、训练到部署和维护的整个生命周期，确保所有变更都有记录和文档支持。参考相关文档，包括用户手册、技术规格和培训材料。

AI 资产管理系统的特点

- **自动发现与目录管理：**能够自动发现和 catalog 各种环境和平台中的 AI 资产。实现 AI 资产的自动发现和目录管理，并与现有的资产管理系统集成。
- **版本控制：**跟踪 AI 模型和数据集的不同版本，以确保可重现性，并在需要时便于回滚。
- **集成能力：**与现有的资产管理、开发、部署和监控工具无缝集成，以提供 AI 资产的统一视图。
- **安全与访问管理：**实施强有力的安全措施和访问控制，以保护敏感信息和知识产权。

AI 资产管理系统集中管理 AI 资产，使组织能够最大化其 AI 项目的价值，同时确保合规性、治理和高效的资源利用。

将 AI 资产管理系统与现有的资产和模型清单集成需要一种战略性的方法，以便与组织流程对齐，利用技术，并确保遵守治理、风险和合规（GRC）标准。

以下是组织如何有效地整合 AI 资产管理系统的方法。

1. 与现有资产管理系统的集成

映射 AI 组件：在现有资产管理框架内识别和映射所有 AI 组件，包括模型、数据集和相关应用程序。这确保了 AI 资产作为更广泛组织资产生态系统的一部分，获得全面的视图。

技术利用：利用现有的资产管理软件或平台，整合 AI 特定属性和元数据。这可以包括模型版本控制、数据来源、部署环境和性能指标。

过程对齐：将 AI 资产清单流程与现有资产生命周期管理协议对齐。这包括采购（或开发）、部署、维护和退役阶段，确保 AI 资产在整个生命周期中得到高效管理。

2. 确保合规性和安全性

合规标准：遵循相关标准和框架，如 NIST AI RMF 和 NIST SSDF，将这些标准纳入资产管理过程。这涉及记录与伦理指南、安全措施和风险管理实践的合规性。

访问控制：为 AI 资产清单系统实施强有力的访问控制措施，以确保有关 AI 资产的敏感信息（如专有模型或数据集）受到保护。这需要在资产管理系统中定义角色和权限，以按需限制访问。

安全措施：在存储和传输 AI 资产数据时，纳入安全措施，包括加密和安全访问协议。定期审计和监控访问日志，以检测和响应未经授权的访问尝试。

3. 持续监控和报告

自动监控：部署自动化工具，持续监控 AI 资产，并跟踪模型版本、数据使用情况和系统配置的变化。这有助于维护一个反映 AI 资产当前状态的最新清单。

报告机制：在资产清单系统内开发报告机制，提供对 AI 资产状况的洞察，包括使用统计、性能指标和合规状态。这将促进知情决策，并支持 GRC 报告要求。

4. AI 资产管理的 RACI 模型

使用 RACI 模型定义明确的角色和责任，以确保对 AI 资产清单系统的有效治理。

执行：IT 和 AI 开发团队更新资产清单中的 AI 资产详细信息。

负责：首席数据官（CDO）和首席信息安全官（CISO）对 AI 资产清单系统的整体管理和安全性负责。

咨询：业务部门和合规团队被咨询，以确保 AI 资产清单符合操作需求和监管要求。

知情： 定期报告使所有利益相关者，包括管理层和战略角色，了解 AI 资产的状态和健康状况。

5. 培训与意识

员工培训： 对参与 AI 开发、部署和管理的员工进行培训，以确保他们理解资产清单系统、流程及其责任。

意识活动： 开展意识活动，强调准确的 AI 资产文档和遵守安全与治理协议的重要性。将 AI 资产清单系统与现有的资产和模型清单整合，有助于提高运营效率和风险管理，支持战略决策，并确保遵循监管标准。组织可以通过利用技术、对齐现有流程、确保透明治理，来维持一个强大且反应灵敏的 AI 资产清单系统。

6. 生命周期问责

评估跨实体影响： 大多数 AI 系统超越单一上下文使用。这尤其适用于那些功能强大的通用系统，因为它们在各个领域都有应用。生命周期分析确保在 AI 开发和实施的价值链中考虑各参与方的角色。这也将有助于确保对 AI 的适当法律责任得到公平和有效的分配。通用 AI 系统的资产清单和风险评估方法通常不如专业 AI 系统成熟，需要更多的努力、时间、资源和专业知识。

