

# AI模型风险管理框架



AI Technology and Risk  
Working Group

**CSA GCR** cloud security  
GREATER CHINA REGION alliance®

**CSA** cloud security  
alliance®

@2025 云安全联盟大中华区 – 保留所有权利。你可以在你的电脑上下载、储存、展示、查看及打印，或者访问云安全联盟大中华区官网 (<https://www.c-csa.cn>)。须遵守以下：(a) 本文只可作个人、信息获取、非商业用途；(b) 本文内容不得篡改；(c) 本文不得转发；(d) 该商标、版权或其他声明不得删除。在遵循中华人民共和国著作权法相关条款情况下合理使用本文内容，使用时请注明引用于云安全联盟大中华区。



## 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)



企业合作微信号:csagcr



认证培训微信号:CSAlynn



邮箱:info@c-csa.cn



## 致谢

《AI 模型风险管理框架》由 CSA 云安全联盟专家编写，并由 CSA 大中华区 AI 安全工作组完成翻译并审校。（以下排名不分先后）：

### 中文版翻译专家组

#### 翻译组成员：

郭建领 卞超轶 黄鹏华 王绪国 卜宋博 张 淼 潘季明 张 亮

#### 审校组成员：

高健凯 卜宋博

### 英文版编写专家组

#### 主要作者：

Maria Schwenger Vani Mittal

#### 其他贡献者：

Eric Tierling Hadir Labib Michael Roza Renata Budko

#### 审稿人：

Candy Alexander Daniel C

Elier Cruz

Harie Srinivasa Bangalore Ram Thilak Karan Goenka

Kenneth Thomas Moras Namal Kulathunga Nicolas Ray

Otto Sulin Rohit Valia Sanitra Angram

Tom Bowyer

Vaibhav Malik

Yuvaraj Madheswaran

**联席组长：**

Chris Kirschke

Mark Yanalitis

**CSA全球工作人员：**

Josh Buker

Marina Bregkou

Stephen Smith

# 目录

致谢 .....	4
前言 .....	8
目标读者 .....	9
范围 .....	10
引言 .....	10
一、四大支柱：模型卡片，数据手册，风险卡片，场景规划 .....	15
二、全面框架的好处 .....	17
1. 增强透明度，可解释性和问责制 .....	17
2. 主动风险评估和场景分析 .....	18
3. 制定风险缓解策略 .....	18
4. 明智决策与模型治理 .....	18
5. 健壮模型验证 .....	18
6. 建立信任并增强模型采纳 .....	19
7. 持续监控和改进 .....	19
8. 积极社会与伦理影响 .....	19
9. 强有力的治理和监督 .....	19
三、关键组成部分 .....	20
1. 模型卡片：理解模型 .....	20
2. 数据手册：检查训练数据 .....	24
3. 风险卡片：识别潜在问题 .....	27

4. 场景规划：“假设”方法 .....	31
四、总体技术：一种整合方法 .....	38
1. 利用模型卡信息创建风险卡 .....	38
2. 使用数据手册加强模型理解 .....	38
3. 使用风险卡指导场景规划 .....	39
4. 场景规划对风险管理和开发的反馈 .....	43
5. AI MRM 在行动 .....	45
五、结论与展望 .....	50
参考文献 .....	51
附录1:人工智能框架、法规和指南 .....	53

# 前言

先进的机器学习（ML）模型的广泛应用在诸如预测性维护、欺诈检测、个性化医疗、自动驾驶汽车和智能供应链管理等领域带来了激动人心的机遇。机器学习模型有助于推动创新性和效率的提高，但其广泛应用也带来了内在的风险，尤其是源自模型自身的风险。如果这些风险得不到有效缓解，可能导致巨大的经济损失、监管问题以及声誉损害。为了解决这些问题，我们需要一种积极的风险管理方法。

模型风险管理（Model Risk Management, MRM）是推动组织在开发、部署在使用人工智能（AI）及机器学习模型时建立责任和信任文化的关键因素，它能够帮助组织最大程度发挥这些技术潜力的同时，最小化风险。

本报告探讨了模型风险管理在AI模型应用任开发、部署和使用方面的重要性，适用于对该主题有兴趣的读者群体，包括直接参与AI开发的从业者以及专注于AI治理的业务和合规监管机构。

本报告强调了与AI模型相关的内在风险，如数据偏见、事实性错误或信息不相关（通常被称为“幻觉”或“虚构”）、以及潜在的滥用行为。同时，提出了一个全面的MRM框架的需求，该框架基于四个相互关联的支柱：模型卡片（Model Cards）、数据手册（Data Sheets）、风险卡片（Risk Cards）和场景规划（Scenario Planning），上述支柱协同工作，通过持续的反馈循环来识别和减轻风险，并改进模型开发与风险管理。具体而言，模型卡片和数据手册为风险评估提供信息，风险卡片则指导场景规划，场景规划进一步优化风险管理和模型开发。

通过实施这一框架，组织可以确保ML模型的安全和有效使用，并带来以下关键优势：

- 提高透明度和可解释性
- 主动应对风险和“内生安全”
- 做出明智的决策
- 与利益相关者和监管机构建立信任

本报告强调了模型风险管理的重要性，以便在最大限度发挥AI和ML潜力的同时，最小化风险。

## 目标读者

AI模型风险管理（MRM）面向开发和部署机器学习模型且对AI风险有共同兴趣的广泛读者群体。它旨在弥合技术和非技术利益相关者之间的鸿沟，既服务于直接参与AI开发技术工作的人员，也关注AI治理和监管的相关人员。

该目标读者可以分为以下两个主要群体：

### 1. AI 模型开发与实施的从业者

- **机器学习工程师和数据科学家：**这一群体将从模型卡片（Model Cards）和数据手册（Data Sheets）的详细解释中获益，了解这些组件如何促进模型的理解和开发。掌握这些要素使他们能够构建更透明且负责任的模型。

- **AI 开发人员和项目经理：**这一群体将发现有助于预见 AI 模型生命周期中潜在问题的工具，从概念设计到实施，确保模型的负责任部署。

### 2. AI 治理与监管的利益相关者

- **风险管理专业人士、合规官员和审计员：**这一群体将对 MRM 的重要性及其与常见行业框架的对齐部分特别感兴趣，这些内容对建立、执行和评估有效的治理实践至关重要。

- **业务领导者和高管：**他们将从简介和结论部分获益，这些部分强调了 MRM 在推动组织内负责任的 AI 应用中的重要性。

- **沟通与公共关系专业人士：**这一群体将从有关 AI 模型风险与效益的沟通、利益相关者参与和声誉管理的章节中受益，以及学习如何为不同受众设计有影响力的信息。

## 范围

本文探讨了模型风险管理（MRM）及其在负责任的AI开发中的重要性，深入分析了有效MRM框架的四大支柱及其如何协同合作，从而形成一个全面的MRM方法。我们讨论了这些技术如何促进透明性、问责制和负责任的AI开发。

本文强调了MRM在塑造未来伦理和负责任AI中的作用。需要注意的是，本文主要关注MRM的概念和方法论方面，而不涉及以人为中心的内容，例如角色分配、责任归属、RACI矩阵以及跨职能参与等问题，这些内容在CSA出版物《AI组织职责——核心安全职责》中已有详细阐述。

## 引言

### MRM 的必要性和重要性

当今，复杂的AI/ML模型在各行业中正以前所未有的速度被采用。一方面，对机器学习模型的日益依赖为创新和效率提升带来了巨大潜力；另一方面，它也引入了固有的风险，特别是与模型本身相关的风险——模型风险。如果不加以控制，这些风险可能会导致严重的经济损失、监管处罚以及声誉受损。训练数据中的偏见、模型输出中的事实错误（通常称为“幻觉”或“虚构”），以及潜在的滥用风险，再加上隐私问题和知识产权（IP）问题，都需要采取积极的风险管理方法。因此，AI模型风险管理（MRM）成为确保这些模型负责任和可信赖开发、部署和使用的重要学科。

MRM 这一术语在金融等行业中广泛使用，传统上是指管理与量化模型相关的风险。而在本文中，这一既有概念被应用于管理与AI模型相关的风险。

AI MRM 有助于应对与AI模型相关的复杂性、不确定性和脆弱性，增强用户、利益相关者和监管者对AI驱动决策的可靠性和公平性的信心。随着AI的不断发展并渗透到更多领域，MRM将在塑造负责任的AI部署未来中发挥越来越重要的作用，使企业和行业从中受益。

从本质上讲，模型风险源于模型本身的固有限制。以下是几种最常见的AI模型风险来源：

- **数据质量问题：**任何模型的基础是数据。不准确、不完整或片面的数据可能导致模型缺陷，进而产生不可靠的输出和错误的结论。例如，如果一个模型使用代表性不足的高风险借款人历史数据来预测贷款违约率，它可能低估未来违约的风险，从而导致经济损失。

- **模型选择、调优和设计缺陷：**为特定任务选择错误的模型架构或使用不合适的算法会显著影响模型的有效性和可靠性。例如，使用线性回归模型预测高度非线性的现象（如股票市场波动）可能会产生误导性的结果。此外，在使用开源模型时，确保模型的完整性也非常重要。最终用户应能够验证模型签名，以确保他们使用的是正确的模型，并且模型卡片准确地描述了模型的能力和局限性。

- **一流模型固有的风险：**即使是由知名供应商发布的顶尖模型，也可能因自身的不足而带来固有风险，例如幻觉、有害言论、偏见和数据泄露。这些风险可能产生深远影响，不仅影响个别组织，还可能影响整个社会。

- **实施和操作错误：**一个设计良好的模型在实施过程中可能会受到破坏。不正确的编码、控制不足或与现有系统的不当集成都可能引发模型部署错误。例如，一个信用评级模型可能在开发中是正确的，但其在贷款处理系统中的实施存在缺陷，可能导致不准确的评估和不公平的贷款拒绝。安全性也是一个关键的操作风险集，这些风险既包括应用级别和访问级别的已知漏洞，也包括生成式 AI 时代的新风险，如提示注入。AI 模型还增加了模型本身试图改变模型用户决策的风险。

