AI组织责任
核心安全责任

AI Controls Framework
Working Group

CSA GCR cloud security
GREATER CHINA REGION alliance ®

“AI 组织责任工作组”的永久的官方网址是

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

联盟简介

云安全联盟（Cloud Security Alliance,CSA）是中立、权威的全球性非营利产业组织，于2009年正式成立，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，推动数字安全产业全面发展。

云安全联盟大中华区（Cloud Security Alliance Greater China Region, CSA GCR）作为CSA全球四大区之一，2016年在香港独立注册，于2021年在中国登记注册，是网络安全领域首家在中国境内注册备案的国际NGO，旨在立足中国，连接全球，推动大中华区数字安全技术标准与产业的发展及国际合作。

我们的工作

联盟会刊下载地址
了解联盟更多信息

加入我们

CSA大中华区官网
(https://c-csa.cn)

点击会员

加入联盟

填写相关申请信息

成为CSA会员

JOIN US

目录

致谢

报告中文版支持单位

联通（上海）产业互联网有限公司于 2018 年 3 月成立，注册资金 10000 万元，是中国联通在上海市成立的具有独立法人资格的全资子公司。公司积极顺应国家推动经济社会数字化、智能化发展的要求，正确认识和把握内外部科技发展环境新变化，坚持把创新作为引领发展的第一动力，聚焦 5G 时代发展的新机遇，以科技赋能差异化竞争与高质量发展。公司通过整合 5G、人工智能、大数据、ICT、IDC、云计算、物联网等能力，聚焦电子政务、工业互联网、智慧教育、智慧城市、智慧民生、文旅、交通物流、新商业等重点领域，为政府和各行各业客户提供从智能连接、云平台到大数据应用的个性化解决方案和集成运营一站式服务。2023 年公司总收入 6.4 亿元，净利润为 7415 万元。2023 年从业人员总数为 432 人，其中研发人员近 70%。全面承接生活数字化转型重点场景建设，标杆项目先行先试，形成示范效应，全面参与上海城市升级、产业升级，是上海建设智慧城市的领跑者之一。

参与本次报告的专家：

朱奕健：联通（上海）产业互联网有限公司 副总经理

李亚杰：联通（上海）产业互联网有限公司 云安全专家

吴振宏：联通（上海）产业互联网有限公司 云安全运营负责人

沈　爽：联通（上海）产业互联网有限公司 云安全专家

联通（上海）产业互联网有限公司是 CSA 大中华区理事单位，支持该报告内容的翻译，但不影响 CSA 研究内容的开发权和编辑权。

报告英文版编写专家

主要作者

Jerry Huang
Ken Huang

贡献者/联合主席

Ken Huang
Nick Hamilton
ChrisKirschke
Sean Wright

审稿人

Candy Alexander
Llango Allikuzhi
Eray Altili
Aakash Alurkar
Romeo Ayalin
Renu Bedi
Saurav Bhattacharya
Sergei Chaschin
Hong Chen
John Chiu
Satchit Dokras
Rajiv Gunja
Hongtao Hao, PhD
Grace Huang
Onyeka Illoh
Krystal Jackson
Arvin Jakkam Reddy
Reddy
Simon Johnson
Gian Kapoor
Ben Kereopa-Yorke
Chris Kirschke
Madura Malwatte
Madhavi Najana
Rajith Narasimhaiah
Gabriel Nwajiaku
Govindaraj Palanisamy
Meghana Parwate
Paresh Patel
Rangel Rodrigues
Michael Roza
Lars Ruddigkeit
Davide Scatto
Maria Schwenger Mj
Bhuvaneswari Selvadurai
Himanshu Sharma
Akshay Shetty
Nishanth Singarapu
Abhinav Singh
Dr. Chantal Spleiss
Patricia Thaine
Eric Tierling
Ashish Vashishtha
Peter Ventura
Jiewen Wang
Wickey Wang
Udith Wickramasuriya
Sounil Yu

CSA 全球员工

Marina Bregkou
Sean Heide
Alex Kaluza
Claire Lehnert
Stephen Lumpe

序言

在数字化时代，人工智能（AI）与机器学习（ML）技术正以惊人的速度重塑我们的工作与生活，并深刻影响全球经济与社会结构。随着 AI 技术的广泛应用，安全性、可靠性和合规性问题日益受到关注。作为云计算和 AI 安全领域的领导者，CSA 致力于推动行业安全标准与最佳实践的制定和实施。我们深知，随着 AI 与 ML 系统的发展，组织在保障安全方面肩负着重要责任。

在此背景下，CSA 发布了《AI 组织责任：核心安全责任》白皮书，深入探讨 AI 和 ML 系统在组织责任、信息安全和网络安全中的关键问题，并提供了全面的框架与指导，旨在帮助企业构建安全、合规的 AI 生态系统。

白皮书详细分析了数据真实性、匿名化、数据最小化等核心问题，并提出了一系列量化的评估标准与实施策略。同时，白皮书中引入了 AI 共享责任模型，明确界定了 AI 平台提供商、应用所有者、开发者与使用者之间的责任分工，并探讨了各方如何协同合作，确保 AI 应用的安全可靠运行。

此外，本白皮书还强调了持续监控和白皮书机制的重要性，呼吁各方遵循行业最佳实践与安全标准。我们相信，通过本报名的指导，组织能够更好地履行 AI 安全责任，共同推动 AI 技术健康发展。

李雨航 Yale Li

CSA 大中华区主席兼研究院院长

前言

本报告是一份工作草案，重点关注开发和部署人工智能（AI）和机器学习（ML）系统的过程中，在信息和网络安全方面的组织责任。本报告综合了专家推荐的核心安全领域的最佳实践，包括数据保护机制、模型漏洞管理、机器学习运营（MLOps）流水线强化以及负责任地训练和部署人工智能的治理政策。

报告中讨论的要点包括：

● 数据安全和隐私保护：数据真实性、匿名化、假名化、数据最小化、访问控制以及安全存储和传输在人工智能训练中的重要性。

● 模型安全：涵盖模型安全的各个方面，包括访问控制、安全运行环境、漏洞和补丁管理、MLOps 流水线安全、AI 模型治理和模型安全部署。

● 漏洞管理：探讨了 AI/ML 资产清单、持续漏洞扫描、基于风险的优先级排序、修复跟踪、异常处理和报告指标的对于有效漏洞管理的重要性。

报告采用可量化的评估标准、RACI（“执行、负责、咨询、知情”）角色定义模型、高层实施策略、持续监控和报告机制、访问控制映射以及对基本准则的遵循，逐一分析了各项责任。这些内容基于行业最佳实践和标准，如 NIST AI RMF、NIST SSDF、NIST 800-53、CSA CCM 等。

本文旨在为企业提供指导，通过在安全及合规的关键领域提出建议，帮助企业在 AI 设计、开发和部署方面履行负责任且安全的义务。

介绍

本报告重点介绍我们所定义的关于企业围绕人工智能（AI）和机器学习（ML）、数据安全、模型安全和漏洞管理的“核心安全责任”。鉴于组织有责任维护安全可靠的人工智能实践，本报告和本系列的另外两份报告为企业履行这些组织责任提供了蓝图。

具体而言，本报告综合了专家推荐的核心安全领域最佳实践，包括以下几个部分：数据保护机制、模型漏洞管理、MLOps 流水线强化以及负责任地训练和部署人工智能的治理政策；另外两篇报告则探讨了企业进行安全人工智能开发和部署的其他要点。本系列通过三份有针对性的报告，在关键的安全和合规领域提出建议，旨在指导企业履行其负责任和安全的开展人工智能设计、开发和部署的义务。

AI 共享责任模型

人工智能共享责任模型勾勒出 AI 平台提供商、AI 应用程序所有者、AI 开发人员和 AI 使用者之间的任务划分，根据服务模型（SaaS，PaaS，IaaS）的不同而有所变化。

人工智能应用程序的安全运行需要各利益相关者协同付出。在人工智能的背景下，责任由三个关键方共担：人工智能服务的使用者、人工智能应用程序所有者和开发者，以及人工智能平台提供商。

在评估人工智能赋能的集成（应用或服务）时，理解共享责任模型并明确各方处理的具体任务至关重要。

AI 赋能应用的关键层

人工智能平台

- 该层为应用程序提供 AI 能力。它涉及构建和保护承载 AI 模型、训练数据和配置参数的基础设施。
- 安全注意事项包括：阻止恶意输入以及避免 AI 模型生成恶意输出。安全的 AI 系统应该具备应对潜在的有害输入并避免有害输出的能力，比如宣扬仇恨，越狱等。
- AI 平台层包括以下任务：
  - 模型安全和防护
  - 模型调优
  - 模型责任
  - 模型设计与实现
  - 模型训练和治理
  - AI 计算和数据基础设施

AI 应用程序层

- AI 应用程序层利用 AI 的能力与用户交互，各应用的复杂性可能存在显著的差异。简言之，独立的 AI 应用就是为一系列 API 搭建的桥梁；这些 API 会处理来自用户的文本提示，并将其传递给底层模型以获得响应。复杂一点的 AI 应用程序能够利用持久层、语义索引或提供更广泛数据源访问的插件等手段，获得额外的上下文来丰富提示的内容。最先进的 AI 应用程序可以与现有的应用程序和系统无缝集成，实现多模态方法，支持文本、音频以及视觉等输入形式并产生多样化的内容输出。

- AI 应用程序的所有者需要确保无缝的用户体验，并处理其他相关的功能或服务。为了保护 AI 应用程序免受不法活动的影响，建立一个强大的应用程序安全系统至关重要。生成式人工智能（GenAI）系统应彻底检查发送到 AI 模型的提示词中的内容。此外，AI 编排所涉及的数据交换和交互也必须仔细审查，包括与插件和功能等附加组件及数据连接器的数据交换，以及与其他 AI 应用程序的交互。对于基于 IaaS 平台或 PaaS 服务的 AI 应用程序开发者来说，建议使用专门的人工智能内容安全功能，并根据具体要求，采用其他增强保护的功能。

- AI 应用程序层包括以下任务：

  - AI 插件和数据连接
  - 应用程序设计和实现
  - 应用程序基础架构
  - AI 安全系统

AI 使用

- AI 使用层描述了 AI 功能的应用方式和使用场景。生成式人工智能引入了与 API、命令提示和 GUI 等传统界面不同的创新的用户/计算机交互模型，这种新的界面具有交互性和适应性，可以根据用户的意图调整计算机的功能。与要求用户适应系统设计和功能的早期界面不同，生成式人工智能界面优先考虑用户交互。这使得用户的输入能够显著地影响系统的输出，强调了安全机制对保护个人、数据和企业资源的重要性。

- AI 使用层的安全注意事项与计算机系统类似，它依赖在身份和访问管理、设备安全、监控、数据治理和管理控制等方面的稳健措施。

- 鉴于用户行为可能对系统输出产生重大影响，有必要更加关注用户行为和责任。必须修订可以被接受的使用政策，并告知用户传统应用程序和 AI 增

强型应用程序的区别。这涵盖了 AI 相关的安全、隐私和道德标准等问题。此外，提升用户对基于 AI 攻击的潜在风险的认知是非常重要的，这种攻击形式可能涉及包含精心伪造的文本、音频、视频和其他用作欺骗的媒体内容。

- AI 使用层包括以下任务：
  - 用户培训和问责机制
  - 可接受的使用策略和管理控制
  - 身份和访问管理（IAM）和设备控制
  - 数据治理

请谨记，共享责任模型有助于划分角色，并确保职责清晰分工，从而促进安全高效地使用人工智能技术。对于不同的 AI 集成的类型，工作负载责任的分配会有所不同。

软件即服务（SaaS）

- 在基于 SaaS 的人工智能集成中，人工智能平台提供商负责管理底层基础设施、安全控制和合规措施。
- 用户的主要重点在于配置和定制 AI 应用程序，以符合相应的特定要求。