7. 适用框架和法规

以下指南为 AI 资产清单系统提供基础：

IEEE 7010-2019： 提供 AI 治理的指导方针，包括资产清单管理和透明度方面。

NIST AI RMF： 该框架提供了一种结构化的方法来管理与 AI 相关的风险，包括资产清单管理和风险评估。

NIST SSDF： 提供安全软件开发实践的指导方针，包括资产清单和漏洞管理。

ISO/IEC 38507:2022： 提供 AI 治理的指导方针，包括资产清单管理、风险管理和合规性。

OCDE AI 原则： 强调透明度、问责制和非歧视。

欧盟人工智能法案：建立了关于人工智能使用的全面法规，重点关注风险管理、透明度和问责制。法案还对高风险 AI 系统规定了具体要求，包括清单管理和合规性。

这些框架为 AI 资产清单系统提供了基础，确保其与行业最佳实践和人工智能治理与管理标准保持一致。

4.2 差距分析

差距分析是一种战略管理技术，旨在带来所需的变更和改进。当应用于防止影子 AI 时，它涉及评估组织内 AI 使用的当前状态，并创建与安全和治理 AI 实施相一致的明确框架的路线图。

准备和背景理解

当前状态评估： 使用 AI RMF 的“MAP”功能分析当前的 AI 系统及其在组织中的使用情况。这包括创建 AI 系统的清单、分析使用模式，并审查现有的治理框架。

定义期望的目标状态： 期望的目标状态专注于建立全面的 AI 治理政策，符合 AI RMF 的“GOVERN”功能。这包括制定明确的 AI 使用指导方针，确保负责任的数据管理，并遵循法律和伦理标准。治理应与行业最佳实践相一致，以确保 AI 系统在定义的边界内运行，并遵守内部和外部法规。

健全的 AI 治理政策和指导方针： 制定全面的 AI 治理政策和指导方针，遵循 AI RMF 的“GOVERN”功能。这应关注明确的使用指示、数据管理协议以及遵循法律和伦理标准。

持续的 AI 监督与控制： 健全的监控和控制流程对 AI 系统的持续监督至关重要。期望的状态涉及实施先进的监控解决方案，如实时分析和异常检测软件，以提供对组织 AI 使用情况的实时可见性。这包括跟踪 AI 的利用情况、数据输入和输出，以及用户交互，以检测未经授权或不当使用。

自动化警报和异常检测机制能够及时识别潜在风险并触发适当响应。定期审核和日志审查进一步确保持续遵守数据保护标准和模型完整性。这一综合的监控和控制框架使组织能够主动管理与 AI 相关的风险，并维护 AI 系统和数据的完整性与安全性。

意识与培训：实施全面的培训项目，与 AI RMF 的重点相一致，强调教育员工有关 AI 风险、数据隐私和伦理考量。

技术控制：健全的技术控制框架对于防止未经授权的 AI 使用和确保数据安全至关重要。这包括实施访问控制措施，以确保仅授权用户可以访问 AI 系统和数据。加密和密钥管理解决方案保护敏感数据在存储和传输过程中的安全，防止未经授权的访问或盗窃。此外，组织利用微分段等先进技术来隔离 AI 系统和数据，增强安全性并促进精细控制。定期的安全评估和渗透测试可以识别漏洞，并确保 AI 基础设施的韧性。这些技术控制为安全的 AI 采用提供了坚实的基础，帮助减轻风险，保护组织的资产和声誉。

差距识别与分析

比较与对比： 通过将当前 AI 治理和安全实践与 AI RMF 定义的期望状态进行比较，识别差距。这一步骤涉及评估 AI 系统的清单、使用监控、政策执行及现有控制措施的有效性。

识别差距： 突出 AI 系统清单管理、使用监控、政策执行和现有控制措施有效性中的差异。特别注意与 AI RMF 中列出的可信特性相关的领域，如安全性、问责制和公平性。