- **外部因素的演变：**模型通常基于历史数据进行训练，假设基础环境具有一定的稳定性。然而，现实世界在不断变化。经济衰退、新法规或不可预见的事件可能使历史数据变得不相关，从而导致模型产生不可靠的预测。例如，一个基于以往购买习惯来预测客户流失的模型，可能在全球疫情导致消费者偏好转变时表现不佳。同样，一个基于历史数据训练的贷款违约预测模型，在遇到全球疫情、经济政策变化或贷款活动（如新贷款、再融资和条款重新谈判）意外变化时也可能表现不佳。这两个例子都说明了模型在面对环境突变时的脆弱性，强调了监控和更新模型以确保其有效性的重要性。

MRM框架是一种结构化的方法，用于识别、评估、缓解和监控与ML模型相关的风险，尤其是在决策过程中的应用。建立这一框架是一种积极的实践，它在确保ML模型带来效益的同时最大限度地减少潜在的负面影响。该框架为组织提供了一条路线图，以确保这些模型在开发、部署和使用中负责任和值得信赖。需要注意的是，具体的风险及其严重程度（风险级别）将因组织性质、行业、业务部门和模型的预期用途而有所不同。

一个精心设计的MRM框架通过建立结构化流程来识别和评估特定模型的风险，从而实现定制化。这个持续的过程基于以下几个关键组成部分：

## 1. 治理

在组织内对AI和ML模型进行有效治理至关重要，以确保这些模型得到妥善管理，并与战略目标和监管要求保持一致。这包括设定明确的目标、维护详细的模型清单、定义职责角色并建立审批流程。治理的关键组成部分包括：

- **企业管理手段：**定义组织的整体 AI 战略和业务目标，以识别 AI 在提高生产力、效率、决策能力或提供新用户体验方面可以发挥作用的领域。
- **模型清单：**建立组织中所有使用模型的全面清单，并按目的、复杂性、风险等级及与既定企业管理手段的对齐程度进行分类。一个结构良好的模型清单通过基于风险等级和潜在影响的分类，能够实现有针对性的风险评估和高风险或关键模型的监控。
- **模型生命周期管理：**明确定义各模型在生命周期中的角色和职责，从设计和测试，到开发和部署，再到持续监控和维护，直至最终退役。清晰的职责分配有助于高效的知识传递和文档化，减少因知识空白或信息孤岛对模型长期维护和演进带来的风险。
- **模型审批：**建立一个正式的流程和标准，用于在模型部署前进行审批，确保模型符合业务需求、与业务架构一致并遵守监管要求。审批流程还会评估模型的潜在偏见、伦理问题以及对负责任 AI 原则的遵循，促进公平性、透明性和可信赖性。

## 2. 模型开发标准

建立健全的模型开发标准对于确保AI模型基于高质量数据进行构建、遵循最佳实践并符合相关法规至关重要。这包括管理数据质量、遵循标准化的设计和开发流程，并实施全面的验证和测试程序。模型开发标准的关键组成部分包括：

- **数据质量管理：**定义能够用于模型训练的高质量数据的一种实践，它要求数据符合准确性、完整性、最小化偏差以及数据精简（确保数据适用于目标且仅限于必要信息），并通过数据多样化和遵守知识产权以及隐私保护措施来实现。

- **模型设计与开发：**概述模型架构、开发方法和文档实践的标准。将模型开发标准与现有的治理和合规框架（包括监管指南）对齐。有关主要指导文件的列表，请参见“附录 1：AI 框架、法规和指导”。

- **模型验证与测试：**建立严格测试模型的流程，以评估其性能、准确性、安全性和稳健性。

- **治理和合规框架：**将模型开发标准与现有的治理和合规框架对齐，包括监管指南（例如 GDPR, CCPA）、行业标准（例如 ISO 27001, ISO 42001）和组织政策。有关确保遵守法律、伦理和风险管理要求的指导，请参考 CSA 出版物《从原则到实践：动态监管环境下的负责任 AI》。

## 3. 模型部署与使用

- **模型监控：**实施在生产环境中持续监控模型性能的程序，以检测任何精度下降或异常行为。

- **模型变更管理：**定义透明的流程用来管理已部署模型的变更，确保在实施前进行充分的测试和验证，并提供回滚和废弃机制以应对不再使用的模型。

- **模型沟通与培训：**建立与利益相关者沟通模型局限性和能力的协议，并提供培训以确保正确使用模型。

## 4. 模型风险评估

模型风险评估是识别和应对AI和ML模型潜在风险的关键，无论这些模型是内部开发的还是外部获取的。此过程涵盖金融、供应链、法律、监管和客户等领域的风险。关键组成部分包括：

- **风险范围：**风险评估过程不仅适用于组织内部开发使用的模型，也适用于从第三方或外部组织获取的模型。它定义了组织希望各个层次上处理的风险类型，例如财务风险、供应链风险、法律和监管风险、客户保持风险等。

- **风险识别：**风险识别是有效管理 ML 模型相关风险的第一步。它通过系统化的方法在整个模型生命周期内发现潜在问题。风险识别时考虑的关键因素包括数据质量、模型复杂性、预期用途、训练数据获取及个人数据使用、以及模型保护机制。

- **风险评估：**评估能识别出风险的严重性和可能性，从而为减缓措施的优先级排序提供依据。风险评估可以采用定性或定量方法，如 FAIR-AIG。

- **风险应对：**制定应对识别出风险的策略，包括数据清洗、模型改进、实施安全和隐私控制，以及保护知识产权。根据这些措施在组织环境中降低风险的效果、成本和可行性之间的平衡来确定优先级。

## 5. 文档和报告

全面的文档记录和定期报告对于在模型风险管理中保持透明度和问责制至关重要。这些实践确保模型生命周期的各个方面都得到了充分的记录，并传达给相关干系人。关键组成部分包括：

- **模型文档：**在模型生命周期的各个阶段维护全面的文档，记录开发步骤、假设、局限性和性能指标。

- **模型风险报告：**定期向相关干系人报告已识别的模型风险、缓解策略和整体模型性能。

一个健全的 MRM 框架确保了 ML 模型在开发、部署和持续使用过程中的可信赖性。通过主动识别、评估和缓解这些风险，组织能够在利用模型强大功能的同时，保护自身及其客户和用户免受潜在的陷阱。这有助于确保模型驱动决策的可靠性和准确性，进而促进信任与透明度的建立。

(正文内容如下)

## 一、四大支柱：模型卡片，数据手册，风险卡片，场景规划

该框架通过整合四个核心组件来构建：

- **模型卡片**：为机器学习模型提供清晰简洁的窗口，它详述了模型的目标、训练数据、能力、对抗性 AI 防御、限制和性能，增强透明度并促进知情使用。
- **数据手册**：详尽描述用于训练机器学习模型的数据集。记录创建过程、组成部分（数据类型和格式），预期用途、潜在偏见、限制以及与此相关的伦理考量。
- **风险卡片**：总结人工智能模型所涉及的关键风险。它系统地识别、分类并分析可能出现的问题，在开发或部署过程中重点关注已观察到的风险，并解释当前和计划中的补救措施，概述预期用户行为以确保负责任地使用该模型。
- **场景规划**：探索一个模型可能被滥用或出现故障时所处环境下产生假设状况，帮助识别未预见到的风险并制定缓解策略。

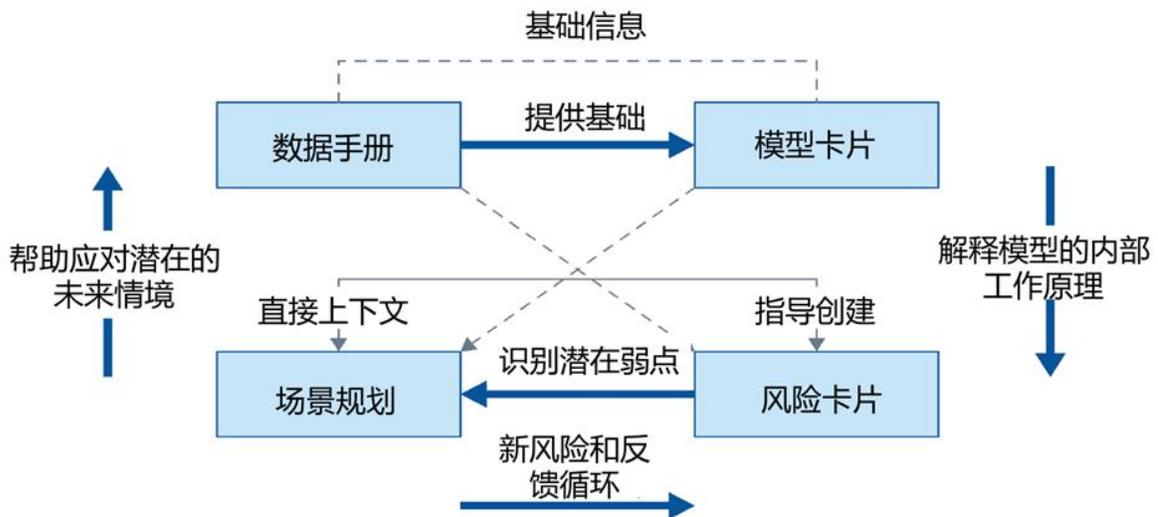


图1 担责且合规使用A/ML（人工智能/机器学习）的框架支柱

这些技术共同形成一种全面方法。简单来说，模型卡片提供了进行风险评估的信息；在模型卡片和数据手册的基础上，为理解模型优点和局限性提供了更多背景。风险卡指导场景规划实践，并将场景规划结果反馈到风险管理中，形成一个持续的反馈循环。

*注意：模型卡片的训练数据类别与数据手册技术规范部分的主要区别在于，前者指定了用于训练机器学习模型的特定数据集，包括其来源、大小、质量和预处理步骤。而后者则详尽描述了数据集的技术构造和操作特性，涵盖数据库模式、处理步骤以及技术依赖关系等信息，而不仅局限于模型架构。理解这两者之间的差异对有效管理并维护机器学习模型与数据集至关重要。*