平台即服务（PaaS）

- 基于 PaaS 的 AI 平台提供了一个中间层。AI 提供商管理着核心的人工智能功能，但用户仍然可以进行定制化的配置和控制。
- 用户有责任确保安全地使用 AI 模型、处理训练数据，以及调整模型（例如权重和偏置）。

基础设施即服务（IaaS）

- 在 IaaS 场景中，用户可以更好地控制基础设施，这也意味着要承担更多的责任。
- 用户负责全栈管理，包括 AI 模型、训练数据和基础设施的安全。

以数据为中心的人工智能系统的基础组件

以数据为中心的人工智能系统的基础组件涵盖了数据和模型管理的整个生命周期。这些基础组件协同工作，以安全高效的人工智能系统，能够处理数据并提供有价值的深度分析或自动化决策。

- 原始数据：从各种来源收集的原始未处理的数据。
- 数据准备：将原始数据清理和组织成结构化格式的过程。
- 数据集：经过整理的数据集，可用于分析和模型训练。
- 数据和人工智能治理：确保数据质量和人工智能使用伦理的政策和程序。
- 机器学习算法：用于解释数据的计算方法。
- 评估：评估机器学习模型的性能。
- 机器学习模型：基于数据集训练的算法的成果。
- 模型管理：监督机器学习模型的生命周期。
- 模型部署和推理：实施模型以做出预测或决策。
- 推理结果：部署模型产生的结果。
- 机器学习运营（MLOps）：部署和维护AI模型的实践。
- 数据和人工智能平台安全：保护系统免受威胁的措施。

数据运营：涉及数据的获取和转换，以及确保数据安全和治理。机器学习模型的有效性取决于数据流水线的完整性和强化的 DataOps 框架。

模型运营：包括创建预测性机器学习模型、从模型市场采购，或使用大型语言模型（LLM）（如 OpenAI 提供的模型或调用各类大模型 API）。模型开发是一个迭代过程，需要系统的方法来记录和评估各种实验条件和结果。

模型部署和服务：涉及模型容器的安全构建、模型的隔离和保护部署，以及对活动模型的自动扩展、速率限制和监视的实施。它还包括为检索增强生成（RAG）应用程序中的高可用性、低延迟服务提供特性和功能，以及为其他应用程序提供必要的特性，包括在平台外部部署模型或需要目录中的数据特性的应用程序。

运营和平台：涵盖平台漏洞、更新、模型隔离和系统控制的管理，以及在安全架构框架内实施授权的模型访问。此外，它还涉及部署用于持续集成/持续部署（CI/CD）的运营工具，确保整个生命周期在独立执行环境（如：开发、预生产及生产）中遵守既定的标准，以实现安全的机器学习运营（MLOps）。

表 1 将运营与以数据为中心的人工智能系统的核心组件相关联，突出了它们的角色和相互依赖性。

表 1：以数据为中心的人工智能系统组件及其相互关联的角色

基础组件	说明
数据运营	数据的摄取、转换、安全和治理。
模型运营	构建、获取和试验机器学习模型。
模型部署和服务	机器学习模型的安全部署、服务和监控。
运营和平台	MLOps的平台安全性、模型隔离和CI/CD。

表 2 提供了 AI/ML 系统每个阶段的潜在风险和威胁的综合视图，以及解决这些问题的示例和建议的缓解措施。

表 2:AI/ML 安全风险概述

系统阶段	系统组件	潜在安全风险	威胁	缓解措施
数据运营	原始数据、数据准备、数据集	数据丢失：未经授权删除或损坏数据。数据中毒：故意操纵数据以损害模型的完整性。合规挑战：未能满足数据保护的监管要求。	数据泄露/中毒：攻击者可能会注入虚假数据或更改现有数据。	实施稳健的数据治理框架。部署异常检测系统。建立恢复协议和定期数据备份。
模型运营	ML算法，模型管理	模型盗窃：窃取专有模型。未经授权的访问：未经许可访问模型。	通过API访问的攻击：利用API漏洞访问或操纵模型。模型窃取（提取）：复制模型以供未经授权的使用。	加强访问控制和身份验证机制。通过加密和速率限制来保护API端点。定期更新和修补系统。
模型部署和服务	模型服务、推理响应	未经授权的访问：未经授权访问模型服务基础设施。数据泄露：错误的系统配置导致暴露敏感信息。	模型欺骗（逃逸）：改变输入以从模型中接收特定输出。训练数据恢复（反演）：从模型中提取私人训练数据。	安全部署实践，包括容器化和网络分段。主动监控和记录模型交互。实施速率限制和异常检测。
运营和平台	机器学习运营、数据和人工智能平台安全	漏洞管理不足:未及时解决已知漏洞。模型隔离问题:未能正确隔离模型，导致潜在的交叉污染。	攻击机器学习供应链:在第三方组件中引入漏洞或后门。模型污染（投毒）：破坏训练数据，导致错	持续的漏洞管理和修补。用于一致部署的CI/CD流程。隔离控制和安全架构设计。

			误分类或系统不可用。	

我们从以下维度分析每项责任：

1. **评估标准**：在我们讨论人工智能责任时，应考虑可量化的指标来评估人工智能系统的安全影响。通过量化这些指标，利益相关者可以更好地了解人工智能技术的相关风险以及如何应对这些风险。组织必须经常评估其人工智能系统，以确保其安全性和可靠性，例如应该评估：系统处理攻击的能力（对抗鲁棒性），是否泄漏敏感数据，出错的频率（假阳性率），以及训练数据是否可靠（数据完整性）等。作为组织安全计划的一部分，评估和监控这些关键措施将有助于提高人工智能系统的整体安全状况。

2. **RACI 模型**：该模型有助于明确在人工智能决策和监督过程中的执行者、负责人、咨询方和知情方（RACI，Responsible, Accountable, Consulted, and Informed；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，文本将 Responsible 翻译为执行，后文中 RACI 即为：执行、负责、咨询和知情）。应用 RACI 模型描述了人工智能治理中的角色和责任分配，这种责任分配对于安全的人工智能系统至关重要。当然，根据组织的规模和业务重点，本报告中描述的具体角色和团队仅供参考。首先，明确描述关键责任是重中之重。其次，组织可以根据职责规划适当的角色，然后为这些角色配备相应的团队，团队之间可能会存在一些职责重叠。本文定义的 RACI 框架旨在提供初始角色和团队配置，以帮助组织开发其定制化的 RACI 模型。然而，各企业实施过程中可能因其独特的组织结构和优先事项而异。

3. **高层实施策略：**本节描述了将网络安全的注意事项无缝集成到软件开发生命周期（SDLC）中的策略。组织必须优先执行信息安全的 CIA 原则：即确保数据和系统的机密性、完整性和可用性。同时应严格实施访问控制机制，以管理用户权限并防止未经授权的访问。必须建立健全的审计机制，以跟踪系统活动并及时发现可疑行为。影响评估应分析判断潜在的网络安全风险，特别是漏洞识别和威胁缓解方面，以保护人工智能系统中的敏感信息。

4. **持续监控和报告：**持续监控和报告能够确保人工智能系统的持续防护、安全并保证性能。关键组件包括实时监控、对模型性能异常或安全事件告警、审计跟踪/日志、和定期报告，以及能够改进或解决问题的措施。持续监控和报告有助于组织保持透明度，提升效率并明确责任归属，以及建立对人工智能系统的信任。

5. **访问控制：**访问控制对于保护人工智能系统至关重要，包括严格的 API 身份验证/授权策略、管理模型注册表、控制对数据存储库的访问、监督持续集成和部署流水线（CI/CD）、处理机密信息以及管理特权访问。通过为 AI 流水线的各个部分定义用户角色和权限，可以保护敏感数据，防止模型被篡改或未经授权访问。通过强身份认证和访问管理不仅可以保护知识产权，还可以明确 AI 工作流的责任归属。

6. **遵守基础治理、风险与合规、防护、安全和道德标准：**

强调遵守基于行业最佳实践和监管要求，如下所示：

- NIST SSDF 用于安全软件开发
- NIST 人工智能风险管理框架（AI RMF）
- ISO/IEC 42001:2023 人工智能管理系统（AIMS）
- ISO/IEC 27001:2022 信息安全管理体系（ISMS）
- ISO/IEC 27701:2019 隐私信息管理系统（PIMS）
- ISO 31700-1:2023 消费者保护：面向消费品和服务的隐私设计

- 面向大型语言模型（LLM）应用的 OWASP 十大安全风险列表 （OWASP Top 10 for LLM Applications）
- NIST SP 800-53 Rev.5 信息系统和组织的安全和隐私控制
- 《通用数据保护条例》（GDPR）关于数据匿名化和假名化的技术规范和相关指导意见
- 关于云服务中令牌化技术（Tokenization）的指导意见

前提条件

本报告采取行业中立的立场，提供适用于各个行业的指导方针和建议，而不会对特定行业产生具体偏见。

目标受众

本报告旨在满足不同受众的需求，每个受众都有不同的目标和兴趣。

1. **首席信息安全官（CISO）：**本报告专门针对 CISO 的关切和责任而设计，为在人工智能系统中集成核心安全原则提供了意见。值得注意的是，首席人工智能官（CAIO）的角色正在许多组织中出现，预计在不久的将来，本报告中定义的大多数相关职责可能会从首席信息安全官转移到 CAIO。

2. **AI 研究人员、工程师、数据专业人员、科学家、分析师和开发人员：**该报告为 AI 研究人员和工程师提供了全面的指导方针和最佳实践，帮助他们开发合乎道德和值得信赖的 AI 系统。它是确保负责任的人工智能发展的关键资源。

3. **商业领袖和决策者：**对于首席信息官、首席产品官、首席数据官、首席风险官、首席执行官和首席技术官等商业领袖和政策制定者来说，报告提供了与人工智能系统开发、部署和生命周期管理相关的网络安全战略的重要的信息和认识。

4. **政策制定者与监管机构：**政策制定者和监管机构将发现这篇报告非常有价值，因为它提供了关键的见解，有助于制定有关人工智能伦理、安全和控制的政策和监管框架。它为人工智能治理领域的知情决策提供了指导。

5. **投资者和股东：**投资者和股东将会欣赏这份报告，因为它展示了一个组织对负责任的人工智能实践的承诺。它强调了确保人工智能道德发展的治理机制，这对投资决策至关重要。

6. **客户和公众：**本报告为客户和公众提供了关于组织在开发安全 AI 模型时的价值观和原则的透明度。

职责角色定义

下表提供了一个通用指南，说明了在集成或操作人工智能技术的组织中常见的各种角色。我们必须认识到，每个组织可能会以不同的方式定义这些角色及其相关职责，反映其独特的运营需求、文化以及其人工智能计划的具体要求。因此，虽然该表提供了对人工智能治理、技术支持、开发和战略管理中潜在角色的基本理解，但仅供参考。我们鼓励各组织调整和定制这些角色，以最好地满足组织的定制化要求，确保结构和职责与其战略目标和运营框架保持一致。随着人工智能技术的发展，可以进一步定义新的角色。

管理与战略

角色名称	角色描述

首席数据官（CDO）	监督企业数据管理、策略创建、数据质量和生命周期。
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）	监督信息安全战略和运营。
业务部门（BU）负责人	指导业务部门，使人工智能计划与业务目标保持一致。
首席人工智能官（CAIO）	负责组织内人工智能技术的战略实施和管理。
管理层	监督和指导整体战略，确保与组织目标保持一致，包括CEO、COO、CIO、CTO、CISO、CAIO、CFO等。
首席云官	领导云战略，确保云资源与业务和技术目标保持一致。
首席架构师	领导架构战略，确保设计技术架构与企业标准、流程、程序和目标保持一致。技术选型，监督设计的质量和实施，在组织内培养高级架构师。