修复策略开发

行动计划： 针对每个识别出的差距，制定包括技术、政策和培训方面的行动计划，以解决缺陷。这些计划应参考 AI RMF 的综合方法来管理与 AI 相关的风险。

利益相关者参与： 吸引利益相关者参与修订修复计划，确保方法与 AI RMF 中涉及相关 AI 参与者和促进负责任 AI 使用的原则保持一致。

实施与持续改进

执行修复计划： 实施旨在弥补识别差距的策略，遵循 AI RMF 中提出的原则，以实现持续的 AI 监督和治理。

评估与调整： 定义衡量已实施措施有效性的指标，包括减少未经授权的 AI 使用、改善数据安全和提高员工合规性。定期评估实施措施的有效性，参考 AI RMF 的持续改进指南，以适应不断发展的 AI 技术和组织目标。

通过系统地遵循这些步骤，并参考 NIST AI RMF，组织可以进行全面的阴影 AI 预防差距分析。这种方法确保与安全和治理 AI 实施的战略对齐，减轻风险并促进负责任的 AI 治理和使用文化。

RACI 模型

- **执行：** IT 安全团队、数据治理官、首席信息安全官、IT 团队
- **负责：** 首席 AI 官
- **咨询：** 法律团队、业务单位领导、AI 开发团队
- **知情：** 管理层（包括 CEO、CTO、CFO 等）

高层战略

- 采取分阶段实施的方法，优先处理关键高风险领域，逐步扩展至全面的 AI 治理。
- 持续改进。

监控与报告

- 建立持续的流程以确保合规性，并适应新兴 AI 开发和威胁。

访问控制

- 实施健全的访问控制措施，以限制 AI 系统的使用和数据访问，仅限授权人员。

适用框架和法规

- 遵循 NIST AI RMF 和 NIST SSDF。

4.3 未经授权的系统识别

定期审核 AI 系统清单，使用资产管理软件或配置管理系统，识别未经授权或未记录的系统。实施网络扫描工具以检测与组织网络连接的未经授权的 AI 系统或设备。建立快速处理未经授权系统发现的协议，包括调查、减轻和执行相关政策和程序。通过实施额外措施和考虑，组织可以增强识别、预防和有效应对未经授权 AI 系统的能力，减轻相关安全风险，确保其数据和资源的完整性和机密性。

1. **持续监控：** 实施持续监控机制，以实时检测未经授权的 AI 系统。这可以包括使用入侵检测系统（IDS）或安全信息和事件管理（SIEM）解决方案，提供对异常或未经授权活动的警报。

2. **用户行为分析/用户行为实体分析（UBA/UEBA）：** 利用用户行为分析技术检测与未经授权系统访问或使用相关的异常行为。

   组织可以通过分析用户活动和模式来识别与未经授权的 AI 系统相关的潜在安全漏洞或政策违规。

3. **终端安全：** 加强终端安全措施，以防止未经授权的 AI 系统访问敏感数据或资源。这可能涉及部署终端保护平台（EPP）或终端检测和响应（EDR）解决方案，以监控和控制设备级别的系统访问。

4. **数据泄漏防护（DLP）：** 在云服务和终端上使用 DLP 解决方案，防止员工和系统将数据发送到未经授权的 AI 系统。

5. **网络分段与隔离：** 通过将网络基础设施分段，将授权的 AI 系统与未经授权的系统隔离，降低未经授权访问或数据外泄的风险。网络分段可以通过虚拟局域网（VLAN）或限制网络段之间通信的防火墙策略实现。

6. **定期漏洞评估：** 定期进行漏洞评估和渗透测试，以识别可能被未经授权的 AI 系统利用的潜在安全弱点。组织可以主动解决漏洞，降低未经授权访问或破坏的可能性。网络分段的具体技术可以是软件定义网络（SDN）或网络功能虚拟化（NFV）。

7. **员工培训与意识：** 提供全面的培训和意识项目，以教育员工有关未经授权的 AI 系统相关风险及遵循组织政策和程序的重要性。员工应报告可疑的活动或设备给相关部门。

8. **事件响应规划：** 制定并定期更新事件响应计划，概述应对未经授权系统发现的程序。这应包括对相关利益相关者的遏制、调查、修复和沟通协议。

9. **定期政策审查：** 定期审查与 AI 系统部署和使用相关的组织政策和程序，确保其保持最新，并有效应对与未经授权系统相关的风险。任何差距或缺陷应通过政策更新或修订及时解决。

10. **增强协作与沟通：** 鼓励与不同业务单位的紧密协作和持续沟通，使 IT/网络安全团队了解其需求，并提供与组织目标一致的解决方案，同时满足安全要求，降低使用未经授权系统的风险。