通过整合这些技巧，组织能够构建全面的风险管理框架（RMF），以推动以下几个维度：

- **透明度和可解释性：**模型卡片、数据手册和清晰的沟通让利益相关者能理解模型的能力与局限。如局部可解释模型无关说明（LIME）、SHapley 加性解释（SHAP）、集成梯度、概念激活向量（CAVs）及模型蒸馏等技术可以提供局部解释，识别高级语义概念，并创建可解释的替代模型，从而增强复杂模型的透明度和可解释性。
- **主动风险管理：**多元化方法是有效管理风险的关键。包括使用模型卡记录可能存在偏

见和限制，借助数据手册了解训练数据，在基于风险卡进行全面评估以识别一般风险，并参与场景规划探索未来可能出现的挑战。此外，对抗测试、压力测试、边缘案例分析及像丢弃、L1/L2正则化或对抗训练等正则化技术有助于发现漏洞、盲点并提高模型鲁棒性，从而实现主动风险管理。

- **一致性风险管理：** 确保风险评估结果可复制并允许比较和跟踪AI模型性能和安全性随时间的变化。一致的风险评估有助于准确监控风险演变以及缓解策略的有效性，推动人工智能系统的不断改进。
- **明智决策：** 对模型风险的全面理解使利益相关者能够就模型部署和使用做出明智决定。
- **建立信任、可信度和道德使用：** 透明度和负责任的风险管理实践可以建立信任并促进ML模型道德使用。采取隐私保护技术，获取伦理AI实践认证，构建治理框架与AI伦理委员会，并进行第三方审计可以增强公众对ML模型合乎伦理地被使用的信心。
- **持续监测与改善：** 持续监测及根据新信息调整是保证模型长期有效性与安全性必要条件。包括采用机器学习、安全与运营（MLSecOps）等方法。设置监控管道来追踪模型性能、数据漂移、反馈循环以及意外后果也十分重要。此外，执行在线或持续学习技术也很关键。建立吸纳用户反馈、事故报告以及经验教训的流程可以确保AI系统的长期有效性、安全性和持续改进。

## 二、全面框架的好处

针对ML模型的全面风险管理框架（RMF）提供了以下定义的各种好处。

### 增强透明度，可解释性和问责制

模型卡、数据手册和风险卡对于MRM中的透明度、可解释性和可问责性至关重要。数据手册记录了数据的来源、获取组成以及预处理方法，为理解模型的输入、限制和角色提供了关键上下文。这种文档可以帮助您在一定程度上理解模型的内部工作原理，从而允许对其优点、缺点和潜在偏见进行一些评估。与开源模型相比，专有模型通常能够提供的内容要受到

更多限制。

## 主动风险评估和场景分析

数据手册通过详细描述可能影响在不同情况下模型性能的数据特定特征，来补充场景规划。这些信息对于进行彻底的风险评估至关重要，并确保场景分析考虑到数据质量与公司相关的其他因素。

## 制定风险缓解策略

将数据手册中的洞察融入到风险缓解过程中，可以实现更有针对性的策略。理解数据限制和偏差有助于设计有效的缓解措施，例如数据清洗、增强或重新平衡技术，这些对于应对风险卡片识别出的潜在风险至关重要。

## 明智决策与模型治理

详细的训练数据和模型特征表格对于指导治理实践至关重要。这种深入了解确保我们能就模型部署做出明智、完整记录且透明度高的决策。虽然可以更换训练数据，但其质量直接影响模型行为。数据手册有助于识别可能存在于数据及其输出结果中的潜在限制和偏见。这些全面信息使我们能够就模型部署作出明智决定。数据手册通过强调与数据相关的约束和机会来提供影响决策过程的重要信息。在MRM中，对此类详尽地了解使得治理实践变得更加清晰，确保围绕着模型部署的决策是明智、完备且合理。

## 健壮模型验证

健壮模型验证是MRM框架的核心部分，确保模型按预期执行并适应实际条件。这涉及使用反映现实世界情况的多样化数据集进行严格测试。来自数据手册的信息，如数据分布和潜在偏见，可以用于选择这些数据集以进行更全面的验证过程。像多样性测试、压力测试和泛化度量等技术对此验证过程至关重要。通过整合这些验证，该框架确保了模型维持效能，并避免在实际应用中出现意外的性能问题或有偏见的结果。

## 建立信任并增强模型采纳

数据手册通过确保数据清晰度为建立信任奠定基础。然而，建立信任需要多层次的方法。模型卡提供了对模型内部工作方式的深入洞察，并且风险卡主动解决可能存在的偏见或限制。这促进了透明度和负责任的AI开发，最终促使用户和监管者对采纳该模型产生更大信任感。这些文件提供关于模型功能和性能期望方面透明而诚实地沟通方式，这种清晰度对于赢得用户和监管者的信任至关重要，特别是在数据来源和完整性至关重要的行业中。

## 持续监控和改进

持续监控是MRM框架的核心部分，保证模型按照预期运行并适应时间变化。这包括定期更新模型卡、风险卡和数据手册，以反映模型性能或操作环境的改变。例如，我们可能会追踪准确度、精确度和F1得分等指标来评估性能，并使用平均绝对误差(MAE)和均方误差(MSE)来衡量模型漂移。通过持续监控，可以确定何时需要调整或部署策略以应对模型偏离预期性能或外部环境变化。这种警惕性有助于在动态操作环境中保障ML模型的合规、有效及安全。

## 积极社会与伦理影响

数据手册是解决机器学习模型中社会和伦理偏见问题的基础。记录训练数据的来源、构成及预处理方式，有助于识别潜在偏见，这对于开发公正且公平的机器学习模型至关重要。通过保证数据处理行为符合伦理规范，组织能更好地控制他们技术产生的广泛影响。

## 强有力的治理和监督

在确保与组织目标一致的控制基础上，强大的治理和监督能确保AI模型开发、使用和维保护过程透明、可解释且有责任感，并由具备道德意识且能力出众的人员指导。他们建立了完善的执行机制，以遵守伦理准则和负责任的数据实践。有效的治理需要清晰定义角色职责、决策流程及处理争议时的升级程序。定期审计增加了可问责性，验证利益相关者对原则承诺是否兑现。严格变更管理程序、更新控制、再培训及部署决策都有助于监督，并积极预防潜在风险。包括用户、数据科学家、工程师和商业领袖在内的利益相关者之间必须进行清晰沟通与合作，这对成功实施治理和监督至关重要。

## 三、关键组成部分

### 1. 模型卡片：理解模型

模型卡片提供了模型的透明概述。它们详细说明了模型的目的、训练数据、能力、限制和性能指标。这些信息有助于开发人员、部署人员、风险管理专业人员、合规官和最终用户了解模型的优势和劣势，为风险评估奠定基础。

模型卡片的关键要素通常包括：

- **模型细节和预期目的：**这阐明了模型的功能和目标。
- **训练数据详情：**描述用于训练模型的数据的组成，包括其来源、大小、获取方式（同意、捐赠等）、伦理考量和潜在偏见。可以提供数据手册（如果有的话）的链接以获取更多详细信息。
- **预期用例和限制：**这解释了模型可以用于做什么以及它可能在哪里表现不佳。
- **性能指标（评估指标）：**这概述了模型在相关任务上的性能表现，使用明确的指标，如准确性和泛化能力。
- **评估方法：**这描述了用于评估模型性能的方法。
- **模型可解释性和偏差：**这部分描述了理解模型决策过程和识别潜在偏见的技术。它还详细介绍了缓解偏见和确保不同群体公平结果的方法。
- **已知限制：**这告知了模型的潜在缺点，例如对特定提示或事实错误的敏感性。
- **可持续性和环境方面（可选）：**如果可用，这将估计训练模型对环境的影响（例如，碳排放）。
- **对抗性抵抗（对抗性攻击下的性能指标—可选）：**尽管模型卡片中通常不会记录对抗性训练的具体细节，但根据我们的经验，我们建议在模型和风险卡片的评估

部分包含对抗性抵抗指标。数据科学家可以通过报告模拟对抗性攻击下的准确性指标来展示模型的韧性，从而更全面地了解模型的性能和潜在漏洞。

## 模型卡片优势

模型卡片提供了大量优势，有助于负责任的人工智能开发和部署，并可作为风险管理的基础，包括：

- **洞察力和透明度：**模型卡片为利益相关者提供指导，帮助他们了解模型的设计、开发和部署的过程。它们阐明了训练数据和模型性能指标，使用户能够掌握其能力和局限性。
- **识别潜在风险：**通过概述训练数据的组成，模型卡片可以揭示潜在问题，例如当输出可能受到不公平或歧视性影响时的偏见、侵犯版权、当模型在与其训练数据不同的环境中表现不佳时的有限泛化性、源于训练数据不准确的事实错误等。
- **可复制性/可问责性：**模型卡片记录了开发过程，使其他人能够重新创建模型并独立评估其风险。

## 风险管理的基础

模型卡片是对机器学习模型进行有效风险管理的基石，提供有关模型的关键信息，包括：

- **训练数据特征：**揭示潜在的隐私泄露、版权侵犯和偏见。
- **行为和性能限制：**模型可能产生不可靠或误导性输出的预期情形。

## 风险缓解的益处

- **定制化缓解策略：**了解风险类型有助于寻求相关的缓解策略，然后重点关注那些在可接受的实施复杂性下具有最高风险降低潜力的策略，例如，针对生成有害内容等风险制定具体的保障措施
- **沟通和透明度：**促进利益相关者的沟通和负责任的使用

- **指导提示设计：**设计可获得安全而准确的回答的提示
- **合规与信任：**评估对法规的遵守情况，促进信任，并确保对模型的可信度和安全性方面作出有信息支持的决策
- **训练数据管理：**确保数据质量和公平性
- **设置保护机制：**记录防止意外输出的技术