治理与合规

角色名称	角色描述	类别名称
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
数据保护专员	监督数据保护策略和遵守数据保护法律法规。	治理与合规
首席隐私官	确保遵守隐私法律法规。	治理与合规
法律团队/部门	提供有关人工智能部署和使用的法律指导。就法律/监管义务进行沟通。确保与人工智能供应商签订的主协议中有适当的条款。	治理与合规
合规团队/部门	确保遵守内部和外部合规要求。	治理与合规
数据治理官	管理组织内的数据治理，确保符合政策、数据隐私法和监管合规要求。	治理与合规

信息安全官	授权官员、管理官员或信息系统所有者，以确保为包括ISSO、ISM和ISS在内的信息系统或项目保持适当的运营安全态势。	治理与合规

技术和安全

角色名称	角色描述
安全运营团队	实施和监控安全协议以保护数据和系统。
网络安全（Network Security）团队	保护网络免受威胁和漏洞的侵害
云安全团队	确保基于云的资源和服务的安全性。
网络空间安全（Cybersecurity）团队	保护组织资产免受网络空间威胁、漏洞及未经授权访问的侵害。
IT运营团队	支持和维护IT基础设施，确保其运营和安全。
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，以实现最佳性能和安全性。

运营与发展

角色名称	角色描述
数据管理人	负责安全保管、运输、数据存储和业务规则的实施。这是代表数据所有者获取、操纵、存储或移动数据的任何组织或个人。
AI开发团队	开发和实施AI模型和解决方案。
质量保证团队	测试并确保AI应用程序和系统的质量。
人工智能运营团队	管理人工智能系统的运行，以确保其性能和可靠性。

应用程序开发团队	开发应用程序，根据需要集成AI功能。
AI/ML测试团队	专门测试AI/ML模型的准确性、性能和可靠性。
开发运营（DevOps）团队	提高部署效率，保持运营稳定。
开发安全运营（DevSecOps）团队	在整个软件开发生命周期（SDLC）中实施安全性。
AI维护团队	确保人工智能系统和模型在部署后得到更新、优化和正确运行。
项目管理团队	监督人工智能项目从启动到完成，确保其达到目标和时间表。
运营人员	支持日常运营，确保人工智能技术的顺利集成和运行。
数据科学团队	收集并准备数据，用于AI模型训练和分析。
容器管理团队	管理容器化应用程序，促进部署和可扩展性。
IT运营团队	确保IT基础设施正常运行，支持人工智能和技术需求。
AI开发经理	领导人工智能开发项目，指导团队成功实施。
人工智能运营主管	指导与人工智能相关的运营，确保人工智能解决方案的效率和有效性。

规范性引用文件

以下列出的文件对于应用和理解本文件至关重要。

- Generative AI safety: Theories and Prac tices
- OpenAI Preparedness Framework
- Applying the AIS Domain of the CCM to Generative AI
- EUAI Act
- Biden Executive Order on Safe, Secure, and Trust worthy Artificial Intelligence
- OWASP Top 10 for LLM Applications
- CSA Cloud Controls Matrix (CCM v4)

- MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems)
- NIST Secure Software Development Framework(S SDF)
- NIST Artificial Intelligence Trust worthiness and Risk Management Framework-
- General Data Protec tion Regulation (GDPR)
- OWASP LLM AI Cybersecurity & Governance Checklist
- OWASP Machine Learning - Top 10
- WEF Briefing Papers
- Building the AI-Powered Organization

一、 将数据安全和隐私纳入人工智能训练

人工智能正在从复杂的以模型为中心的方法转向以数据为中心的方法。人工智能现在不再主要依赖于在小数据集上训练的复杂模型，而是利用海量数据集和开放式数据流。然而，这种以数据为中心的范式也引发了人们对数据隐私、安全、偏见和正当使用的合理担忧，AI 社区必须负责任地应对这些问题。数据正在改变人工智能，但我们必须确保它的来源是符合伦理且受控的。

以下部分讨论了与确保人工智能组织中培训数据的安全性和隐私性相关的重要类别。这些类别包括数据真实性、匿名/假名化、数据最小化、数据访问控制，以及安全存储与传输。每个类别都通过可量化的评估标准、通过 RACI 模型明确界定的责任、高层实施策略、持续监控和报告机制、访问控制映射以及基于行业最佳实践的基本准则进行了彻底分析。这种全面的方法确保了一个结构化、负责任和高效的框架，用于管理推动人工智能进步的重要资产，同时也符合道德要求和卓越运营。

1.1 数据真实性和许可管理

AI 中的数据真实性是指保证用于训练、测试和部署 AI 模型的数据是真实、准确且可靠的。这涉及验证数据没有被篡改或更改，以免误导人工智能算法或导致不准确、有偏见或不可靠的模型输出。

确保数据真实性至关重要，因为 AI 模型严重依赖数据质量和完整性。如果数据不真实或被操纵，模型可能会学到不正确的模式，导致性能不佳，并可能基于预测做出有害的决策。数据真实性在基于 AI 模型的决策具有重大影响的对各个行业领域尤为重要，例如教育、医疗保健、金融服务、零售、制造、政府服务和网络安全。

此外，在为人工智能应用程序收集和处理个人数据时，必须获得适当的数据使用同意并遵守《通用数据保护条例》（GDPR）等法规。GDPR 要求组织在收集和处理个人数据之前获得个人的明确同意，并赋予个人访问、更正和删除其数据的权利。

- 评估标准：定期测量经过真实性审核的数据百分比，目标是在规定时间内进行 100%验证。此外，监控数据许可和 GDPR 法规的合规情况。
- 在任何可能的情况下，个人都应有权更正有关他们的数据。例如：联邦贸易委员会在保险方面的要求， 联邦贸易委员会推行人工智能监管，禁止有倾向性的算法。
- RACI 模型：数据管理团队（执行）、首席数据官（负责）和法律合规部门（咨询）、安全团队（告知）。
- 高层实施策略：实施定期数据真实性审计的政策，其中可能涉及数据来源检查和异常检测等技术。此外，建立获取数据同意、确保数据隐私和遵守 GDPR 规定的流程。
- 持续监测和报告：用于认证的已核实数据的定期报告、未经授权的数据变更以及遵守数据使用同意和 GDPR 规定。

通过确保数据真实性、获得适当的数据使用同意并遵守 GDPR 等法规，组织可以在尊重个人隐私和个人数据权利的同时建立值得信赖的 AI 模型。

1.2 匿名化和假名化

匿名化和假名化在值得信赖的人工智能系统中保护个人数据隐私，如下所示：

- 匿名化永久地从数据中删除标识符，这使得重新识别个人身份变为不可能，有助于遵守数据保护法，在许多情况下，例如根据 GDPR，甚至可以将匿名数据排除在数据保护法规的范围之外。
- 假名化将标识符替换为系统生成的标识符或人工标识符假名。个人仍然与他们的数据保持联系，但真实身份受到保护。根据某些数据保护法规，如 GDPR 和 HIPAA，假名化是一项要求。
- 评估标准：通过匿名化和假名化技术，将可识别的个人数据减少 99%。区分直接标识符（如全名、SSN 和信用卡号）和准标识符非常重要。直接标识符需要更严格的削减措施，因为能够使用它们来高度确定地识别个人。此外，信用卡号等直接标识符可能导致盗窃，而 SSN 等标识符可能导致身份盗窃。准标识符（年龄、邮政编码和性别）虽然对隐私很重要，但可以进行不太严格的削减，以确保数据保护和可用性之间的平衡。尽管如此，许多准标识符可能会在人工智能中引发偏见（例如，年龄、地点、性别、种族、性取向等）。此外，某些准标识符可能属于特殊类别的个人数据（GDPR 第 9 条-特殊类别个人数据的处理）（https://gdpr-info.eu/art-9-gdpr/），需要特别谨慎，包括宗教信仰、政治派别、性取向和种族血统。准标识符也可能被组合起来重新识别一个人。虽然准标识符对于人工智能系统的可用性和功能至关重要，特别是在医疗保健或营销等领域，但它们存在重新识别的风险。当不同的数据集被组合在一起时，即使它们已经被匿名或假名化，这些准标识符也可能被匹配，从而重新识别个人。例如包含匿名医疗记录的数据集可以与公开可用的数据集相结

合，如选民登记记录。如果两个数据集都包含详细的人口统计信息，则有可能基于这些准标识符匹配记录，从而重新识别匿名数据集中的个人。为了降低这种风险，需要采取平衡的方法。不断评估重新识别的风险至关重要，尤其是随着数据处理技术的发展和变得更加复杂。采用如差分隐私等先进技术，在数据中添加统计噪声以防止重新识别，可以进一步加强隐私保护。此外，定期审计和合规检查对于确保数据匿名化和假名化过程符合不断发展的数据保护法律法规至关重要。

- RACI 模型：数据保管人（执行）、数据保护官（咨询）、首席隐私官主管（负责）、法律团队（咨询）、IT 团队（告知）、安全团队（咨询）、数据治理团队（咨询）和数据专家（告知）。

- 高层实施策略：实施最先进的匿名化和假名化技术涉及利用先进的加密方法，如差分隐私、同态加密和安全多方计算来保护敏感数据，同时保持其分析效用。例如，公司可以采用 k-匿名（k-anonymity）、l-多样（l-diversity）和 t-相近（t-closeness）等技术来确保个人身份隐藏在数据集中，同时仍然允许进行有意义的分析。此外，可以采用令牌化（tokenization）和数据屏蔽等技术，用不敏感的等效数据替换敏感数据，进一步加强隐私保护。

- 持续监测和报告：定期评估这些措施的有效性技术。

- 访问控制映射：限制对编辑或匿名化规则和去假名化工具的访问。

- 基本准则：遵循《通用数据保护条例》（GDPR）关于数据匿名化和假名化的指导方针，以及基于云服务的令牌化指南。

1.3 数据最小化

数据最小化是指只使用实现特定目的或功能所需的必要数据量的做法。这种做法是许多数据保护法规（如 GDPR）的要求。这也是可用于防止匿名数据重新识别

的技术之一。这种技术限制了为机器学习目的收集、存储和使用的数据的数量和类型。这种做法有助于保护数据主体的隐私和安全，并提高机器学习模型的性能和效率。在机器学习中，数据最小化涉及仔细选择对模型训练和性能至关重要的特征和数据点，同时排除无关或过多的数据。这与可解释性、公平性、透明度和隐私等可信赖的人工智能的基础支柱有关。

- 评估标准：根据组织的业务目标和责任，力求减少非必要数据的收集，降幅至少达到一个显著的百分比。
- RACI 模型：数据收集团队（执行）、数据隐私办公室（咨询）、数据治理委员会（负责），合规团队（咨询），安全团队（告知），数据专家（告知）。
- 高层实施策略：制定严格的数据收集指南，重点关注最小的数据采集。
- 持续监测和报告：跟踪收集的数据量并评估其必要性。
- 访问控制映射：对谁可以授权额外的数据收集实施控制。
- 基本准则：采用 GDPR 中的隐私设计原则。