评估标准

- 对未经授权 AI 系统检测机制的有效性
- 识别未经授权或未记录的 AI 系统的速度和准确性
- 网络扫描和监控工具的全面性
- 对 AI 系统清单的审计频率和彻底性
- 对未经授权系统发现的响应时间
- 员工对未经授权系统报告培训的有效性

- 未经授权系统检测与整体安全基础设施的集成
- 检测方法对新兴 AI 技术和威胁的适应性

未经授权系统检测责任矩阵（RACI 模型）

- **执行：** IT 安全团队、网络安全团队
- **负责：** 首席信息安全官（CISO）
- **咨询：** AI 开发团队、系统管理员、法律与合规部门
- **知情：** 首席技术官、首席 AI 官、业务单位领导

高层实施策略

1. 实施持续监控机制以实现实时检测。
2. 部署网络扫描和发现工具以检测未经授权的 AI 系统。
3. 建立处理未经授权系统发现的协议。
4. 实施用户行为分析以检测异常活动。
5. 加强终端安全措施。
6. 部署数据泄漏防护解决方案，如数据丢失防护（DLP）或云访问安全代理（CASB）。
7. 实施网络分段和隔离技术。
8. 定期进行漏洞评估和渗透测试。
9. 制定和维护全面的事件响应计划。

持续监控与报告

1. 实施实时警报以检测未经授权的 AI 系统。
2. 定期审计 AI 系统清单。
3. 监控用户行为，识别与未经授权系统相关的异常活动。
4. 跟踪和报告未经授权系统事件及其解决方案。
5. 实施持续的漏洞扫描和报告。
6. 定期生成关于未经授权系统检测措施有效性的报告。
7. 监控并报告员工对 AI 系统使用政策的遵守情况。

访问控制映射

- **IT 安全团队和网络安全团队：** 完全访问检测工具和系统日志
- **CISO：** 不受限制地访问所有未经授权系统检测数据和报告
- **AI 开发团队：** 访问批准的 AI 系统清单和部署日志
- **系统管理员：** 访问网络和系统配置数据
- **法律与合规部门：** 访问事件报告和政策违规数据
- **首席技术官和首席 AI 官：** 访问高层级的未经授权系统报告
- **业务单位领导：** 访问各自单位的未经授权系统事件摘要

适用的框架和法规

- 联邦信息安全现代化法案（FISMA） - 美国
- 网络和信息系统（NIS）指令 - 欧盟
- 网络安全信息共享法（CISA） - 美国

- 一般数据保护条例（GDPR）- 欧盟（针对数据保护方面）
- ISO/IEC 27001:2013 信息安全管理系统
- NIST 特别出版物 800-53 信息系统和组织的安全与隐私控制
- NIST 网络安全框架
- CIS 关键安全控制

4.4 访问控制

实施强有力的访问控制机制，以根据用户角色、权限和身份验证凭据限制对 AI 系统、模型和数据集的访问。利用多因素身份验证（MFA）、基于角色的访问控制（RBAC）和最小权限原则等技术，确保只有授权用户和系统可以访问 AI 系统和资源。以下是可以应用的一些访问控制措施：

1. **基于角色的访问控制（RBAC）：** 实施 RBAC，以根据用户在组织中的角色和责任分配访问权。根据特定角色（如系统管理员、AI 开发人员、研究人员、数据科学家和模型培训师），授权人员应基于工作要求访问 AI 系统及相关资源。
2. **最小权限原则：** 应用最小权限原则，限制用户执行任务所需的最低访问权限。仅为执行职责的授权用户授予必要的访问权限，从而降低对 AI 系统和敏感数据未经授权访问的风险。
3. **访问控制列表（ACLs）：** 利用 ACLs 为各个用户或用户组定义对 AI 系统和资源的特定访问权限。限制对授权人员和明确拒绝访问未经授权用户或实体以防止未经授权使用。

4. **网络分段：** 对网络基础设施进行分段，以将 AI 系统和其他关键资源与未经授权的访问隔离。使用 VLAN、防火墙或软件定义网络（SDN）等网络分段技术，创建独立的网络段，以限制授权用户与未经授权设备或系统的通信。