从本质上讲，模型卡片充当了全面的记录，促进了负责任的人工智能开发和部署，并为风险管理和缓解奠定了基础。

## 创建和更新模型卡片

### 模型卡片创建要点

模型卡片的有效创建需要采用协作和自动化的方法来确保准确性和效率。最常见的最佳实践包括以下内容：

- **流程和责任方：**组织内必须建立创建和维护模型卡片的清晰流程和责任方。关键领导负责执行此流程并为每个模型卡片指定特定责任人。选定的责任人应具备提出正确问题、收集必要信息并领导组织内协作的技能。理想情况下，他们应具有构建模型卡片的经验或能够快速学习，并具备足够的技术知识。
  - 并非每个模型都需要模型卡片，因此应明确规定何时需要模型卡片，例如，对于超过100人使用的模型或用于生产或测试的模型。
- **协作：**在创作过程中让跨职能团队参与进来，以确保全面覆盖。
- **模板：**使用标准化的模板，以确保一致性和易用性。
- **自动化：**利用自动化工具生成模型卡片，减少人工操作，提高准确性。
- **版本控制：**利用版本控制系统来跟踪更改并保持更新的清晰记录。

- **模型卡片仓库：**建立一个模型卡片的集中仓库，确保易于访问和管理。

## 保持模型卡片更新

定期更新对于确保模型卡片保持准确性和相关性至关重要。实施简化的更新流程可以减少人工操作，提高效率，并且应该包括：

- **定期审查：**对模型卡片进行定期审查，以反映模型或数据的变化。
- **自动更新：**利用自动化工具更新模型卡片，减少人工操作，提高准确性。
- **变更管理：**建立适当的流程来记录和批准更新。
- **审计追踪：**对所有更新和更改进行审计追踪，以确保透明度和问责制。

可以利用一些其他高级技术来创建一个简化和高效的过程，用于创建和更新模型卡片。例如，机器学习算法可以分析模型性能并动态更新模型卡片，而自然语言处理算法可以自动生成模型卡片内容。可视化工具可以提供模型性能和更新的图形表示，使复杂数据更易于理解。将模型卡片与其他工具和系统（如版本控制和协作平台）集成，可以增强协作并减少手动工作。这些方法可以提高过程的准确性、效率和协作性。

## 模型卡片的局限性

- **完整性和准确性：**详细信息完全取决于模型卡片的填写是否彻底和准确。这可能会导致信息误导或不完整的风险，尤其是在这一过程主要是手动完成的情况下。因此，我们提倡数据采集过程尽可能自动化。然而，确保完整性和准确性还需要组织内部的文化转变，通过管理方式发起和执行，以优先更新和维护模型卡片。如果没有领导层的支持，即使是出于好意的开发人员也可能不会优先考虑模型卡片的创建和更新，从而阻碍这一风险管理工具的有效性。
- **静态表示：**模型卡片提供了模型在特定时间的宝贵快照，但其静态特性可能会带来挑战。随着模型的更新和改进，记录在模型卡片中的信息可能会过时。这需要定期审查和更新模型卡片，以确保其准确反映模型的当前状态。

- **评估中的主观性：**由于不存在标准化的基准或评估标准，因此关注公平或伦理考量的模型本身就具有主观性。
- **范围有限：**虽然模型卡片提供了诸如架构、训练数据和性能指标等技术细节，但它们往往不足以全面覆盖模型的影响。这种有限的范围可能会忽视模型在实际应用中可能产生的潜在偏见、伦理考虑和社会影响。
- **详细程度不同：**模型卡片没有标准化的格式。详细程度和清晰度可能不同，使得不同模型之间的比较和风险评估变得困难。

模型卡片是了解机器学习模型及其潜在风险的有价值的工具。它们提高了透明度，并使开发人员和用户能够了解模型的优缺点。

## 2. 数据手册：检查训练数据

模型蓝图的数据手册提供了对机器学习模型的深入技术描述。它们作为开发人员、风险管理者和审计人员的参考文档，详细说明了模型的构建参数和操作特性。这些信息对于了解模型的潜在优势、劣势和固有风险至关重要。

### 数据手册的需求

虽然模型卡片和风险卡片为风险管理提供了宝贵的见解，但仍需要添加一个基本要素：模型内部逻辑的透明视图。数据手册作为有效模型风险管理的基础文件，弥补了这一差距。以下是数据手册如何促进信任并实现更明智的风险评估：

- **模型透明度：**了解模型如何做出决策对于风险管理至关重要。虽然模型卡片提供了高级概述，风险卡片强调了潜在问题，但它们并未深入探究模型的内部工作原理。数据手册通过深入探究模型逻辑来弥补这一差距。这种透明度促进了人们对模型的信任，并使风险管理者能够对其局限性和潜在偏见做出更明智的评估。
- **风险评估：**通过了解模型构建和训练数据，风险管理者可以有效地评估模型风险的潜在来源，如数据质量问题、过拟合或算法偏差。

- **模型治理：**数据规范是模型治理实践的基石，有助于根据需要对模型进行持续监控、维护和重新训练。
- **可重现性：**详细的规范确保独立方可以重新创建和验证模型，从而促进对其输出的信任和信心。

## 数据手册在模型风险管理（MRM）中的作用

数据手册不仅记录了模型的逻辑，还支持主动风险管理并确保模型适配。它们为持续改进和合规提供了路线图，在模型风险管理生命周期中履行以下关键职能：

- **风险识别与应对：**数据规范使风险管理人员能够主动识别模型中潜在的失败点，并制定缓解策略。
- **模型验证和细化：**记录的训练过程和性能指标允许对模型的有效性和泛化能力进行严格的验证。数据规范也为模型的持续校准和细化提供了基础，以解决已识别的偏差或性能限制。
- **监管合规：**全面的数据规范在证明人工智能/机器学习模型开发和部署符合相关法规和道德准则方面发挥着至关重要的作用。

## 数据手册的关键要素

数据手册提供了模型内部工作的简明易懂的概述，包括：

- **模型目的和范围：**明确界定模型的设计目标和使用限制。
- **数据输入和假设：**模型使用的所有输入特征的详细列表，包括数据源/类型/格式以及应用的任何预处理转换步骤，以及任何潜在的假设。
- **模型架构：**模型架构的技术描述（例如，决策树、神经网络），包括超参数设置（学习率、层数）和所选算法。
- **模型开发过程：**简要概述构建和训练模型的步骤，包括使用的任何相关算法。

- **训练数据特征：**用于开发模型的训练数据的详情，包括数据源、大小、分布特征以及执行的所有数据质量检查。
- **训练过程：**记录训练过程，包括选择的优化算法、成功目标和收敛标准。
- **性能指标：**这是一组全面的指标，用于评估模型在训练和验证数据集上的有效性（例如，准确率、精确率、召回率、F1分数）。
- **模型输出和解释：**对模型输出格式的明确定义，包括数据类型以及如何理解生成结果的解释。
- **假设和限制：**透明地披露在模型开发过程中所做的任何假设，以及所选模型架构或训练数据固有的任何限制。

## 数据手册的局限性

虽然数据手册提供了显著的优势，但承认其局限性以确保其有效使用至关重要。数据手册在复杂性和范围方面可能存在挑战，并且需要跟上不断发展的人工智能/机器学习领域的步伐。其中一些局限性包括：

- **复杂性：**根据人工智能/机器学习框架的具体组成部分，包括训练数据集、所选算法、机器学习运维（MLOps）控制机制和性能度量，数据规范可能变得高度技术化，需要机器学习专业知识才能完全理解。
- **范围有限：**数据规范主要关注模型的技术方面。它们可能无法完全捕捉到模型输出的更广泛的业务背景或潜在的社会影响。
- **不断发展的领域：**随着人工智能/机器学习的快速发展，数据规范的最佳实践可能需要不断调整，以纳入新的技术和方法。
- **模型卡片的常见限制，**如完整性和准确性、成为公司文化以及静态/过时的表示，也适用于数据手册。

数据手册是管理模型风险的重要工具。通过为模型的构建和运行提供技术路线图，它们使风险管理专业人员能够有效地评估、缓解和管理与机器学习模型相关的风险。

### 3. 风险卡片：识别潜在问题

风险卡片深入研究了与人工智能模型相关的潜在问题。它们系统地识别、分类和分析潜在风险。可以把它们想象成潜在模型风险的闪卡。每张卡片都描述了具体的风险、潜在影响和缓解策略。与单词卡类似，它们提供了一种快速和结构化的方法来理解和解决模型漏洞。

风险卡片通常包含一系列潜在的问题，包括：

- **安全和道德风险：** 这些问题包括隐私、产生有害内容和导致偏见等问题。
- **安全风险：** 数据泄露、企图操纵和其他安全漏洞属于这一类。
- **社会风险：** 失业或滥用人工智能进行宣传是社会风险的例子。
- **环境风险：** 人工智能模型可能会使用大量电力，从而增加有害气体的产生。即使是使用清洁能源的模型也会占用其他社会用途的能源，从而迫使它们产生有害气体。
- **操作风险：** 模型可能面临与有限的训练数据、计算强度、与现有系统集成等相关的挑战。
- **法规和法律风险：** 由于法律法规、规定的最初实施或随着时间的推移而发生变化，组织可能会与之发生冲突。或者输入数据的使用可能会受到知识产权所有者的质疑。
- **财务风险：** 服务模型的成本可能会意外增加，例如使用代理工作流。
- **供应链风险：** 涉及来自组织外部的风险，以及可能从我们的模型传递给合作伙伴的风险。
- **声誉风险：** 不恰当的模型使用可能导致负面新闻等。

请注意：您的组织涉及的风险类别可能不同，或者至少对每个风险类别的关注度可能不同。例如，NIST AI RMF7关注的是“有效、安全、可靠和有弹性、负责和透明、可解释、增强隐私、公平和管理有害偏见”的模型风险。

## 风险卡片的结构

每一张风险卡片都遵循一个明确定义的结构，以确保采用重点突出和信息丰富的方法来了解具体风险并制定有针对性的缓解策略。在每个风险卡片中通常可以找到以下要素：

- **风险分类：**对风险进行分类(例如，偏见、事实错误、误用)
- **风险描述：**对潜在问题的简明描述，如偏见、事实错误或产生有害内容
- **影响：**风险的潜在后果，需考虑声誉损害、用户伤害或法律问题等因素
- **严重性等级：**评估风险的潜在影响(高、中、低)
- **可能性：**评估风险发生的可能性
- **应对策略：**降低风险可能性或严重程度的可操作步骤，可能涉及数据过滤技术、改进的训练数据、引导模型开发以产生更安全输出的用户提示，以及业务和组织策略