1.4 数据访问控制

机器学习中的数据访问控制涉及管理和限制谁可以访问用于训练、测试和部署机器学习模型的数据并与之交互。此过程确保只有授权的个人或系统才有能力查看、修改或使用数据。在机器学习环境中，有效的访问控制对于保护敏感信息、维护数据完整性和遵守隐私法规至关重要。它通常涉及验证用户身份的认证机制、根据用户角色授予用户特定访问权限的授权协议，以及跟踪数据访问和使用情况的审计系统。在组织中，AI 模型通常是在从各种数据源或系统聚合的数据上运行的。因此，AI 模型应尊重底层数据源系统的访问控制定义和策略。这意味着，AI 模型应该只能访问他们被授权处理的特定数据，这些授权来自数据管理人或系统管理员定义的访

问控制规则和权限。保持适当的访问控制可确保人工智能基础设施的数据隐私、安全，并符合监管要求，同时防止未经授权访问或 AI 模型滥用敏感或机密数据。

- 评估标准：每年未经授权的数据访问事件低于 0.5%
- RACI 模型：安全团队（执行）、首席信息安全官（负责），数据治理机构、数据管理人、IT 团队（咨询）、运营团队（告知）。
- 高层实施策略：实施分层安全模型，作为访问控制高层实施策略的一部分。该模型不仅应集成健全的身份验证和授权协议，还应集成多因素认证（MFA）、基于角色的访问控制（RBAC）和最小特权原则（PoLP）等先进技术。
- 持续监控和报告：监控访问日志并进行审计。采用工具基于风险标记对关键模型、生成模型和数据的访问。
- 访问控制映射：定期监控和管理访问权限。
- 基本准则：实施 ISO/IEC 42001、ISO/IEC 27001、ISO/IEC 27701、NIST 800-53 和 OWASP Top 10 A07:2021（身份识别和验证失败）中的最佳实践。

1.5 安全存储和传输

在机器学习中，安全存储和传输对于保护敏感数据至关重要。安全存储涉及加密静态数据以防止未经授权的访问，采用健全的访问控制，并定期进行安全审计。为了安全传输，传输中的数据使用传输层安全（TLS）、字段或信封加密等协议进行加密。这确保了数据在系统或网络之间传输时保持机密性和完整性。通过防止未经授权的访问、使用和泄露数据，以及恶意或意外修改、删除或损坏数据，增强了数据和 ML 模型的安全性。

- 评估标准：确保在传输和静止状态的数据全部采用 256 位 AES 或更高级别的加密标准。
- RACI 模型：安全团队（执行）、首席信息安全官（负责），合规和法律团队（咨询），管理层（告知）。
- 高层实施策略：投资先进的加密技术，并根据策略自动删除人工智能数据和模型。
- 持续监控和报告：使用工具进行实时安全监控。
- 访问控制映射：将安全存储和传输与访问控制集成在一起。
- 基本准则：遵循美国国家标准与技术研究院（NIST）的指导方针和隐私法，保护传输和静止的数据。

二、 模型安全

模型安全是一项多方面的任务，包括各种各样的组件。这些包括对模型 API 的访问控制、身份验证和授权框架、速率限制、模型生命周期管理、安全的模型运行环境、基于硬件的安全功能、网络安全控制、操作系统强化、安全配置，以及容器和云安全。我们应该探索这些关键领域中的每一个评估标准，基于 RACI 模型分配责任，描绘高层实施策略，建立持续监控和报告机制，配置访问控制，并参考 NIST AI RMF、NIST SSDF、NIST 800-53 和 CSA CCM 等标准中的基本准则。

2.1 模型访问控制

访问控制对于保护 AI 模型至关重要，确保只有授权人员和系统才能与敏感数据和功能交互。在 AI 模型治理领域，访问控制措施必须稳固健全、适应性强，并与组织和行业安全标准保持一致。

从身份验证和授权，到速率限制和生命周期管理，AI 模型的完整性取决于强大而灵活的访问控制协议。这些协议规定了谁可以访问 AI 模型，何时可以访问，以及在什么情况下可以访问。随着组织应对人工智能部署的复杂性，实施全面的访问控制战略势在必行，以降低风险、保护知识产权和遵守监管合规标准。此外，我们强调将 AI 模型访问控制与组织现有的安全框架相结合，以增强整体系统的韧性。这涉及制定细粒度访问策略，规定谁可以与 AI 模型交互以及在什么情况下交互。此外，实施健全的身份验证机制，如多因素认证和基于角色的访问控制，可以进一步加强人工智能系统的安全。定期审计和监控访问日志对于及时发现和应对任何未经授权的访问尝试至关重要。通过将 AI 模型访问控制与现有安全框架紧密结合，组织可以加强对网络威胁的防御，并确保其人工智能系统和数据的完整性和机密性。

2.1.1 身份验证和授权框架

机器学习模型的身份验证和授权框架确保了对机器学习模型和相关数据的访问受到严格控制和管理，对于安全至关重要。身份验证通常验证用户或系统的身份，常用密码、令牌或生物特征验证等方法。同时，通过访问级别的授权，定义了谁可以根据既定的角色和权限查看、编辑或使用模型。这对于保护敏感信息、保持数据完整性以及遵守隐私和安全法规至关重要，从而防止对机器学习模型及其数据的未经授权的访问或修改。在人工智能中，特定的验证是所有用户和实体根据目的和背景批准、适当地使用人工智能数据和模型。这与数字版权相融合，是访问和授权的基础。

- 评估标准：对所有未通过 API 访问的 AI 模型，应确保认证和授权框架实现 100% 覆盖（参见 1.2.1）。
- RACI 模型：安全团队（执行）、首席信息安全官（负责），法律团队（咨询），人工智能开发团队（告知）。
- 高层实施策略：开发和实施安全的 AI 模型访问的综合框架。
- 持续监控和报告：定期审核身份验证和授权机制。

- 访问控制映射：根据模型特定要求自定义访问。
- 基本准则：使用 NIST 800-207、NIST 800—53、NIST SP 800—63 和 NIST AI RMF 进行风险管理。

2.1.2 模型接口速率限制

机器学习中的模型接口速率限制涉及限制用户或系统在给定时间范围内向模型发出的请求数量。这种做法对于管理模型上的负载、防止滥用（如拒绝服务攻击）以及确保用户之间的公平资源分配至关重要。速率限制可以在用户与 ML 模型交互的各类接口层级实现，例如：API 或 web 界面。控制请求率有助于保持模型的性能、稳定性和可用性，确保其即使在高需求或潜在攻击场景下也能持续高效可靠地运行。

- 评估标准：减少因拒绝服务（DoS）或分布式拒绝服务（DDoS）攻击导致的停机时间。
- RACI 模型：平台支持团队（执行）、解决方案负责人（负责）、数据专家（咨询）、风险管理团队（咨询）和 AI 模型用户（告知）。
- 高层实施策略：实施速率限制，防止过度使用或滥用 AI 模型接口。
- 持续监控和报告：跟踪使用模式并相应调整速率限制。
- 访问控制映射：实施基于用户的速率限制策略。
- 基本准则：遵循 OWASP LLM 04：模型拒绝服务。

2.1.3 模型生命周期管理中的访问控制

在机器学习模型的全生命周期管理中，访问控制涉及在模型生命周期的各阶段（开发、部署和维护阶段）管理和规范对 ML 模型的访问和交互。此过程确保只有授权人员或系统才能在各个阶段与 ML 模型交互，从而保护模型免受未经授权的访

问或更改，这种访问或更改可能会导致模型完整性受损或出现性能问题。实施健全的访问控制对于维护机器学习模型的安全性和有效性至关重要，因为它有助于防止可能的数据泄露、模型滥用，并能确保符合数据隐私和安全的相关法规。在模型生命周期的所有阶段实施控制访问，组织可以保护他们的机器学习资产，同时构建安全高效的机器学习的开发环境。模型生命周期管理中的访问控制能够提高机器学习模型的透明度和责任归属，能够为数据访问使用提供一致和明确的策略和程序，并记录数据来源地址和目的地址。这一领域与隐私性、透明度和责任归属等可信赖的人工智能的基础支柱有关。

- 评估标准：确保在模型生命周期的所有阶段，对 AI 模型和数据的访问仅限于授权用户和系统。
- RACI 模型：AI 模型治理团队（执行），首席数据官（责任），安全团队、法律团队、合规团队（咨询），运营人员（告知）
- 高层实施策略：
  - 根据敏感度对数据和模型进行分类。
  - 为模型生命周期的每个阶段定义明确的访问控制规则，并关联到各用户角色。
  - 将访问控制与现有的 IAM（身份和访问管理）解决方案集成。
  - 记录并审核所有访问请求以及数据/模型使用情况。
- 持续监测和报告：
  - 在未经授权的访问尝试时发送警报。
  - 执行用户访问审查和重新认证。
  - 定期对访问和控制进行审计。

- 为可疑活动建立警报阈值。

- 访问控制映射：
  - 开发阶段：仅允许数据科学家和机器学习工程师访问
  - 测试阶段：为质量保证团队添加访问权限
  - 生产阶段：授予严格控制的生产系统访问权限

- 基本准则：
  - 符合 NIST 800-53、NIST AI RMF 框架等标准
  - 根据 CSA CCM 等最佳实践框架验证控制措施。

2.2 安全的模型运行环境

为安全的 AI 模型运行环境构建能够具有韧性的系统需要融合强大的硬件、网络和软件安全控制。在全面关注保护 AI 部署免受持续演进的威胁侵扰的同时，组织应该精心设计和增强其运行时环境，以维护其完整性、机密性和可用性。从基于可信执行环境的硬件安全功能到防火墙和网络隔离等网络安全控制机制，每个组件都精细地融入在深度防御策略的体系中。依据 NIST AI RMF、NIST 800-53 和 CSA CCM 等行业标准，我们应进行团队的跨学科合作，协调实施、监控并管理这些关键的安全措施。

2.2.1 基于硬件的安全功能

机器学习模型的基于硬件的安全特性包括计算硬件中的物理环境和整体架构，这些元素增强了机器学习应用程序的安全性。具体包括：可信执行环境（TEE），用于隔离和安全处理的机密计算，用于保护敏感代码和数据的安全隔离区（Secure Enclaves），用于安全加密操作的硬件安全模块（HSM），确保可信软件初始化的安全

引导机制，以及防止未经授权的物理访问的物理防篡改机制。这些功能在提供基础安全层方面至关重要，尤其是在金融、医疗保健或国防等高风险行业；机器学习模型通常都处理敏感数据，并且非常需要针对各种威胁（包括篡改和未经授权的访问）进行强有力的保护。

- 评估标准：在适用的情况下，公司定义的人工智能系统使用基于硬件的安全功能的百分比。
- RACI 模型：硬件安全团队（执行）、首席技术官（负责）、采购部门（咨询）、系统管理员（告知）。
- 高层实施策略：在 AI 系统中集成可信执行环境（如 NVIDIA 的机密计算方法）、GPU、TPU 和其他硬件安全措施。
- 持续监控和报告：定期检查硬件安全。
- 访问控制映射：确保只有授权人员才能访问硬件安全设置。
- 基本准则：实施 NIST AI RMF 和 NIST 800-53 指南。

2.2.2 网络安全控制

机器学习模型的网络安全控制是为保护机器学习模型及其相关数据免受基于网络的威胁和漏洞而实施的措施和协议。通过采用零信任架构，将人工智能系统与更广泛的网络隔离开来。它通过隔离人工智能系统和机器学习模型来减少攻击面，使攻击者更难在网络中横向移动。这些控制对于在传输过程中保护机器学习模型中使用的数据、防止未经授权的访问以及确保机器学习通信的完整性和机密性至关重要。网络安全的关键控制包括：使用下一代防火墙进行监测、控制流入、流出的网络流量，使用 TLS 等加密协议保护传输中的数据，使用入侵检测与预防系统（IDPS）以及 Web 应用程序防火墙（WAF）来识别和缓解攻击，使用虚拟专用网络（VPN）来创建安全通信通道，以及使用安全的 API 网关来管理和验证对机器学习模型的 API 调用。在机器学习环境中，数据和模型安全性是首要的；因此，这些措施对于机器