5. **双因素身份验证（2FA）：** 实施 2FA 机制，以增强对 AI 系统和敏感数据的访问安全。要求用户使用多个因素进行身份验证，例如一次性密码（OTP）和生物识别验证，以降低未经授权访问的风险。

6. **加密与安全通信协议：** 加密授权用户与 AI 系统之间的数据传输，使用诸如 SSL/TLS 等安全协议。确保用户与 AI 系统之间交换的敏感信息被加密，以防止被未经授权方拦截或窃听。

7. **访问监控与审计：**部署访问监控和审计机制，以跟踪和记录与 AI 系统和资源相关的用户访问尝试。监控访问日志以查找可疑或未经授权的访问尝试，并定期进行审计，以识别潜在的安全漏洞或政策违规行为。

8. **用户培训与意识：**提供全面的培训和意识项目，教育用户关于访问控制政策、程序和最佳实践。确保用户了解他们在访问 AI 系统方面的责任，以及未经授权访问的后果。

通过实施这些访问控制措施，组织可以增强其 AI 系统和基础设施的安全态势，降低未经授权访问的风险，并保护敏感数据免遭未经授权的使用或利用。

评估标准

- 基于角色的访问控制（RBAC）实施的有效性
- 基于属性的访问控制（ABAC）实施的有效性
- 在所有 AI 系统和资源中应用最小权限原则
- 多因素身份验证（MFA）机制的强度

- AI 资源的访问控制列表（ACL）的全面性
- AI 系统隔离的网络分段有效性
- 加密和安全通信协议的稳健性
- 访问监控和审计流程的彻底性
- 员工对访问控制政策的理解和遵守程度

责任矩阵（RACI 模型）

- **执行：** IT 安全团队、网络安全团队
- **负责：** 首席信息安全官（CISO）
- **咨询：** AI 开发团队、系统管理员、人力资源
- **知情：** 首席技术官、首席人工智能官、业务部门领导

高级实施策略

1. 为 AI 系统和资源实施基于角色的访问控制（RBAC）
2. 在所有用户帐户和系统中应用最小权限原则
3. 开发和维护全面的访问控制列表（ACL）
4. 实施网络分段以隔离 AI 系统
5. 部署多因素身份验证（MFA）以确保所有 AI 系统访问
6. 实施加密和安全通信协议
7. 建立访问监控和审计机制
8. 开发并提供关于访问控制政策的用户培训项目

持续监控与报告

1. 实施对 AI 系统访问尝试的实时监控。
2. 定期审计用户的访问权限和特权。
3. 监控和分析访问日志以发现可疑活动。
4. 跟踪并报告多因素身份验证（MFA）的采用和使用情况。
5. 定期生成访问控制政策合规性报告。
6. 监控网络流量以发现潜在的分段违规行为。
7. 跟踪并报告数据在传输和静态状态下的加密使用情况。

访问控制映射

- **IT 安全团队和网络安全团队：** 完全访问安全工具和日志
- **首席信息安全官（CISO）：** 对所有访问控制数据和报告的无限制访问
- **AI 开发团队：** 访问开发环境，适当限制
- **系统管理员：** 对系统配置的提升访问，受限于最小权限原则
- **人力资源：** 访问员工角色信息以管理 RBAC
- **首席技术官和首席 AI 官：** 访问高级别的访问控制报告
- **业务单元领导：** 访问其单元的访问控制措施摘要

适用的框架和法规

- 一般数据保护条例（GDPR）- 欧盟
- 加州消费者隐私法（CCPA）- 美国

- 联邦信息安全现代化法案（FISMA） - 美国
- 健康保险可携带性和责任法案（HIPAA） - 美国
- 支付卡行业数据安全标准（PCI DSS）
- ISO/IEC 27001:2013 信息安全管理系统
- NIST 特别出版物 800-53 安全与隐私控制
- NIST 网络安全框架
- SOC 2 信任服务标准
- ISO/IEC 27559:2022
- ISO 31700-1:2023

4.5 活动监控

影子 AI 是指未经授权在组织内使用 AI 工具和模型，不局限于单一的实施者群体。**对影子 AI 的有效监控需要对可能参与此类活动的组织内部各种角色有细致的理解**。以下是可能涉及的不同角色的简要说明：