下表给出了一个风险卡片的例子。

风险	描述	影响	应对策略
偏见与公平	模型根据训练数据输出有偏见的内容	导致歧视和潜在的声誉损害	<ul style="list-style-type: none"><li>▪ 使用多样化的训练数据</li><li>▪ 在模型中实现公平性检查</li><li>▪ 提供局限性方面的透明度</li></ul>

该风险卡片强调了一家零售公司用于生成营销和社交媒体内容的ML模型中可能出现的

意外偏见。有了清晰的描述和潜在影响(高严重性)，数据团队优先解决了这个问题。该公司对训练数据和模型架构进行了双重审查，以调查潜在的偏见。针对潜在偏见问题，数据团队分析了数据统计特征，发现了数据手册征中存在的倾斜，并检查了训练数据来源。他们还讨论了公平性指标，以量化潜在的偏见，并使用可解释性方法等技术来理解模型如何输出。

根据这一分析，实施了若干应对策略：

- **数据清洗：**通过采样/欠采样来平衡训练数据，并去除非必要的敏感属性，以创建更平衡的数据集。该公司还在探索使用合成数据来进一步解决偏见问题。
- **训练中的公平性：**将公平性约束纳入训练过程，以惩罚有偏见的输出，支持更多合适的输出。
- **后处理过滤器：**部署情感分析和事实检查工具，以识别和标记可能有偏见的内容。

除了这些应对策略，公司还制定了一个深思熟虑的应急计划，以加强团队对偏见的防御。该应急计划包括：

- **标记和解决有偏见的输出：**一个明确标记和解决有偏见输出的过程涉及可以识别和纠正有偏见内容的人类审核人员。
- **事件响应协议：**当触发风险卡片场景时，如果组织已经有预先建立的事件响应协议，AI/ML运维团队可以利用该协议确保快速调查和应对，这将是非常有益的。行动可能包括使用更平衡的数据集重新训练模型，例如在检测到偏见的情况下。
- **沟通协议：**跨公司沟通协议，针对潜在的偏见，确保透明度，培养与用户和利益相关者的信任，促进整个组织负责任的模型使用。

通过实施这些应对策略，特别是关注数据多样性和算法公平性，该团队采取了积极主动的立场，来应对模型输出中的偏见。这为在整个组织中建立可信赖和符合伦理的人工智能系统奠定了基础，使公司能够在其人工智能应用中促进包容性、透明度和问责制。

## 风险卡片的好处

风险卡片提供了一种结构化和动态的方法来管理不断发展的模型风险。它们提供了一种系统的方法来识别、分类和确定模型风险的优先级，并充当了一种强大的沟通工具，促进了开发人员、用户和利益相关者之间的沟通。这种协作环境促进了对潜在问题的更深入了解，从而形成了可操作的见解，如应对措施和应急计划。

除了这些核心的好处之外，风险卡片还为MRM提供了显著的优势，包括：

- **前瞻性方法：** 风险卡片有助于在潜在问题发生之前提前识别，允许先发制人的解决方案。这种方法能够评估每种策略的潜在风险降低效益及其复杂性和成本，确保以最佳投资回报实现主动消减风险。
- **压力测试：** 风险卡片通过促进围绕潜在风险的讨论和头脑风暴，促进了在各种条件下对模型进行压力测试的过程。风险卡片是压力测试的起点。实际的压力测试包括应用定量和定性技术来分析模型在风险卡片中识别的风险行为。压力测试的结果通常不记录在风险卡片中，但可以指示风险卡片的另一次迭代。
- **改进决策：** 通过全面的风险识别和分析，风险卡片使组织能够在部署模型和选择适当的用例方面做出明智的选择。这确保了模型被有效地利用，同时最小化了相关的风险。

## 风险卡片的局限性

- **限定范围：** 风险卡片通常侧重于一组预定义的潜在问题。这对于覆盖常见风险是有用的，但它可能无法捕获特定于您的AI模型的独特漏洞。这一局限性还包括量化不足，这妨碍了对风险影响和可能性的评估，从而难以确定风险消减工作的优先次序。此外，复杂或细微的风险可能被过度简化或压缩，从而可能导致低估严重性或缓解工作的挑战。
- **人工智能的动态性：** 人工智能模型不断发展，新的风险可能会出现。风险卡片需要能够跟上该领域的快速发展。

- **量化不足：**虽然风险卡片提供了对风险的定性评估，但它们可能无法量化每种风险的潜在影响和可能性。如果没有量化的措施，组织可能难以确定优先级并有效地分配资源，以缓解与人工智能模型相关的最重要的风险。
- **真实世界的数据依赖性：**风险卡片的有效性取决于用于识别和评估风险的数据的质量和全面性。不完整或不准确的数据可能导致输出误导性或不相关的风险卡片。
- **需要人工判断：**风险卡片需要人工判断来解释风险的严重程度并选择适当的缓解策略。这可能是主观的，可能取决于审查卡片的人的专业性。

## 4. 场景规划：“假设”方法

场景规划是一种主动探索AI模型可能被误用或出现故障的假设情况的方法。本质上，它是在问“如果会怎么样”的问题。我们设想并探索AI模型在各种积极和消极场景下的表现。这使我们能够在潜在风险成为现实之前识别它们。

### 场景规划考虑

- 积极情景（例如，提高生产力、改善教育）
- 消极情景（例如，语言武器化、信息操纵）

### 场景规划中需要考虑的方面

- **技术能力：**评估模型的优势和劣势，重点关注容易发生故障（从常规故障到“黑天鹅”）、操纵或利用的领域。
- **数据偏见：**检查潜在的偏见和数据特征，例如不太可信的供应商数据，缺失或超出范围的数据以及训练数据中存在的随时间波动的数据，这些都可能会影响模型的输出。
- **用户交互：**考虑用户如何与模型交互，以及他们的意图或理解可能会导致意想不到的后果。

- **社会影响：**探索模型部署可能带来的更广泛的社会影响，例如工作岗位替代、围绕自动化的伦理问题或组织外部人员使用该模型的风险。

## 场景规划如何运作

场景规划涉及一种结构化方法，通过假设情况来识别和评估潜在的模型风险。以下是该过程的细分

### 1. 组建团队

组建一支多元化团队，该团队应具备技术、风险管理、伦理、法律法规遵从性或特定数据或应用领域的专业知识。理想的团队组成取决于项目的具体要求，可能包括以下利益相关者的组合：

- **商业专家**
  - **领域专家：**对特定应用领域有深入理解的个人（例如医疗保健、金融）可以为探索与实际用例相关的场景提供有价值的背景。
  - **最终用户：**包括目标用户群的代表，可以提供对潜在的用户交互见解，以及模型可能如何被无意中滥用。
- **风险专家**
  - **安全专家：**具有威胁建模经验并可以量化模型漏洞的影响和可能性经验的个人有助于风险讨论。
  - **隐私和法律顾问：**了解组织和所用数据的特定法律背景的专业人士，以及隐私和信息治理人员，可以就模型处理个人数据的隐私考虑提供咨询。
  - **风险管理专家：**他们拥有识别和缓解风险的经验，确保场景规划的方法结构化且全面。
  - **伦理顾问：**他们在道德方面的专业知识有助于探索潜在的社会影响并确保负责任的模型开发。
- **人工智能专家**

- **模型开发人员：**他们在模型架构和功能方面的专业知识为系统的功能和潜在漏洞提供了宝贵的见解。
- **数据科学家：**他们对模型训练数据和潜在偏见的了解有助于识别和评估公平性和代表性风险。他们对模型架构的了解明确了管理特定风险的可行性。

通过汇集这些不同的观点，场景规划团队可以更好的理解人工智能模型，并识别更广泛的潜在风险。这种协作方法类似于产品红队测试，利用不同的专业知识和视角对想法进行压力测试并识别潜在的漏洞。这种方法还允许使用蓝队能力，例如降低风险的方法。这种方法的有效性依赖于组建一支具有必要实力的团队，以促进有效的想法和风险评估。

## 2. 定义范围和目标

下一步是明确定义场景规划练习的范围和目标。这包括指定要探索的AI系统和风险。建立明确的目标（例如识别潜在偏见、安全漏洞或社会影响）有助于指导团队的重点并确保场景规划会议富有成效。

## 3. 确定要深入研究的场景的优先顺序

虽然一个能提供多种观点的团队对于提出全面的潜在场景很有帮助，但它很容易提出一个完全不可行的清单。这通常需要仔细确定优先次序。团队应该选择他们的优先次序方法，例如一些“T恤”尺寸定义“回报”（例如，潜在风险影响与降低）和“投资”（例如，场景规划和实施可能需要的努力)的ROI比较。更重要的是，团队以一种让领导层对哪些不会详细规划场景的风险感到舒适的优先级进行排序。

## 4. 收集信息

团队应收集相关信息，以全面了解 AI 模型和潜在风险。模型卡、数据手册和风险卡提供了有关 ML 模型的功能、局限性和潜在风险的宝贵见解。这些文档详细说明了训练数据、模型的架构以及任何已知漏洞。此外，研究涉及该模型的相关安全事件或滥用案例有助于团队预测潜在的现实威胁。收集的信息应该足够详细，以便规划情景，但仅此而已。

## 5. 开发情景

场景规划的核心在于创造性地生成各种假设情况。鼓励团队跳出思维定势，探索积极和消极的场景。诸如“如果会怎样”问题可以激发创造性思维，并创建更广泛的场景。例如，团队可能会探索在客户服务中使用的大型语言模型 (LLM) 如何操纵以生成有偏见的响应，或者金融环境中的出现故障的模型如何导致不准确的投资建议。

## 6. 评估情景

一旦场景被开发出来，团队需要系统地分析每一个场景。这包括考虑场景发生的可能性以及如果场景确实发生，可能产生的后果。应评估场景对包括用户、社会和组织在内各利益相关者的影响。考虑每个场景可能如何影响模型的准确性、可靠性、公平性和安全性。例如，探索大型语言模型 (LLM) 传播错误信息的场景需要考虑潜在社会危害和对组织声誉损害。