学习环境中非常关键，尤其是在通过网络（包括互联网或云环境）访问或管理模型时。

- 评估标准：在整个 AI 系统中实现 100%符合网络安全策略。
- RACI 模型：网络安全团队（执行）、首席信息安全官（负责），IT 运营（咨询），所有网络用户（告知）。
- 高层实施策略：实施全面的网络安全措施，如防火墙和网络分段。
- 持续监控和报告：定期监控网络入口和出口流量以及基础设施内的流量，并确保控制合规。定期进行渗透测试和漏洞评估是一种积极主动的方法，有助于在攻击者利用这些潜在弱点之前发现网络安全控制中的漏洞，从而确保机器学习模型及其数据得到强有力的保护。
- 访问控制映射：根据特定角色和模型要求定制网络访问控制。
- 基本准则：与网络安全相关的 CIS 控件 V8（https://www.cisecurity.org/insights/white-papers/cis-controls-v8）

2.2.3 操作系统级加固和安全配置

机器学习模型的操作系统级加固和安全配置涉及强化运行机器学习模型和应用程序的底层操作系统（OS），以降低风险和减少漏洞。这个过程对于为机器学习操作创建一个安全的环境至关重要，因为操作系统是这些应用程序的基础层。关键方面包括：

定期更新和补丁管理：使操作系统及其组件保持最新的安全补丁和更新，以防止已知的漏洞。

最小化安装：删除或禁用操作系统中对于 ML 运营不必要的服务、应用程序和功能，最小化潜在的攻击面。

配置安全设置：调整操作系统设置以增强安全性，例如启用防火墙、配置用户权限以及实施安全策略，以规定系统的访问和使用方式。

用户访问控制：实施严格的用户访问控制，确保只有授权用户才能访问机器学习系统，并应用最小权限原则，即只授予用户执行任务所必须的访问权限。

监控和审计：设置监控和审计工具来跟踪操作系统中的活动和变化，这可以帮助检测和应对安全事件。

安全通信协议：确保与机器学习系统之间的所有通信都是加密和安全的。

这些措施有助于为机器学习系统创建稳健的安全态势，保护其免受各种威胁，这些威胁可能损害机器学习模型及其数据的完整性、机密性和可用性。

- 评估标准：确保 100%的 AI 系统在加固的操作系统和安全配置下运行。
- RACI 模型：系统管理团队（执行）、首席信息安全官（负责）、安全团队（咨询）、最终用户（告知）。
- 高层实施策略：在操作系统加固和安全配置设置中应用最佳实践。
- 持续监控和报告：监控合规性和具有安全威胁的漏洞。
- 访问控制映射：限制谁可以更改系统配置。
- 基本准则：利用 NIST 800-53、CIS 和 DISA STIG 基准进行安全配置。

2.2.4 K8s 与容器安全

K8s 和机器学习的容器安全指的是一套用于保护容器化的 ML 应用程序及其部署环境的实践和工具。K8s 是一个容器编排平台，容器化技术被广泛用于部署和管理机器学习模型及工作负载。在这种情况下，具体的安全防护包括：确保在良好的风险管理框架下安全配置容器并减少漏洞，实施强大的 K8s 集群安全（包括网络策略、访问控制和 pod 安全），以及保护 K8s 环境中的通信通道。此外，包括管理容器

权限、定期扫描漏洞，以及执行治一系列安全策略，包括：在 ML 工作流中治理容器安全运行及交互，保护 ML 模型和数据免受安全威胁，如：未经授权的访问、漏洞、及容器化部署环境中的其他安全威胁等。

- 评估标准：目标是降低 AI 系统中容器环境的安全漏洞率。
- RACI 模型：容器管理团队（执行）、CTO（负责），安全团队（咨询）、DevOps 团队（咨询）、应用程序开发团队（告知）。
- 高层实施策略：确保在容器化环境中安全部署人工智能应用程序。
- 持续监控和报告：对容器编排工具进行定期安全评估。
- 访问控制映射：为容器环境定义并执行严格的访问策略。
- 基本准则：遵循 OWASP K8s TOP10、NIST SSDF、CNCF 安全报告、CIS 基准、NIST 800-190 和 NIST 800-53 最佳实践。

2.2.5 云环境安全

机器学习模型的云环境安全包括为保护基于云的基础设施中的机器学习模型及其相关数据而实施的策略和措施。这涉及保护云上的数据存储和处理、管理对云资源的访问控制、加密静态和传输中的数据以及确保部署在云平台上的机器学习模型的安全性。它还包括定期漏洞评估、遵守特定于云的安全标准，以及实施身份和访问管理的最佳实践。这种安全性对于防止未经授权的访问、数据泄露和其他网络威胁至关重要，可确保在云环境的动态和分布式特性下，机器学习模型和数据的完整性和机密性。

- 评估标准：努力在 AI 部署中 100%遵守云安全策略。
- RACI 模型：云安全团队（执行）、首席信息安全官（负责），IT 治理（咨询），云服务用户（告知）。
- 高层实施策略：为人工智能系统实施稳健的云安全措施。

- 持续监控和报告：使用特定于云的监控工具来检测威胁并发出警报。
- 访问控制映射：为基于云的 AI 应用程序自定义访问控制。
- 基本准则：采用 CSA CCM。坚持并优先考虑云原生应用程序保护平台 CNAPP。

2.3 漏洞和补丁管理

本章节中的职责和方法应纳入组织更广泛的安全和开发流程，从而确保安全有效地开发、部署和维护人工智能系统。对这些流程的定期审查和更新将有助于适应不断发展的威胁和技术，这种定期审查可以识别软件缺陷、漏洞和弱点，并且对其进行排序，处理系统故障、报错和失效，让人工智能系统免于暴露遭受攻击。以下列举的这些技术位于组织中责任的各个方面，能够确保和提高人工智能系统的安全性、可靠性和可信度.

2.3.1 机器学习代码完整性保护

机器学习代码完整性保护是指为确保机器学习应用程序中使用的源代码的安全性和完整性而采取的措施和实践。这涉及保护代码免受未经授权的修改，确保其真实性，并在整个开发和部署过程中保持质量。最佳实践包括实施版本控制系统来跟踪和管理更改，使用签名来验证代码的真实性，定期进行代码审查和审计以检测漏洞，以及使用静态和动态代码分析工具来识别潜在的安全问题。这些保护对于维护机器学习应用程序的可信度、防止恶意代码注入以及确保机器学习模型按预期运行和不受损害来说至关重要。

- 评估标准：
    - 完整性保护覆盖的 ML 代码百分比
    - 完整性检查的频率
    - 已解决的关键漏洞百分比

- 责任（RACI 模型）：人工智能开发团队（执行）、人工智能开发经理（负责）、安全团队、DevOps 团队（咨询）、合规团队（告知）。
- 高层实施策略：
  - 对机器学习模型实施自动完整性检查。
  - 将完整性检查集成到 CI/CD 流水线中。
  - 为机器学习代码建立定期漏洞评估。
  - 实施针对机器学习算法的安全编码实践。
  - 对机器学习模型中的异常行为实施运行时监控。
- 持续监测和报告：
  - 利用监控工具检测机器学习模型行为中的异常。
  - 为检测到的异常建立事件响应程序。
  - 定期报告完整性检查结果和发现的任何异常。
- 访问控制映射：
  - 根据角色和职责授予对 ML 代码仓库的访问权限。
  - 实现访问 ML 代码的最小特权原则。
  - 对敏感的机器学习代码库使用多因素认证。
- 基本准则：
  - 有关保护机器学习系统的指南，请参考 NIST AI RMF。
  - 实施 NIST 800-53 中建议的与机器学习系统相关的安全控制。
  - 遵守 CSA CCM，以考虑云特定的安全因素。

2.3.2 机器学习训练和部署代码的版本控制系统

机器学习训练和部署代码的版本控制系统是管理机器学习项目的重要工具，能使团队能够跟踪和管理用于训练和部署机器学习模型的代码库的更改。这些系统通过维护更改历史、允许轻松跟踪修改以及在需要时支持回滚到以前的版本，促进了开发人员和数据专家之间的协作。它们在处理各种版本的机器学习模型及其相关数据集方面至关重要，可确保机器学习实验和部署的一致性和可重复性。通过使用版本控制，团队可以有效地管理机器学习模型的生命周期，从开发和测试到部署和维护，同时确保整个过程中代码和模型的完整性和可追溯性。

- 评估标准：
  - 受版本控制的代码百分比
  - 提交和更新的频率
  - 遵守版本控制策略
- 责任（RACI 模型）：
  - 执行：开发团队
  - 负责：开发经理
  - 咨询：DevOps 团队、QA 团队
  - 告知：安全团队
- 高层实施策略：
  - 实施集中式版本控制系统（例如 Git）
  - 执行分支和合并策略
  - 在合并之前自动进行代码审查和检查

- 实施版本标记以进行发布管理

- 持续监测和报告：
  - 监控提交活动并识别异常模式
  - 对未经授权的更改或异常活动建立警报
  - 定期审查版本控制日志以确保合规性

- 访问控制映射：
  - 根据角色定义存储库的访问控制列表
  - 对存储库访问实施双因素身份验证
  - 定期审核存储库访问以确保合规性

- 基本准则：
  - 遵循 NIST SSDF 的安全软件开发指南。
  - 实施 NIST 800-53 中概述的与版本控制和变更管理相关的控制措施。

2.3.3 利用代码签名验证获批版本

在机器学习环境中，安全实践之一是利用代码签名验证获批版本，其中数字签名用于验证机器学习模型中使用的软件代码的真实性和完整性。此过程通常是在代码经过审查和批准部署后，将加密签名附加到代码上。签名相当于签章，可以验证代码自签名以来是否被更改或篡改。在机器学习工作流中，代码签名对于确保用于训练、测试和部署机器学习模型的代码是准确的、授权的版本并且没有被恶意修改非常重要。这种做法有助于保持对机器学习软件供应链的信任，特别是当模型分布在不同的环境中或在多个团队或组织之间共享时。

- 评估标准：

- 用批准的证书签名的代码百分比
- 遵守代码签名政策
- 代码签名检查的频率

- 责任（RACI 模型）：
  - 执行：开发团队
  - 负责：开发经理
  - 咨询：安全团队、发布管理团队
  - 告知：合规团队
- 高层实施策略：
  - 在构建过程中实现代码签名
  - 安全地管理代码签名证书
  - 部署前自动执行代码签名检查
  - 对签名代码实施时间戳，以防止重放攻击
- 持续监测和报告：
  - 监控代码签名活动和证书使用情况
  - 对未经授权的代码签名尝试实施警报
  - 定期审查代码签名日志以确保合规性
- 访问控制映射：
  - 仅限授权人员访问代码签名基础设施

- 对代码签名证书实施基于角色的访问控制
- 定期查看代码签名基础设施的访问日志

- 基本准则：
  - 参考与代码签名和完整性检查相关的 NIST 800-53 控制

2.3.4 基础设施即代码方法

机器学习模型的基础设施即代码方法（IaC）是指通过机器可读的定义文件来管理和配置计算基础设施，而不是通过物理硬件配置或交互式配置工具。这种做法能够以一致和可重复的方式自动设置、配置和管理机器学习模型所需的基础设施，如服务器、存储和网络资源。基础设施即代码的方法允许机器学习团队在各种环境（如云、本地或混合设置）中快速部署和扩展他们的模型，只需最少的人工干预。这种方法提高了基础设施的效率和可靠性。通过部署机器学习模型，确保基础设施状态可维护、版本受控并符合定义，从而改善协作并降低机器学习项目中的运营风险。