1. 拥有直接模型访问权限的内部人员：

- **数据科学家和开发人员**：这些人具备技术能力，可以为未授权目的利用现有模型。他们可能会修改现有模型用于个人项目、绕过数据访问控制，或将模型用于超出其预期范围的任务。
- **模型版本控制和更改跟踪**：实施系统以跟踪模型的更改，识别未经授权的修改。

- **数据访问日志记录与审计**：监控数据访问日志以识别异常模式或绕过访问控制的尝试，使用专门的日志记录和审计工具，如 Splunk 或 ELK。
- **模型使用监控**：使用监控工具（如 TensorBoard 或 MLflow）跟踪模型的使用情况，包括频率、数据输入和输出。这有助于检测异常使用情况。
- **代码审查和安全培训**：将安全最佳实践整合到开发生命周期中，包括漏洞代码审查以及对数据科学家和开发人员的安全培训。
- **IT 专业人员**：具有基础设施和数据管道访问权限的 IT 专业人员可能会部署未经授权的 AI 工具，或通过操纵数据输入来改变批准的模型输出。
- **网络流量监控**：使用网络流量分析工具（如 Wireshark 或 Suricata）监控网络流量，以检测异常数据传输或与未经授权服务器的连接，表明可能部署了未经授权的 AI 工具。
- **终端安全工具**：使用终端安全软件检测员工设备上安装的未经授权的软件，可能揭示出影子 AI 工具的使用。
- **数据血统追踪**：实施数据血统追踪系统，映射整个组织的数据流动。这有助于识别与影子 AI 活动相关的意外数据移动。

2. 组织内的外围用户：

- **业务单元**：市场营销或销售团队可能会使用现成的基于云的 AI 工具（如客户细分或线索生成），而不遵循适当的审批程序。这可能导致数据安全风险或违反合规规定。
- **用户意识与培训**：提供全面的培训计划，介绍批准的 AI 工具和资源，强调安全最佳实践和遵循既定程序的重要性。

其他监控影子 AI 的方法

- **用户活动监控**：实施用户活动监控工具，谨慎使用并遵循隐私法规。这些工具可以跟踪应用程序使用情况，并识别员工使用未经授权的 AI 工具的情况。
- **数据丢失防护 (DLP)**：部署专门设计的 DLP 解决方案，检测和防止 AI 应用程序泄露敏感数据。这可以帮助检测由影子 AI 活动引发的潜在数据泄露。
- **员工反馈和举报程序**：
    - 鼓励员工报告任何可疑的影子 AI 活动，营造透明的文化，使员工愿意提出担忧。
    - 建立保密的举报程序，让员工可以在不担心报复的情况下报告影子 AI 活动。
- **数字版权管理 (DRM)**：考虑对敏感数据集使用 DRM 控制，以限制未经授权的访问和对 AI 模型的使用。

重要说明：

- 在安全性与员工隐私之间取得平衡至关重要。
- 过度侵入性监控会损害员工信任和士气。
- 优先考虑清晰的沟通与员工教育，以配合监控工作。

评估标准

- 检测并缓解未经授权的 AI 工具使用的百分比
- 检测并响应影子 AI 事件的时间
- 完成 AI 安全意识培训的员工数量
- 网络流量异常检测的频率和覆盖范围

- 符合批准的 AI 工具使用政策的比例
- 数据丢失防护措施在阻止未经授权数据传输方面的有效性

责任矩阵 (RACI 模型)

- **执行**：IT 安全团队，网络安全团队：完全访问活动监控工具和日志
- **负责**：首席信息安全官 (CISO)：无限制访问所有活动监控数据和报告
- **咨询**：数据保护官，AI 开发团队，业务单元领导：访问与 AI 模型开发和部署相关的活动监控数据
- **知情**：管理层：访问活动监控的摘要报告和仪表板

高级实施策略

1. 制定全面的影子 AI 监控框架。
2. 实施强大的网络流量监控和异常检测系统。
3. 建立一个集中的 AI 资源平台用于批准的工具。
4. 定期开展 AI 安全意识培训计划。
5. 部署聚焦 AI 相关风险的数据丢失防护 (DLP) 解决方案。
6. 实施具有隐私考虑的用户活动监控工具。
7. 建立明确的沟通渠道和举报程序，用于报告可疑的 AI 活动。