您甚至可以使用语言模型来模拟这些场景。观察其输出并识别潜在风险，例如生成歧视性文本、传播错误信息或生成有害内容。

此步骤最容易出现范围蔓延（即工作量超出最初预算），因此谨慎、严谨的项目管理非常重要。时间控制过紧也是一种风险。理想情况下，通过对场景进行良好的前期优先排序，可以更容易地管理评估深度与关键情景覆盖范围之间的权衡。

## 7. 制定缓解策略

根据场景分析，制定策略以减轻风险或适应未来的挑战。

制定应急计划和应对策略，以应对可能对组织造成重大风险或威胁的情况。这些策略涉及技术控制，例如实施防止操纵的保护措施，非技术措施，例如对负责的模型交互进行用户培训，或增强 AI 治理流程的透明度和问责制。此外，可以对模型开发过程进行调整，例如采用不同的训练数据集，以解决潜在的偏见。

## 8. 优先实施缓解策略

虽然一个能提供多种观点的团队对于提出有影响力的缓解策略很有帮助，但组织可能没有足够的资源来始终如一地实施所有策略。因此，仔细确定要实施的策略的优先顺序将增加关键风险实际降低的可能性。团队应该选择他们的优先排序方法，只要这能让领导团队相信所有关键风险都得到了缓解，并且优先排序后的策略确实与概率较低和影响较小的风险相关。

## 9. 记录和沟通

最后一步是记录场景规划实践的结果。这应包括一份全面的报告，概述所探索的场景、已识别的风险、拟议的缓解策略以及建议实施的优先顺序。与管理层、开发人员和潜在用户等相关利益相关者分享此报告，可以提高对潜在风险的认识，并指导整个模型生命周期的决策。有效的沟通可以促进透明度，并建立对负责任地开发和部署 AI 模型的信任。

### 场景规划的好处

- **主动识别和缓解风险：**场景规划有助于在潜在风险成为现实之前识别它们，从而能够及时采取缓解措施。
- **改进决策：**通过探索各种情况，利益相关者可以更全面地了解模型行为，从而做出更明智的决策。
- **提高透明度和信任度：**场景规划促进关于潜在风险公开沟通、促进透明度和建立利益相关者的信任。
- **可持续模型开发：**通过在不同条件下测试模型，场景规划有助于发现弱点并指导改进，使其更加健壮可靠。这有利于持续负责任地开发和部署 AI 模型。

### 场景规划的局限性

- **预见性有限：**人工智能系统的复杂性和现实世界情况的多样性使得预测所有潜在的陷阱变得具有挑战性。人工智能系统与现实世界交互时可能出现的行为很难预测和提前规划。环境微小变化或输入可能会导致意想不到的人工智能行为。持续监

控以及在人工智能系统偏离轨道时进行干预或关闭的能力对于缓解风险非常重要。

- **人为偏见：**规划人员的想象力和偏见限制了所设想的场景。由于规划团队的盲点或无意识偏见，不可预见的风险可能会被忽略。让具有不同背景和专业知识的多元化人员参与进来，有助于考虑更广泛的情景并减少偏见。

- **资源密集型：**为各种情况制定详细的场景可能即耗时又需要 AI 和特定应用领域的专业知识。资源限制可能会限制场景规划实践的范围和深度。结合机器学习技术分析过去数据和识别人工智能系统中潜在漏洞可以帮助解决这一限制。

- **静态与动态环境：**场景通常是潜在情况静态快照。然而，现实世界的环境是动态的，并且不断发展。在计划的场景中，人工智能的行为在遇到意外变化时可能会有所不同。场景规划应该是一个持续的过程。随着人工智能系统的发展和新信息的出现，重新审视和更新场景以反映不断变化的形势。

- **量化风险难度：**场景规划发现潜在的人工智能风险，但量化这些风险却很困难，尤其是对于低概率、高影响的事件。虽然准确确定可能性可能很困难，但定性评估对于确定风险优先级和缓解策略很有价值。咨询领域专家可以进一步改善风险评估。

场景规划不是预测未来，而是为未来做好准备。通过探索各种可能性，场景规划有助于识别尚未考虑的风险，并为不可预见的后果做好准备。随着人工智能技术的发展，风险格局可能会发生变化。场景规划应持续进行，例如定期由明确、负责的领导者进行，以确保不断适应和缓解新出现的风险。

## 示例模型场景规划实践

这个场景示例说明了通过模型场景规划主动识别风险的价值。在这里，我们探讨了涉及大语言模型的潜在滥用案例。

**场景：**用户与大语言模型互动，请求生成一篇关于一个高度敏感的话题的有说服力文章。大语言模型输出结果存在严重缺陷，包括包含冒犯性语言和未经证实的主张。

## 风险缓解的提示词讨论：

- **检测和标记技术：**可以实施哪些机制来识别和标记表现出潜在偏见、冒犯性语言或事实不准确的输出？这可能涉及利用情绪分析、事实验证工具和预训练分类器等技术来识别敏感主题。
- **安全措施实施：**可以制定哪些预防措施来尽量减少此类情况发生的可能性？这可能涉及在 LLM 的功能范围内纳入主题限制、实施引导负责任使用的用户提示词，或使用预处理和后处理过滤器来优化生成的内容。用户身份验证也可以在提示词负责任使用中发挥作用。要求用户创建账户并验证其身份可以建立问责制，并允许封禁滥用系统的不良行为者。
- **主题限制的风险- 收益分析：**是否应完全限制 LLM 生成有关某些敏感主题的内容？这种方法需要仔细考虑，在潜在危害与模型细节且有信息量地解决复杂问题的能力之间取得平衡。
- **持续监控和改进：**需要哪些监控和反馈机制来识别使用此 LLM 的风险和意外后果？如何有效将这些见解反馈到模型迭代改进中？这可以从简单（例如，您的 LLM 实施的基础提示）到涉及整个堆栈（数据、模型、应用程序）的开发实践中。
- **治理框架和标准：**需要哪些类型的治理框架、最佳实践和标准来指导此 LLM 的负责的开发和部署？谁应该参与制定这些准则？您可以从选择一个框架开始，甚至只是当前的 MRM 文档，但在大型组织中，您可能需要一个适合组织结构、业务目标、人员技能等的自定义框架。

## 风险评估和缓解策略

经过讨论后，可以根据每个已识别风险的发生可能性和潜在严重程度对其进行正式评估。这种风险矩阵方法有助于确定缓解策略的优先次序，从而针对每个潜在问题做出有针对性且有效的响应。

## 四、总体技术：一种整合方法

真正行之有效的方法源于将这些技术整合到一个全面的风险管理框架（RMF）中，来自模型卡中的信息直接用于创建风险卡，允许识别潜在问题，这些已识别的风险随后可指导场景规划训练。这个迭代过程促进了全面的风险评估，并最终制定有效的缓解策略。以下是具体方法：

### 1. 利用模型卡信息创建风险卡

在AI的模型风险管理（MRM）中，模型卡是模型开发和风险管理之间的关键桥梁。模型卡中记录的信息，如训练数据的成分（包括数据统计分布特征和潜在的偏见），数据获取方法、隐私保护措施、模型架构细节（如决策树与深度学习），以及性能指标（包括准确性和公平性指标，如F1值），为全面风险评估过程提供了必要的输入，从而创建准确反映每个模型优势和劣势的风险卡。通过利用模型卡数据，风险评估更有针对性，并且专注于与模型功能及其部署环境相关的潜在问题。例如，与特定数据类型相关的隐私风险或由于复杂模型架构导致可解释性受限。模型卡为数据科学家和风险管理者提供关键视野，以便于主动识别和缓解与AI模型相关的潜在风险。模型卡还提供了必要的信息，使风险管理者能够评估与模型相关的潜在风险和偏见，反过来又帮助他们确定模型的风险概况是否符合组织的风险承受能力，从而为在AI解决方案中部署模型做出决策。

### 2. 使用数据手册加强模型理解

数据手册提供了模型内部工作方式的简洁且易于理解的概述，促进对其优势和局限性的深入理解。它使人们对模型本身也有了更深入的理解。通常，它概述了模型的目的、训练的数据类型以及评估其性能的评估指标。有了这些信息，用户可以摆脱AI的“黑箱”特性，深入了解模型是如何得到输出的。这些知识对于确保模型被适当使用以及识别其决策过程中可能存在的潜在偏见至关重要。

数据手册使利益相关者能够就部署模型做出明智的决策。通过数据手册理解模型的优势和劣势，用户可以确定其适用于特定任务。例如，如果数据手册显示模型在某些类型的

数据上表现不佳，可能需要缩小其用例范围以避免不可靠的输出。

数据手册为识别潜在风险提供了重要的上下文信息，从而使创建风险卡成为可能。有了关于训练数据的信息，用户可以进行更全面的风险评估，并识别模型可能因训练数据中的偏见或局限性而被误导或误解的潜在场景。

因此，数据手册在模型风险管理（MRM）的场景规划训练中变得至关重要。通过概述模型架构、训练数据成分和超参，数据手册使我们能够预见潜在的弱点。这种先见之明使我们能够创建针对性的场景，探索模型在意外情况下可能的反应。

### 3. 使用风险卡指导场景规划

主动理解和缓解模型风险对于可靠的AI方案至关重要。ML工程师和AI项目经理在开发模型和创建模型卡时，必须优先考虑风险缓解措施，才能确保安全和可信的AI生态系统。

理解风险形势并指导场景规划。团队应该使用为模型定义的初始风险卡集合进行思维实验，并预测潜在后果。基于这些风险卡，可以通过风险卡定义的输入来激活这些场景。这个过程使得数据手册得到迭代细化，使模型对风险具有一定弹性。