- 评估标准：
  - 按代码管理的基础设施百分比
  - 遵守基础设施即代码最佳实践
  - 基础设施更新和审查的频率

- 责任（RACI 模型）：
  - 执行：DevOps 团队
  - 负责：DevOps 经理
  - 咨询：开发团队、安全团队
  - 告知：运营团队

- 高层实施策略：
  - 利用 IaC 工具，如 Terraform 或 AWS CloudFormation
  - 对基础架构代码实施版本控制
  - 自动化基础设施变更的测试和验证
  - 实施基础设施漂移检测和修复
- 持续监测和报告：
  - 监控基础设施的变化和漂移
  - 对未经授权或意外的更改实施警报
  - 定期审查基础设施代码是否符合标准
- 访问控制映射：
  - 根据角色限制对基础架构代码仓库的访问
  - 为基础设施即代码工具实施基于角色的访问控制
  - 定期审核对基础架构代码仓库的访问权限
- 基本准则：
  - 遵循 NIST 800-53 指南，对基础设施进行安全配置和管理。
  - 为云基础设施实施 CSA CCM 中概述的安全控制。

2.4MLOps 流水线安全

MLOps 流水线安全的关键领域包括源代码漏洞扫描、测试模型对攻击的鲁棒性、验证每个阶段的流水线完整性以及监控自动化脚本。

2.4.1 源代码漏洞扫描

机器学习中的源代码漏洞扫描涉及使用自动化工具系统地检查机器学习应用程序的源代码中的安全漏洞和编码缺陷。这种做法对于早期发现和修复可能危及机器学习系统的潜在安全漏洞至关重要。扫描通常会检查常见的漏洞，如缓冲区溢出、注入缺陷、调用不安全的库以及可能导致意外行为或性能问题的编码实践。通过定期扫描机器学习代码，开发人员和数据专家可以确保代码库符合安全最佳实践和标准，从而降低漏洞利用和入侵的风险。

这种主动的方法对于维护机器学习应用程序的完整性和可信度至关重要，特别是在处理敏感数据或在关键系统中使用时。

- 评估标准：通过针对模型训练和部署中使用的源代码的百分比率和扫描的频次多少来作为评估标准。
- RACI 模型：
  - 执行：模型开发团队
  - 负责：首席信息安全官（CISO）
  - 咨询：应用安全团队
  - 告知：开发运营（DevOps）团队
- 高层实施策略：利用自动化工具实施定期的源代码漏洞扫描，并将这些工具集成到开发生命周期中。
- 持续监控和报告：为代码扫描建立持续监控系统，并实时报告结果。
- 访问控制映射：确保只有授权人员才能访问和修改源代码和扫描工具。
- 基本准则：以 NIST 800-53 标准和 CSA CCM 为指导原则。

2.4.2 测试模型对攻击的鲁棒性

在机器学习中测试模型对攻击的鲁棒性来评估机器学习模型，可以确定它们能够承受和应对各种对抗性攻击或操纵的能力。该测试对于识别机器学习模型中的潜在漏洞至关重要，这些漏洞可能被利用，从而产生不正确的结果，导致系统故障或泄露敏感信息。它通常包括用人为制作的输入（对抗性样本）探测模型，以评估其应对此类攻击的能力，分析模型在不同威胁场景下的行为，并验证其在面对意外或恶意输入时保持性能和准确性的能力。鲁棒性测试有助于确保机器学习模型的可靠性和安全性，特别是需要稳固的决策能力的应用中，如自动驾驶汽车、金融系统或医疗诊断。

- 评估标准：通过攻击测试的模型百分比和测试的频率来衡量有效性。
- RACI 模型：
  - 执行：AI/ML 测试团队
  - 负责：AI/ML 开发主管
  - 咨询：安全分析师
  - 告知：AI/ML 开发团队
- 高层实施策略：为模型开发稳健的测试框架，重点是识别和减轻潜在的攻击媒介。
- 持续监测和报告：实施持续评估模型鲁棒性的机制，并定期向利益相关者通报结果。
- 访问控制映射：适当地控制对测试框架和模型的访问。
- 基本准则：参考 NIST AI RMF、NIST AI 100-2 E2023《对抗性机器学习：攻击与缓解的分类和术语》、以及其他相关标准中的最佳实践。

2.4.3 验证每个阶段的流水线完整性

在机器学习的每个阶段验证流水线完整性是指确保机器学习流水线的每个阶段（从数据收集和预处理到模型训练、评估和部署）正确安全运行。这涉及在每个阶段进行彻底的检查和验证，以防止数据损坏、未经授权的访问和其他可能损害流水线性能和模型准确性的漏洞。此类验证包括验证数据质量和一致性，确保安全的数据处理实践，评估模型训练过程的可靠性和可重复性，以及确认部署机制是安全的并按预期运行。这种全面的验证方法对于维护机器学习流水线的整体完整性和有效性至关重要，特别是在复杂或高风险的环境中，机器学习模型的准确性和可靠性就显得尤为重要。

- 评估标准：重点是仔细监测 MLOps 流水线的完整性。通过对每个阶段的检查采用百分比评估，并验证评估过程的深度和完整度，可以确保流水线的每个阶段都能够按预期运行，严格遵守既定的标准和最佳实践。
- RACI 模型：
  - 执行：DevOps 团队负责验证每个流水线阶段的日常任务。他们是验证过程的主要执行者，确保 MLOps 流水线的每个阶段都经过彻底检查和验证。
  - 负责：工程主管对 MLOps 流水线的整体完整性负有最终责任。该角色涉及监督验证过程，并确保流水线符合必要的标准和要求。
  - 咨询：质量保证（QA）团队是咨询性的，为验证过程提供专家建议和意见。他们的参与对于确定验证标准和审查验证结果至关重要。
  - 告知：项目经理随时了解流水线验证过程的状态和结果。这确保他们了解可能影响项目时间表或交付成果的任何潜在问题或变化。
- 高层实施策略：系统验证 MLOps 流水线每个阶段的完整性。该战略包括为数据和流程完整性建立明确的程序和标准。它涉及对每个流水线阶段定义

特定的验证测试和检查，确保从数据采集到模型部署的每个环节都能正确安全地运行。

- 持续监测和报告：验证流水线完整性的重要组成部分正在实施一个持续验证和实时报告系统。该系统应在出现任何差异或异常时进行检测，以便立即采取行动纠正问题。持续监控确保流水线始终保持安全高效。

- 访问控制映射：严格的访问控制对于维护每个文件的完整性至关重要MLOps 流水线的阶段。这涉及定义和执行谁有权访问流水线的各个部分，在什么条件下，以及以什么级别的权限。此类控制对于防止可能损害流水线完整性的未经授权的访问或修改至关重要。

- 基本准则：为了确保遵循最佳实践，重要的是要使流水线验证过程符合既定的行业标准和指南，如 NIST《安全软件开发框架》（SSDF）中描述的标准和指南。遵照此框架体系可以为安全性和效率提供基准，指导稳健可靠地进行 MLOps 流水线的开发和维护。

2.4.4 监控自动化脚本

这项任务涉及对所有脚本的密切监控，利用这些脚本，机器学习生命周期的各个阶段（从数据预处理到模型部署和管理）都可以自动化处理。

- 评估标准：监控自动化脚本的有效性由两个主要指标来量化：持续监控下的自动化脚本的百分比和监控活动的频率。此评估有助于确保所有脚本正确有效地运行，并及时发现和解决潜在问题。

- RACI 模型：

  - 执行：IT 运营团队主要负责 MLOps 流水线内自动化脚本的日常监控。他们的职责包括监督脚本的执行，确保其性能和安全性，并识别操作问题。

- 负责：首席技术官（CTO）对自动化脚本的管理和安全负全部责任。CTO 确保监控策略得到有效实施，并与组织的技术目标保持一致。
- 咨询：DevOps 团队提供关键的意见和专业知识，特别是在脚本部署和运营效率方面，对于加强流水线内使用的监测流程和工具至关重要。
- 告知：MLOps 流水线中的所有利益相关者，包括数据专家、机器学习工程师和项目经理，都会随时了解自动化脚本的状态和性能。这确保了流水线所有阶段的连贯性和透明性。为了保证整个 MLOps 流水线的连贯性和透明性，所有利益相关者（包括数据专家、机器学习工程师和项目经理）都必须充分了解自动化脚本的状态和性能。这种做法不仅促进了流水线所有阶段的统一方法，而且确保了决策是基于最新和准确的信息，提高了人工智能部署的整体安全性和效率。

- 高层实施策略：为所有自动化脚本实施全面的监测系统是至关重要的。该系统应跟踪脚本的性能和效率，以确保其符合既定的标准和实践。它应该无缝集成到 MLOps 流水线中，提供对自动化脚本行为和输出的实时状态与分析。
- 持续监测和报告：及时发现并解决自动化脚本的问题的关键是持续监测。监控系统应能够生成实时警报和报告，及时提供有关脚本性能、错误或安全问题的信息。这种持续的反馈对于维护 MLOps 流水线的操作完整性至关重要。
- 访问控制映射：严格的访问控制对于保护自动化是必要的脚本和整个流水线。这涉及定义谁可以访问、修改或执行脚本。访问应基于特定角色的要求，确保只有授权人员才能进行更改，从而降低未经授权或有害修改的风险。
- 基本准则：采用 CSA CCM 等既定框架中的最佳实践以及 NIST 提供的相关指导。

2.5AI 模型治理

AI 模型治理包括几个关键领域：模型风险评估、业务审批程序、模型监控要求和新的模型验证流程。

2.5.1 模型风险评估

机器学习中的模型风险评估涉及系统性的评估机器学习模型在部署和使用中的潜在风险。该评估旨在识别和量化模型不准确、偏差或失效可能带来的不利影响。主要关注领域包括：评估模型在不同数据集和场景中的准确性和泛化能力，评估有偏见或不公平结果的可能性，以及了解模型在极端情况或对抗条件下的行为。模型风险评估还考虑了模型失效的后果，特别是在医疗保健、金融或公共安全等关键应用中。这一过程对于识别和减轻风险至关重要，以确保通过清楚地了解机器学习模型的局限性和潜在影响，负责任地安全部署机器学习模型。

- 评估标准：评估正在进行风险评估的模型的百分比以及这些评估的全面性。
- RACI 模型：
  - 执行：风险管理团队、数据治理委员会
  - 负责：首席风险官（CRO）
  - 咨询：人工智能伦理委员会、法律顾问
  - 告知：数据科学团队
- 高层实施策略：开发一个框架来评估与 AI 模型相关的风险，包括偏见、公平性和数据隐私。
- 持续监控和报告：作为软件供应链的一部分，利用工具实施持续的风险监控，并构建报告风险的协议，。

- 访问控制映射：确保对风险评估工具和数据的访问受到严格控制和监控。
- 基本准则：与 NIST AI RMF 和 NIST 800-53 在风险管理实践方面保持一致。

2.5.2 业务审批程序

这包括组织批准机器学习模型部署到生产中所遵循的正式流程和协议。这些程序确保任何机器学习模型都与业务目标保持一致，符合监管和伦理标准，并达到所需的性能基准。通常，这涉及一个多步骤的审查过程，涉及各种利益相关者，包括数据科学家、业务分析师、风险管理团队，有时还有法律和合规部门；过程中将评估模型的有效性、可靠性和潜在的业务影响。经常评估的关键方面包括模型的预测准确性、验证数据集的性能、潜在的偏见或伦理问题，以及数据隐私法的合规性。这些程序的目的是建立一种可控且知情的方法来部署机器学习模型，最大限度地降低业务风险，并确保负责任地使用人工智能技术。比如，OpenAI 的准备框架（Preparedness Framework）。