持续监控与报告

1. 设置实时警报，以检测未经授权的 AI 工具安装或异常数据访问模式。
2. 定期审计 AI 模型使用情况及其修改。

访问控制映射

1. 限制模型修改权限，仅限授权的数据科学家和开发人员。
2. 实施基于角色的访问控制，适用于 AI 工具和敏感数据集。
3. 建立审批工作流，用于访问外部 AI 工具。
4. 根据工作角色和职责限制网络访问权限。
5. 实施多因素身份验证，以访问关键的 AI 系统和数据。

适用框架和法规

- **通用数据保护条例 (GDPR)**：确保 AI 系统中的个人数据得到适当处理
- **加利福尼亚州消费者隐私法 (CCPA)**：规范商业使用 AI 应用程序时的数据收集与使用
- **健康保险可携性和责任法案 (HIPAA)**：规范在医疗保健系统中使用 AI 时，受保护健康信息的披露
- **AI 法案 (欧盟)**：旨在根据风险级别监管 AI 系统

4.6 变更控制流程

AI 系统的变更应持续记录、测试、批准和归档。通过实施健全的变更控制流程，组织可以有效地管理其 AI 系统的演变，同时确保遵守监管要求，维护数据完整性，并减轻与变更相关的风险。应建立正式的文档化且已批准的变更管理政策和流程，以管理和监督 AI 系统、模型、算法或数据集生命周期中的任何变更。这些流程对于确保 AI 解决方案的可靠性、性能和完整性，同时最大限度地降低意外后果或中断的风险至关重要。以下是 AI 变更控制流程的结构：

1. 文档记录与跟踪：

- 维护所有 AI 组件的全面文档，包括模型、算法、训练数据和相关元数据。
- 使用版本控制系统或专用变更管理工具跟踪对 AI 工件的变更，记录变更请求人、变更详细信息、批准人、实施人、实施时间以及变更的性质。

2. 变更请求提交：

- 建立正式流程，使用流行的变更管理工具（如 JIRA 或 ServiceNow）提交变更请求，利益相关者可以在此提出对 AI 系统的修改或更新。
- 确保记录提议变更的详细信息，包括变更请求的理由、对性能或功能的潜在影响、实施计划、测试计划、回滚计划及任何相关风险或依赖关系。

3. 审查与审批工作流程：

- 实施结构化的审查和审批工作流程，以评估变更请求并评估其对 AI 系统的潜在影响。可使用工作流管理工具（如 Asana 或 Trello）来帮助实现这一点。
- 审查过程中应涉及相关利益相关者，包括数据科学家、领域专家、业务用户、IT、安全和隐私团队。此协作方法将确保审查提议的变更与业务目标和技术要求保持一致，提供有关流程有效性的保证。

4. 测试与验证：

- 进行严格的测试和验证程序，以评估拟议变更对 AI 系统性能、准确性和可靠性的影响。
- 使用 A/B 测试、交叉验证和压力测试，评估在不同条件和场景下的变更影响，并使用测试框架（如 Pytest 或 Unittest）进行测试。

5. 风险评估与缓解：

- 进行风险评估，识别拟议变更的潜在风险，如模型退化、数据漂移或合规性问题，使用 NIST 或 ISO 27001 等风险管理框架。

- 制定缓解策略和应急计划，以应对识别出的风险，包括回滚程序和应急机制，以防出现意外后果。

6. 变更实施与监控：

- 实施已批准的变更，遵循既定的部署流程和变更窗口，最大限度地减少对生产环境的干扰。
- 在实施变更后密切监控 AI 系统，使用监控和警报机制来检测预期行为或性能的偏差。

7. 文档记录与沟通：

- 记录变更控制流程的结果，包括已批准的变更、测试结果、风险评估和实施活动的详细信息。
- 向相关利益相关者（包括终端用户、管理层和监管机构）传达变更及其影响，必要时进行沟通。

8. 影响评估与优先级排序：

- 确保每个变更请求都记录变更对不同利益相关者和 AI 组件的潜在影响，以便在变更评估和审批流程中咨询所有相关利益相关者。
- 另外，应识别并评估每个变更对整个 AI 项目的优先级。

9. 变更请求关闭：

- 在变更过程的实施阶段结束时，应将变更的文档、日志和与变更相关的沟通存储在常规访问的位置，以供后续访问。
- 对所有参与变更流程的利益相关者举行闭会会议，可能是个不错的实践。