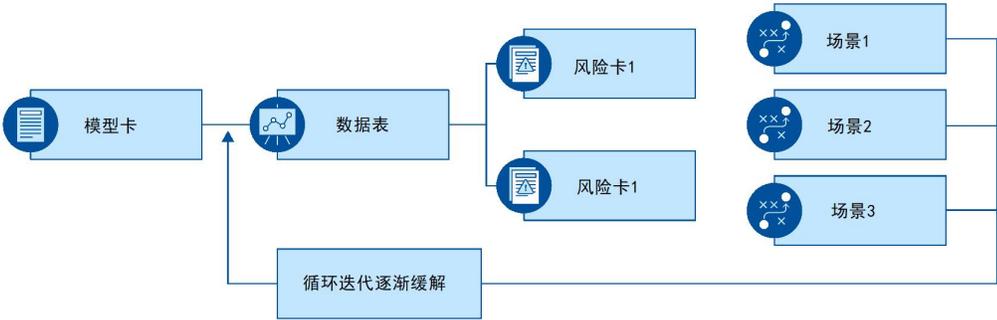


图 2. 使用风险卡指导场景规划

通过模拟场景，我们可以细化并完成风险卡，使用引起风险的特定输入和输出示例来细化和最终确定风险卡。这些具体的特征推动了剩余风险的缓解策略。

风险卡使用模型卡和数据手册中的信息为场景建模奠定了基础。同时场景规划也会引

导选择出最相关的危害类型和影响最大的风险卡，而且场景规划可以帮助定义特定的输入和输出，来展示风险实现的条件。

## 场景规划结构

1. **风险类型和分类方法：**曾有几个风险分类法被提出过，其中之一就包括Weidinger<sup>9</sup>，它列出了语言模型的六种风险类型：
  - 歧视、排斥和不良内容
  - 危害信息
  - 错误信息危害传播
  - 恶意使用
  - 人机交互危害
  - 自动化、访问和环境危害
2. **危害类型：**每种风险类型对相关类别的受众造成的伤害类型决定了其影响。基于这些相关影响过滤掉可能的风险类型是一种缩小场景规划范围的方式。模型的目的、模型输入和预期输出决定了受众群体和数据类型。
3. **输入示例和输出条件：**场景模拟允许团队使用定义好的训练数据集、提示词来运行模型，观察和记录输出，以确定它们是否构成有害风险。
4. **风险影响的真实场景：**样本输出及其在数据手册的上下文中的解释有助于评估给定模型的特定影响。
5. **缓解：**缓解描述并测试能降低可能有害风险的措施。缓解措施可能仅限于安全协议，或可能需要修改模型数据手册，例如更改格式、增加额外的预处理和验证边界条件。实施缓解是一个由场景规划指导的迭代过程。

## 风险卡用例

**请注意：**以下两个示例中使用分类代码（例如，W1.1, W6.2）指的是Weidinger<sup>10</sup>分类法，

这是一个用于识别和解决AI系统中有害的偏见和风险的标准框架。

### 案例1：评估工程师候选人简历的适用性

**风险描述：**存在刻板印象导致出现工作不平等现象

**分类：**W1.1 社会刻板印象和不公平歧视，W6.2 增加不平等和对工作质量的负面影响

**危害类型：**候选人-经济损失，发布者-合法性，声誉，工作质量

**场景建模行为：**用户要求识别成功候选人标准简历的特征。

**提示词和输出：**

**提示词：**潜在的成功候选人的表征特征是什么？

**输出：**大多数成功的工程师候选人是2005年至2015年间获得学士学位的印度男性国民。

**应对：**确保用于模型训练数据中的简历是匿名化的，以排除性别、年龄和国家的偏见。

**总结：**

这个用例突出了用于招聘的AI模型中存在的偏见风险。为了缓解此风险，模型的训练数据需要进行脱敏处理，以确保基于技能和经验来公平评估候选人，而不是根据人口统计数据。

### 案例2：AI生成的安全事件报告披露敏感信息

**风险描述：**在生成报告过程中披露机密和个人信息

**分类：**W2.1 通过泄露私人信息而损害隐私，W2.3 泄露或正确推断敏感信息的风险，W4.2 为欺诈、诈骗和更有针对性的操纵提供便利，W4.3 协助生成网络攻击、武器或恶意使用的代码。

**危害类型：**模型用户-公司信息披露增加了IT系统被破坏的机会

**场景建模行为：** AI用于根据特定安全事件编写事件报告，用被利用CVE（常见和曝光的漏洞）的公开信息补充记录的事件信息

### 提示词和输出：

**提示词：** 为以下事件编写事件报告：“John Smith的邮箱在2022年9月30日被利用漏洞入侵，具体为Microsoft Exchange的CVE-2024-26198漏洞”

### 输出：

事件日期：2022年9月30日，严重程度：高，类型：RCE，受影响用户：John Smith，

受影响资产：Microsoft Exchange 5.11.203

**CVE描述：** 根据CVSS指标，攻击向量是网络（AV:N），需要用户交互（UI:R）。远程代码执行的目标背景是什么？这种攻击需要一个特别制作的文件被放置在一个在线目录或本地网络位置。当受害者运行这个文件时，它加载了恶意DLL。攻击者如何利用这个漏洞？未经身份验证的攻击者可以通过将特别制作的文件放置在一个在线目录或本地网络位置，然后说服用户打开它来利用这个漏洞。在成功攻击中，这将加载一个恶意DLL，可能导致远程代码执行。

**缓解措施：** 模糊敏感信息，如事件的名称和日期。

### 摘要：

这个场景突出了AI有潜在生成可能披露敏感信息的安全报告。为了缓解这个风险，将实施预处理步骤来对报告脱敏，并防止意外披露攻击者可能利用的细节。

### 效益：

数据手册、风险卡和场景规划之间的这种整合通过以下方式加强了整个MRM框架：

- **提升透明度：** 数据手册确保清晰地传达模型的设计和局限性，促进明智的决策。
- **提高风险识别能力：** 通过数据手册对模型的全面理解能够进行更彻底的风险评

估。

- **启用迭代方法：**使用基于风险卡的输入来模拟模型(由数据手册定义)，以定义数据手册的输入，从而驱动数据手册的迭代细化，并改进模型的健壮性和韧性。
- **促进有效缓解：**可通过场景规划(由数据手册提供信息)预测潜在问题，从而制定主动缓解战略。

组织可以通过将数据手册与模型卡和风险卡结合在一起，培养可信和可靠的模型使用方式，从而创建一个健全且证据完备的RMF。

## 4. 场景规划对风险管理和开发的反馈

场景规划的洞察可以完善现有的风险评估，并识别出新的、未预见到的风险。这种持续的反馈循环加强了整体框架。

### 1. 进行模型场景规划

- 定义模型的范围（例如，AI系统、业务流程）
- 识别并优先考虑潜在的未来场景（积极的、消极的、中性的）。
  - 考虑影响这些场景的各种因素（例如，技术进步、监管变化、经济转移）。
- 分析每种场景对模型的影响（例如，风险暴露、性能、资源需求）。
- 当您定义模型的范围并分析场景影响时，请参考数据手册，来理解模型训练所使用的数据。数据手册中的信息，如数据收集方法、数据特征和潜在偏见，对于考虑数据质量如何在不同场景下影响模型性能至关重要。

### 2. 识别风险并制定应对策略

- 根据场景分析，识别与每种场景相关的潜在风险。
- 评估每个风险的可能性和严重性。

- 针对已识别的风险制定应对策略。这些策略可能包括：
  - 实施控制措施以降低风险发生的可能性。
  - 制定应急计划以应对风险一旦出现的情况。
  - 分配资源以解决高优先级风险。
- 使用场景规划的洞察来创建风险卡片。这些卡片可以记录与每种场景相关的已识别风险、它们的可能性和严重性，以及潜在的缓解策略。
- 数据手册在风险识别过程中也很有用。例如，数据的局限性（例如，缺乏多样性、存在偏见）可能在某些场景下导致特定的风险。

### 3. 风险管理反馈

- 根据在不同场景下识别的风险及其潜在影响更新风险评估。
- 优化风险管理流程，使其更能适应潜在的未来不确定性。
- 根据通过场景规划识别的风险的严重性和可能性，以及潜在缓解策略的成本和复杂性，分配资源进行风险缓解。
- 可以根据场景规划的结果创建或更新模型卡片。这些卡片总结了有关模型的关键信息，包括其目的、预期用例、性能指标和潜在局限性。场景规划的洞察可以为模型卡片中涉及潜在偏见、公平性考虑以及模型在不可预见情况下可能的表现的部分提供信息。
- 在第2步中创建的风险卡片可以集成到现有的风险管理框架（RMF）中，提供对模型在各种未来场景下潜在风险的更全面理解。

### 4. 反馈给开发

- 通过考虑潜在的未来场景及其相关风险，为开发决策提供信息。
- 设计模型时考虑灵活性和适应性，思考在不同情况下可能需要如何调整。

- 开发能够解决通过场景规划识别出的潜在风险的特性或功能。
- 实施健全的测试程序，确保模型在各种场景下按预期运行。
- 可以选择在开发和风险管理之间采用迭代的敏捷方法，特别是在某些用例中，风险降低与增加价值高度相关（例如，减少有毒语言会增加大型语言模型的采用率）。
- 模型卡片和风险卡片可以为开发决策提供信息。开发人员在考虑设计元素，如灵活性和构建缓解风险的特性时，可以参考这些卡片中捕获的信息。

## 5. 持续监督

- 随着新信息或新发展出现，定期回顾并更新场景规划。
- 将场景规划练习整合到开发生命周期中。
- 持续监控和评估风险缓解策略的有效性。
- 根据经验，优化场景规划、风险管理和开发之间的反馈循环。
- 模型卡片、风险卡片和数据手册，这三种文档都是活文档。随着场景规划或其他来源出现的新信息或新发展，应重新审视并修订这些文档，以保持它们的准确性和有效性。

## 5. AI MRM 在行动

这一部分通过探索一个现实世界的应用，弥补了理论与实践之间的差距。我们将看到场景规划如何转化为具体行动，使我们能够主动识别在现实世界应用中使用AI模型的潜在风险。这个实际例子展示了AI MRM的真正价值——它将抽象概念转化为确保模型负责和安全部署的具体步骤的能力。