- 评估标准：跟踪批准部署的 AI 模型的百分比以及对批准指南的遵守情况。
- RACI 模型：
  - 执行：项目管理团队
  - 负责：首席人工智能官
  - 咨询：业务部门负责人
  - 告知：所有 AI 利益相关者
- 高层实施策略：建立明确的 AI 模型审批程序，让利益相关者参与决策过程。
- 持续监控和报告：维护审批流程和决策的记录，并建立定期审查机制。
- 访问控制映射：控制对审批文件和决策工具的访问。
- 基本准则：遵循最佳实践，如 NIST SSDF 和 CSA CCM。

2.5.3 模型监控要求

模型监控要求指的是在部署到生产环境后跟踪和评估机器学习模型性能的持续过程。这种监控对于确保模型在不同条件下能否长时间如预期运行至关重要。模型监控的关键方面包括：跟踪模型的预测准确性，检测模型输入或输出中的任何漂移（数据漂移或概念漂移），监控预测中的任何偏差或不公平迹象，以及关注机器学习系统的整体健康和性能。此外，当模型性能发生重大变化或异常时，监控应提醒各利益相关方。持续的监控有助于及时识别和解决模型退化、底层数据模式的变化或新出现的偏差等问题，确保机器学习模型保持有效、公平，并与预期目的保持一致。

- 评估标准：根据监控活动的频率和深度评估模型。
- RACI 模型：
  - 执行：人工智能运营团队
  - 负责：人工智能运营主管
  - 咨询：质量保证团队
  - 告知：业务分析师
- 高层实施策略：实施一个全面的模型监控系统，跟踪性能、准确性和合规性。最近的机器学习监控技术包括数据和概念漂移检测、模型性能跟踪、特征属性分析、偏差检测和实时警报等功能。以下一些示例说明了这些功能，比如：Sage Maker 模型监视器（SageMaker Model Monitor）可以捕获实时推理数据并将其与基线对比；谷歌云人工智能平台预测监控（Google Cloud AI Platform Prediction Monitoring 可以提供模型预测和数据漂移的深度分析；其他各类人工智能监控和可解释性平台，可以支持团队监控、解释和分析生产环境中的机器学习模型，从而检测数据漂移、模型漂移和偏差等问题。

- 持续监测和报告：建立持续数据收集和分析系统，并对性能下降或异常发出警报。
- 访问控制映射：限制对监控工具和敏感数据的访问。
- 基本准则：利用 NIST 800-53 的监测协议指南。

2.5.4 新模型验证过程

新模型验证过程涉及在将新的机器学习模型部署到生产环境之前，对其进行严格测试和验证的系统程序。这些过程旨在确保模型满足预定义的准确性、可靠性和公平性标准，并且确保没有导致不正确或不公平结果的缺陷或偏差。验证通常包括对不同数据集进行广泛测试，以评估模型的性能和泛化能力，检查潜在的偏见或伦理问题，并评估模型在对抗性攻击或数据异常等情况下的鲁棒性。此外，验证过程通常涉及对模型文档和开发实践进行审查，以确保符合行业标准和最佳实践。这些验证过程旨在建立对新模型的能力和部署准备的信心，确保它们按预期运行，并按照业务目标和道德准则提供价值。

- 评估标准：衡量验证过程的全面性，即：完成全面验证的模型百分比。
- RACI 模型：
  - 执行：人工智能开发团队
  - 负责：首席数据官或首席技术官
  - 咨询：IT 安全团队
  - 告知：高级管理层
- 高层实施策略：制定严格的流程来验证新模型，包括测试准确性、偏差和安全漏洞。
- 持续监测和报告：为部署后新模型的持续评估建立协议。

- 访问控制映射：确保严格控制谁可以批准和部署新模型。
- 基本准则：与 NIST AI RMF 保持一致，以验证最佳实践。

2.6 安全模型部署

这涉及一系列实践，以确保部署过程安全、可控并符合组织标准。关键领域包括部署授权程序、灰度发布（canary releases）、蓝绿部署、回滚功能和模型退役。

2.6.1 灰度发布

灰度发布是一种用于将新机器学习模型引入生产环境时最小化风险的技术，即在广泛部署新 ML 模型之前，先逐步向一小部分用户推出。这种技术能够让团队在实际数据和用户交互的真实环境中测试和监控模型的性能，但规模有限。

- 评估标准：通过早期部署的成功率和检测到的问题来监控灰度发布的有效性性。
- RACI 模型：
  - 执行：DevOps 团队
  - 负责：人工智能运营主管或首席技术官
  - 咨询：质量保证（QA）团队
  - 告知：产品管理团队
- 高层实施策略：将灰度发布作为部署过程中的一个步骤，在实际环境中逐步测试模型。
- 持续监控和报告：对灰度发布进行实时监控，以快速识别和解决问题。
- 访问控制映射：确保只有指定的团队成员可以启动和监控灰度发布。

- 基本准则：遵循 NIST SSDF 和 NIST 800-53 的部署指南。

2.6.2 蓝绿部署

蓝绿部署是软件部署（包括机器学习模型的部署）中的一种策略，通过运行两个相同的生产环境（称为“蓝色”和“绿色”）来减少停机时间和风险。这种方法在机器学习中特别有用，因为部署新模型会对应用程序性能和用户体验产生重大影响。

- 评估标准：根据过渡过程的平顺程度和部署期间的停机时间来评估部署策略。
- RACI 模型：
  - 执行：IT 运营团队
  - 负责：首席技术官（CTO）
  - 咨询：DevOps 团队
  - 告知：最终用户
- 高层实施策略：采用蓝绿部署策略，以减少与部署新模型相关的停机时间和风险。
- 持续监控和报告：持续监控“蓝”、“绿”两套环境的性能，以及相关问题的解决方案。
- 访问控制映射：管理两个环境的访问控制，确保安全性和完整性。
- 基本准则：利用 NIST 相关指南中的最佳实践。

2.6.3 回滚功能

在机器学习模型的环境中，回滚功能是指当新部署的模型表现出意外行为、性能不佳或导致其他问题时，在生产环境中回退到 ML 模型的先前版本或检查点的过

程。这是部署策略的一个关键方面，即使新模型未能达到预期，也能确保系统的稳定性和性能。

- 评估标准：在需要时，通过回滚的速度和成功率来衡量有效性。
- RACI 模型：
  - 执行：部署团队（DevOps 团队）
  - 负责：人工智能运营主管或首席技术官
  - 咨询：IT 支持团队
  - 告知：商业利益相关者
- 高层实施策略：确保部署过程包括高效的回滚功能，以便在必要时恢复到以前的版本。
- 持续监控和报告：监控部署以快速检测需要回滚的问题。
- 访问控制映射：控制对回滚工具和过程的访问。
- 基本准则：与 CSA CCM 和其他安全框架保持一致。

2.6.4 模型退役

模型退役是指从活动生产环境中安全、系统地删除机器学习模型的过程。当模型过时，被更高级的版本取代，或者不再满足不断发展的业务需求或合规标准时，模型退役过程变得至关重要。退役是机器学习模型生命周期管理中的关键步骤，以确保过时的模型不会危及系统的完整性或效率。

- 评估标准：通过正确处理退役模型的百分比和遵守退役协议的情况来评估这一过程。
- RACI 模型：

- 执行：人工智能维护团队（DevOps）
- 负责：数据治理官或首席技术官
- 咨询：法律与合规团队
- 告知：人工智能开发团队

- 高层实施策略：制定明确的程序，安全地退役过时或冗余的 AI 模型，并确保数据得到安全处理。
- 持续监控和报告：实施监控以确保正确遵循退役流程。
- 访问控制映射：限制对退役工具和数据的访问。
- 基本准则：遵循 NIST AI RMF 和其他相关指南的退役实践。

三、 漏洞管理

人工智能漏洞管理是保护人工智能和机器学习系统的关键组成部分，可以确保系统的安全性、功能性和合规性。本节讨论此类别中的关键项目。

3.1AI/ML 资产清单

AI/ML 资产清单系统地记录并更新 AI/ML 环境中的所有资产。这不仅包括模型和数据集，还包括在创建、训练和部署软件供应链的 AI/ML 组件时，软件供应链中涉及的 API、算法、库以及任何支持软件或工具。该清单清晰地展示了正在使用的资源，这对于识别潜在漏洞和有效管理风险至关重要。根据机器学习系统的不同，使用的资产清单可以是模型卡、数据卡和模型注册表等格式。了解存在哪些资产以及它们是如何相互关联的，对于识别潜在的漏洞至关重要。

- 评估标准：AI/ML 资产清单的有效性是通过其全面性和更新的规律性来衡量的。全面的清单涵盖了 AI/ML 环境的各个方面，没有遗漏任何组件。更新的频率同样重要，这将确保清单反映 AI/ML 生态系统的当前状态，包括任何新开发的模型、新获取的数据集或软件环境的变化。

- RACI 模型：

  - 执行：IT 运营团队（或 DevOps 团队）负责清单的日常管理和更新。

  - 负责：首席信息官（CIO）或首席技术官（CTO）监督流程，确保清单得到准确维护，并在漏洞管理中得到有效使用。

  - 咨询：AI/ML 开发团队提供必要的见解和信息，以确保清单保持最新且具有相关性。

  - 告知：高级管理层了解 AI/ML 资产的状态和健康状况，从而能够在更高级别做出明智的决策。

- 高层实施策略：应制定计划，定期审查并更新 AI/ML 资产清单。这可以自动化工具来实现，即：跟踪 AI/ML 环境变化并提醒负责团队更新清单。该过程还应包括定期审计，以确保准确性和完整性。

- 持续监控和报告：实施实时监控系统有助于快速识别 AI/ML 资产的变化。这可以包括新模型部署、现有模型更新、数据集变更或软件环境更改。持续监测有助于维护最新的清单，这对有效的漏洞管理至关重要。

- 访问控制映射：限制对 AI/ML 资产清单的访问对于维护其完整性和保密性至关重要。访问权限应仅限于授权人员，根据他们与清单交互的需要，不同角色有不同级别的访问权限。

- 基本准则：遵守 NIST AI RMF（风险管理框架）等框架，确保 AI/ML 资产清单的管理符合行业最佳实践和监管要求。这些框架提供了有效编目和管理 AI/ML 资产的指导方针，有助于制定强有力的漏洞管理策略。

3.2 持续漏洞扫描

持续漏洞扫描是指对所有 AI/ML 资产进行持续的检查，以识别安全漏洞。这包括扫描模型、数据集、相关基础设施（过时的库或不安全的 API）以及 AI/ML 环境中的任何其他组件。扫描的目的是发现可能被利用的漏洞，从而预先解决潜在的安全问题。

- 评估标准：此扫描过程的有效性通过两个主要指标来衡量：被扫描的 AI/ML 资产的百分比和扫描的频率。理想的漏洞扫描程序应确保没有任何组件被遗漏，并能够定期进行扫描，以捕捉由于环境更新或变化而可能出现的新漏洞。
- RACI 模型：
  - 执行：安全运营团队执行扫描过程，利用各类工具和技术进行全面评估。
  - 负责：首席信息安全官（CISO）监督整个过程，确保有效进行扫描并及时解决漏洞。
  - 咨询：AI/ML 团队提供有关 AI/ML 资产的具体要求和配置的详细信息，帮助实现更有针对性的扫描。
  - 告知：IT 管理部门会收到最新的扫描结果以及任何可能影响 IT 基础设施的关键漏洞信息。
- 高层实施策略：实施自动化扫描工具可执行高效和有效的漏洞扫描。这些工具应配置为定期扫描所有 AI/ML 资产，并在新威胁出现时进行更新。安排定期评估可确保 AI/ML 环境长期推移保持安全。