评估标准

- 遵循正式变更控制流程的变更百分比
- 审查和批准变更请求的平均时间
- 检测到的未经授权变更数量
- 导致事故或回滚的变更百分比
- 变更文档的完整性
- 变更后的验证成功频率
- 利益相关者对变更控制流程的满意度

责任矩阵 (RACI 模型)

- **执行**：AI 开发团队，DevOps 团队，质量保证团队
- **负责**：首席技术官
- **咨询**：业务单元领导，数据保护官，法律与合规部门
- **知情**：管理层，IT 安全团队，运营人员

高级实施策略

1. 制定正式的 AI 系统变更管理政策和流程。
2. 实施强大的 AI 工件版本控制系统。
3. 创建标准化的变更请求提交流程。
4. 开发结构化的审查和审批工作流程。
5. 设置全面的测试和验证程序。
6. 实施风险评估与缓解策略。

7. 建立监控和警报机制进行变更后的监控。

8. 创建变更管理的文档记录和沟通协议。

持续监控与报告

1. 跟踪变更请求的数量、批准和拒绝情况。

2. 监控每个变更控制流程阶段的时间指标。

3. 设置警报，以检测未经授权的变更或流程偏差。

4. 定期审查变更后 AI 系统的性能指标。

5. 生成有关变更控制有效性和合规性的定期报告。

6. 进行利益相关者调查，评估对变更管理流程的满意度。

访问控制映射

1. 限制变更实施权限，仅限授权的 DevOps 和 AI 开发团队。

2. 为变更管理工具和文档实施基于角色的访问控制。

3. 对高影响变更的批准权限仅限于高级管理人员或指定变更控制委员会。

4. 确保审计员和合规团队只能访问只读的变更日志和文档。

5. 实施多因素认证以访问关键的变更管理系统。

适用框架与法规

- **萨班斯-奥克斯利法案 (SOX)**：要求财务报告中的内部控制，适用于 AI 系统中的财务流程。

- **FDA 21 CFR 第 11 部分**：管理制药行业的电子记录和电子签名，适用于药物开发或制造中的 AI 系统。

- **ISO/IEC 27001**：为信息安全管理提供框架，包括变更控制流程。
- **信息技术基础架构库 (ITIL)**：尽管不是法规，但提供了 IT 服务管理的最佳实践，包括变更管理。

结论

本白皮书探讨了 AI 治理、风险管理和组织文化在 AI 实施背景下的关键方面。文档分为四个主要部分，每个部分重点关注组织在采用和管理 AI 技术时需要关注的关键领域。

在整篇文章中，对于每个责任项，始终一致地解决了六个跨领域关注点：

1. 评估标准
2. 责任矩阵（RACI 模型）
3. 高层实施策略
4. 持续监控与报告
5. 访问控制映射
6. 适用的框架和法规

文章首先通过定义组织各职能的责任角色，设定了有效 AI 管理的框架。随后深入探讨了风险管理策略，涵盖了威胁建模、风险评估、攻击模拟、事件响应规划和数据漂移监控等重要主题。

第二部分探讨了治理和合规性，概述了 AI 安全政策的制定、审计流程、董事会报告机制以及如何在复杂的监管环境中导航。还讨论了可测量控制和模型治理的实施。

第三部分专注于培养安全文化并提供全面培训，涵盖基于角色的教育、提高意识、负责任的 AI 培训以及有效的沟通策略。

最后一部分讨论了“影子 AI”防范的挑战，探讨了维护 AI 系统清单、进行差距分析、识别未授权系统、实施访问控制以及建立强有力的变更控制流程的方法。

本白皮书通过为每个责任项一致应用六个跨领域关注点，提供了关于 AI 治理的全面和结构化的方法。该框架确保组织能够全面评估、实施和管理其 AI 项目，同时解决关键领域，如问责制、实施策略、监控、访问控制和合规性。

CSA GCR cloud security alliance GREATER CHINA REGION ®

Cloud Security Alliance Greater China Region

扫码获取更多报告

云安全联盟大中华区
官网：https://c-csa.cn 电话：0755-86548359
地址：上海市浦东新区东育路255弄5号3FB30（前滩国际经济组织集聚区）/
广东省深圳市南山区粤海街道北京航空航天大厦1号楼905