- 持续监测和报告：建立警报系统至关重要，以便及时将已发现的新漏洞通知相关团队。该系统将检测到的弱点通知相关团队，以便快速响应并修复。
- 访问控制映射：访问漏洞扫描工具和结果应严格控制。只有授权人员才能进行扫描并访问详细结果，以确保扫描过程中暴露的敏感信息安全。
- 基本准则：遵守既定的安全标准，如 NIST 800-53 确保漏洞扫描过程符合行业最佳实践。这些标准提供了有效识别和解决漏洞的指导方针，增强了 AI/ML 系统的整体安全态势。

3.3 基于风险的优先级排序

基于风险的优先级排序是根据漏洞对 AI/ML 资产的潜在影响和被利用的可能性，对其进行评估和排序。这一过程有助于组织将资源和精力集中在优先缓解最关键的漏洞上，从而有效地降低其 AI/ML 系统的整体风险。

- 评估标准：这种方法的有效性是通过已成功解决的高风险漏洞与已识别漏洞总数的比例来衡量的。高百分比表明对最严重的风险的优先级排序并修复是有效的。
- RACI 模型：
  - 执行：风险管理团队的任务是根据风险水平评估漏洞并确定其优先级。
  - 负责：首席信息安全官（CISO）监督该流程，确保最关键的漏洞被发现并及时得以解决。
  - 咨询：合规团队提供意见，特别是关于漏洞的监管和合规方面。
  - 告知：AI/ML 开发团队了解影响其资产的漏洞的优先级和状态。

- 高层实时策略：制定一个全面的风险评估框架至关重要。该框架应包括评估漏洞严重性的标准，例如对机密性、完整性、可用性的潜在影响以及被利用的可能性。

- 持续监测和报告：实施一个持续监测漏洞及其风险水平的系统至关重要。定期更新并报告漏洞的风险状态，确保所有利益相关者了解当前的威胁形势和修复工作的进展。

- 访问控制映射：应严格控制对风险评估工具和漏洞数据的访问。只有授权人员才能对漏洞进行评估、分类并进行优先级排序，这样可以保持流程的完整性。

- 基本准则：遵守 NIST《人工智能风险管理框架》（RMF）等指导方针，可确保该过程与行业最佳实践保持一致，并为管理 AI/ML 系统风险提供结构化方法。

基于风险的优先级排序是人工智能漏洞管理的重要组成部分，使组织能够有效地分配资源，以减轻其 AI/ML 系统中最紧迫的安全风险。

3.4 修复跟踪

修复跟踪过程涉及对流程的持续监控和管理，以解决并修复 AI/ML 系统中已识别的漏洞。它包括跟踪漏洞缓解措施的实施情况，确保漏洞被及时解决。

- 评估标准：修复跟踪的有效性基于两个主要指标进行评估：修复漏洞所需的时间和已解决问题的百分比。修复时间越短，已解决漏洞的百分比越高，表明修复工作高效且成功。根据前文定义的基于风险的漏洞优先级排序，修复时间必须依据服务级别协议（SLA）进行跟踪。

- RACI 模型：

- 执行：IT 运营团队执行修复漏洞所需的操作。
- 负责：首席信息安全官（CISO）对修复的有效性负全面责任，确保漏洞得到及时解决。
- 咨询：AI/ML 开发团队提供意见和帮助，帮助理解与 AI/ML 资产相关的漏洞修复的具体要求。
- 告知：随时告知高级领导层漏洞修复工作的状态及其对组织的潜在影响。

- 高层实施策略：实施可靠的跟踪系统对于进行高效的修复跟踪十分关键。这些系统应记录已识别漏洞的详细信息、采取的修复措施、责任方和解决问题的时间表。
- 持续监测和报告：维护修复活动的详细记录，包括进度更新和完成状态。持续监控确保漏洞得到积极跟踪和管理，直到成功解决。
- 访问控制映射：与修复活动相关的文档的访问权限应得到严格控制，防止未经授权的访问或篡改。这保护了修复过程的完整性。
- 基本准则：参考 NIST 等既定的网络安全标准 800-53 确保修复跟踪过程与行业最佳实践保持一致，并提供了一种结构化的方法来管理和记录漏洞修复。

3.5 异常处理

异常处理是有效管理与既定安全协议和程序发生偏差或异常的情况的过程。这些例外情况可能是由于独特的情况、运营需求、遗留的旧系统或其他需要偏离标准安全实践的因素造成的。

- 评估标准：根据处理的异常数量和解决方案的整体有效性来评估异常处理的有效性。管理良好的异常处理过程应尽量减少异常的数量，并确保在发生异常时，使用异常处理来妥善解决安全问题。
- RACI 模型：
  - 执行：安全团队管理并解决出现的安全异常。
  - 负责：首席信息安全官（CISO）全面负责对异常处理流程的有效性，确保异常管理符合安全政策和法规。
  - 咨询：法律和合规团队提供指导和建议，以确保在法律和监管要求的范围内管理例外情况。
  - 告知：管理层了解异常情况及其解决方案，确保透明度并与整体运营保持一致。
- 高层实施策略：建立明确且有文件记录的安全异常处理协议是至关重要的。这些协议应定义发现、评估和解决异常的过程，同时确保安全仍然是首要任务。
- 持续监控和报告：异常处理应记录在案，包括异常情况、为解决异常而采取的行动以及采取的任何修复措施。持续的监控和报告有助于确保长期有效地管理异常。
- 访问控制映射：访问异常处理过程和相关文档应仅限于授权人员。这确保了异常的处理保持安全，并符合既定的协议。
- 基本准则：确保与云安全联盟（CSA）发布的《云控制矩阵》（CCM）等行业标准保持一致，有助于建立处理异常的最佳实践，并确保按照公认的指导方针进行管理。

异常处理是人工智能漏洞管理的关键组成部分，使组织能够有效地应对独特的情况，同时保持整体安全和合规标准。

3.6 报告指标

报告指标是指用于评估和量化人工智能漏洞管理工作有效性的具体衡量标准和关键绩效指标（KPI）。这些指标为 AI/ML 系统内的安全状态提供了有价值的参考。

- 评估标准：报告的准确性和及时性对于评估报告指标的有效性至关重要。及时准确的报告确保决策者可以获得可靠的信息，以便做出明智的安全决策。
- RACI 模型：
  - 执行：报告团队负责收集、分析和呈现漏洞管理指标。
  - 负责：首席信息安全官（CISO）负责报告流程的整体有效性，并确保指标与安全目标保持一致。
  - 咨询：AI/ML 和 IT 部门提供输入和背景情况，以确保指标准确反映 AI/ML 环境的安全态势。
  - 告知：高级管理层了解报告指标得出的结果和见解，从而做出战略决策。
- 高层实施策略：制定标准化的报告程序至关重要。这些程序应定义如何收集、分析和呈现指标，以确保一致性和准确性。
- 持续监测和报告：定期更新和分发报告对于让所有利益相关者了解人工智能漏洞管理的现状十分关键。持续报告确保了安全问题得到及时识别和解决。

- 访问控制映射：应控制对报告工具和数据的访问，以防止未经授权的访问或篡改指标。限制访问可以保障报告过程的完整性。
- 基本准则：遵循公认框架的最佳实践，如美国国家标准与技术研究院（NIST）发布的《安全软件开发框架》（SSDF）和云安全联盟（CSA）发布的《云控制矩阵》（CCM），是有助于建立有效开发和管理报告指标的行业标准指南。

结论

本报告探讨了组织在开发和部署人工智能和机器学习系统时必须承担的核心安全责任。通过关注数据安全、模型安全和漏洞管理，我们描绘了一个全面的框架，以确保人工智能系统在其整个生命周期内的安全性、隐私性和完整性。

在数据安全和隐私领域，本报告强调了数据真实性、匿名化、假名化、数据最小化、访问控制以及安全存储和传输的重要性。这些措施对于保护敏感信息和遵守数据保护法规至关重要。

关于模型安全，本报告讨论了访问控制、安全运行环境、漏洞和补丁管理、MLOps 流水线安全、AI 模型治理和安全模型部署的重要性。通过实施健全的安全控制和治理流程，组织可以降低与 AI 模型相关的风险，并确保其可靠和可信地运行。

漏洞管理是人工智能安全的另一个关键方面。本报告强调了维护 AI/ML 资产清单、进行持续漏洞扫描、确定风险优先级、修复跟踪工作、处理异常和建立报告指标的必要性。这些做法能够使组织主动识别和解决漏洞，最大限度地减少安全漏洞问题的可能性，并确保人工智能系统的持续安全。

在整个报告中，我们从以下方面分析了每一项责任：可量化的评估标准，角色定义的 RACI 模型，高层实施策略，持续监控和报告机制，访问控制映射，以及根据行业最佳实践和标准（如 NIST AI RMF、NIST SSDF、NIST 800-53、CSA CCM 等）制定的基本准则。

通过采用本报告中描述的建议和最佳实践，组织可以为安全和负责任的人工智能开发和部署奠定坚实的基础。然而，我们必须认识到，人工智能安全是一个持续的过程，随着技术和威胁的发展，需要不断的监控、适应和改进。

当组织面对人工智能落地带来的复杂性时，在包括管理层、技术团队、治理机构和最终用户在内的所有利益相关者之间塑造安全协作的文化至关重要。通过共同

努力并遵守相关原则和实践，组织可以释放人工智能的变革潜力，同时确保所有利益相关者的安全、隐私和信任。

缩略语

AI	人工智能
AI RMF	人工智能风险管理框架
AIMS	AI管理系统
AIOps	用于IT运营的人工智能
API	应用程序编程接口
AWS	亚马逊网络服务
CAIO	首席人工智能官
CCM	云控制矩阵
CDO	首席数据官
CEO	首席执行官
CFO	首席财务官
CI/CD	持续集成/持续部署
CIO	首席信息官
CIS	互联网安全中心
CISO	首席信息安全官
CNAPP	云原生应用防护平台
COO	首席运营官
CPO	首席隐私官
CSA	云安全联盟
CTO	首席技术官
CVE	常见漏洞和暴露
CVSS	常见漏洞评分系统
DASD	动态应用程序安全测试
DataOps	数据操作数据操作
DDos	分布式拒绝服务

DevSecOps	开发安全与运营
DISA	国防信息系统局
ENISA	欧盟网络安全局
GDPR	通用数据保护条例
HSM	硬件安全模块
IaaS	基础设施即服务
IAM	身份和访问管理
IDPS	检测和预防系统
IDS	侵入检测系统
IEC	国际电工委员会
IPS	入侵防御系统
ISM	信息安全经理
ISMS	信息安全管理体系
ISO	国际标准化组织ISS 信息系统安全
ISSO	信息系统安全官
K8s	Kubernetes开源容器编排系统
KPI	关键绩效指标
LLM	大型语言模型
MFA	多因素认证
ATT&CK	对抗策略、技术和常识
ML	机器学习
MLOps	机器学习操作
NIST	国家标准与技术研究所
OS	操作系统
OWASP	开放网站应用程序安全项目
PaaS	平台即服务
PIMP	隐私信息管理系统

PoLP	最小特权原则
QA	质量保证
RACI	执行、负责、咨询、告知
RBAC	基于角色的的访问控制
SaaS	软件即服务
SASTS	静态应用程序安全测试
SDLC	软件开发生命周期
SLA	服务水平协议
SSDF	安全软件开发框架
STIGs	安全技术实施指南
TEET	可信执行环境
TLST	传输层安全
VPN	虚拟专用网络
WAF	Web应用程序防火墙

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