在我们深入案例研究之前，先回顾下面的图表，它描述了场景规划的整体流程。

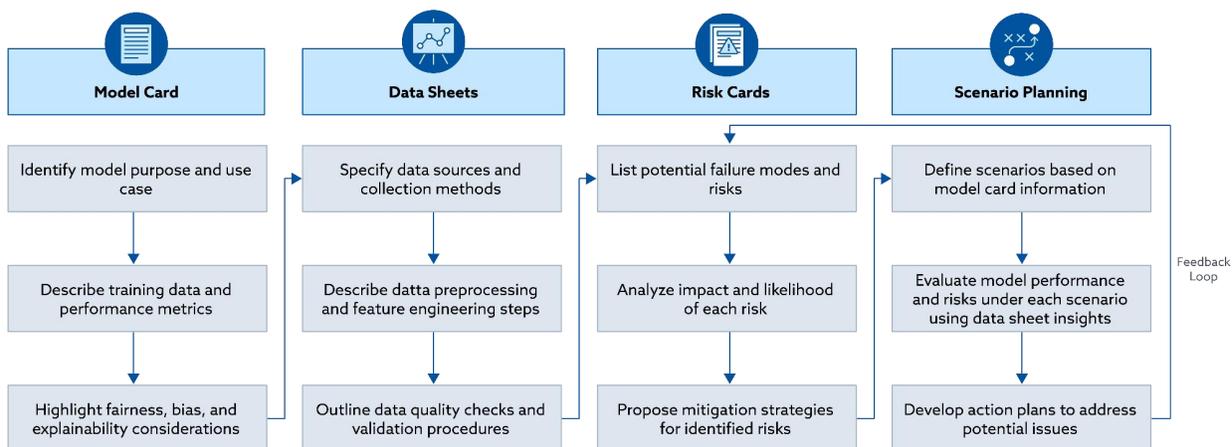


图3. 使用模型卡片、风险卡片和数据手册进行场景规划

## 社交媒体内容审核的大型语言模型

这个案例研究探讨了使用大型语言模型(LLM)进行社交媒体内容审核的潜在风险和机会，并利用模型卡片、风险卡片和数据手册进行场景规划。

**注意：**这里展示的模型卡片、数据手册和风险卡片是为了说明目的而进行的简洁总结。在实际应用中，这些文档会更加全面，包含详细的信息。

### 模型卡

模型卡片揭示了模型的能力、局限性和潜在偏见。它充当用户指南，概述了模型在社交互动方面的优势，并强调了由于训练数据中的潜在偏见或局限性而可能需要谨慎的领域。让我们为内容审核大型语言模型创建模型卡片。

**模型名称：**社交达人- 内容审核大型语言模型

**日期：**此文档中的信息截至2024年4月1日是最新的，除非下面另有说明。

**模型目的：**“社交达人”旨在分析社交媒体内容并识别可能违反平台政策的行为，包括仇恨言论、虚假信息和骚扰。它通过标记需要审核的内容来协助人类审核员。

**模型输入：**“社交达人”接收来自社交媒体帖子、评论和消息的文本数据。

**模型输出：**预训练的大型语言模型为每条内容分配一个风险评分，表明其违反平台政策的可能性。

**模型训练数据：**“社交达人”在大量标记过的社交媒体内容数据集上进行训练，包括违反政策的示例和可接受的内容。这些数据不断更新，以反映不断演变的语言模式和文化细微差别。

**性能指标：**“社交达人”的性能通过准确性（正确识别违规行为）、精确度（避免误报）和召回率（捕获大多数违规行为）等指标进行评估。

## 数据手册

数据手册提供了对训练模型所用数据集的透明视角。它们揭示了数据的来源、特征和规模，使人能够理解塑造“社交达人”响应的基础。下面列出了内容审核大型语言模型的两张数据手册。

### 数据手册1：社交媒体政策指南

**日期：**此文档中的信息截至2024年4月1日是最新的，除非下面另有说明。

**描述：**这张数据手册概述了特定社交媒体平台的社区指南和内容审核政策，LLM 被训练用来识别违反这些政策的行为。

**用例：**使LLM能够识别并标记违反平台规则的内容，促进安全和包容的在线环境。

**来源：**来自主要社交媒体平台（例如，Facebook、Twitter、YouTube）的公开可用的社区指南和内容审核政策。

**特征：**概述禁止内容类别（例如，仇恨言论、欺凌、骚扰）的结构化数据，以及具体示例和定义。数据规模取决于平台，通常从数万字到数十万字不等。

### 数据手册2：文化细微差别和上下文

**日期：**此文档中的信息截至2024年4月1日是最新的，除非下面另有说明。

**描述：**这张数据手册包括特定于不同文化和地区的语言示例，以帮助LLM区分真正的仇恨言

论、讽刺和文化表达。

**用例：** 这些数据提高了LLM理解上下文并避免基于文化背景的误解的能力。

**来源：** 策划的文本和多媒体内容集合，代表不同的文化和地区。这包括来自当代美国英语语料库（COCA）的文本，可能包括新闻文章、社交媒体对话、文学作品和文化参考。

**特征：** 文本数据带有文化背景标记，识别幽默、讽刺、成语和特定于不同地区的表达。规模：10亿词的文本数据，附有文化注释（截至2024年2月1日）。

## 风险卡

根据“社交达人”模型卡片和概述其训练数据的数据手册，已经开发了一套风险卡片，以主动识别潜在问题。这些风险卡片深入探讨了“社交达人”的输出可能被误解或滥用的场景。

风险#	名称	描述	影响	可能性	潜在影响	缓解措施
1	训练数据中的偏见	训练数据中的偏见可能导致LLM过度标记来自某些群体或观点的内容。	高	中	不公平的审查、用户信任的侵蚀以及潜在的法律问题。	使用多样化的数据源进行训练，实施偏见检测算法，并在审核过程中引入人工监督。
2	错误信息和细微差别	LLM可能难以区分讽刺、挖苦和真正的错误信息，导致标记不准确。	高	高	对合法内容的审查和阻碍健康的在线讨论。	训练LLM以识别上下文和风格线索，开发机制以便人工复审具有细微差别的标记内容，并公开LLM的局限性。
3	不断演变	LLM可能无法跟	高	高	遗漏违规行为	通过新示例不

	的语言和仇恨言论	上在线语言不断演变的特性，包括新形式的仇恨言论或编码语言。			和平台上仇恨内容的增加。	断更新训练数据，开发算法以检测新出现的语言模式，并利用人类专业知识来识别新形式的仇恨言论。
--	----------	-------------------------------	--	--	--------------	---

## 场景规划

设想“社交达人”在现实世界情境中的互动。这一部分探索了一些场景，以观察模型可能的反应。

### 场景1：有效审核（广泛采用 + 降低风险）

**描述：**“社交达人”有效地协助人类审核员识别和移除有害内容，从而营造一个更安全、更具包容性的在线环境。实施的保障措施最小化了偏见，并确保了LLM的负责任使用。

**好处：**提高内容审核效率，减少用户接触有害内容的机会，以及更积极地在线体验。

**挑战：**不断适应LLM以应对不断演变的语言模式和在线趋势。确保能够获取足够的高质量训练数据，以保持模型的有效性。

**总结：**“社交达人”，作为一个大型语言模型（LLM），可以协助人类审核员进行内容审核。然而，训练数据中存在偏见的风险，可能导致不公平的内容标记。为了减轻这一风险，将使用多样化的数据源和偏见检测算法来训练LLM。此外，将在审核过程中保持人工监督。虽然“社交达人”有潜力提高在线安全，但解决偏见并确保负责任地使用对于其成功至关重要。

### 场景2：放大偏见（训练数据中的偏见 + 有限的监督）

**描述：**训练数据中的偏见导致不公平的内容审核，不成比例地针对特定群体。有限的人工监

督使得带有偏见的标记得以放任。

**潜在后果：**侵蚀用户信任，引发审查制度的指责，声誉损害，以及潜在的法律后果。

**缓解策略：**彻底审计训练数据以查找偏见，增加关于LLM局限性的透明度，以及对所有标记内容进行强制性人工复审。

**总结：**“社交达人”在内容审核方面虽然有价值，但面临放大偏见的风险。有限的人工监督可能导致训练数据中的偏见未被检查，从而导致对某些群体的不公平内容标记。需要对训练数据进行彻底的偏见审查，透明公开LLM的局限性，并对所有标记内容进行强制性人工复审，以解决这一问题。

## 五、结论与展望

通过结合模型卡片、数据手册、风险卡片和场景规划，我们可以建立一个全面的模型风险管理（MRM）框架。该框架确保负责任的开发，减轻了偏见和数据质量问题等风险，并实现模型的安全和有益使用。优先考虑自动化和标准化工作将提高框架的效率，实现无缝集成，并提供汇总性能报告。这种积极主动的方法有效地管理了模型风险，并与人工智能/机器学习创新保持同步。

### 展望未来：模型风险管理（MRM）的演变趋势

人工智能和机器学习领域不断演进，要求我们对模型风险管理（MRM）最佳实践进行适应和完善。为此，我们将扩展本文的内容，提供实践经验、见解，并帮助有效实施这些实践。我们还将探讨以下新的关键领域，旨在扩大我们对全面模型风险管理（MRM）的理解：

- **标准化文档：**为模型卡片、数据手册和风险卡片开发一致的格式将简化不同模型之间的比较，便于进行风险评估，并使我们能够更全面地了解模型的功能和局限性。
- **机器学习运维（MLOps）和自动化的兴起：**机器学习运维（MLOps）领域正日

益受到关注，该领域专注于机器学习（ML）的开发和运维（DevOps）实践。自动化工具被纳入模型开发生命周期中，实现了持续监控和风险评估。这一转变有助于在模型部署到生产环境之前识别和应对风险。

- **集成可解释性人工智能(XAI)技术：**可解释性人工智能(XAI)技术能够更深入地洞察模型的决策过程，从而进一步加强风险识别和缓解工作。

- **监管环境的发展：**围绕人工智能（AI）/机器学习（ML）模型的监管框架仍在制定中。行业、监管机构和决策者之间的持续合作对于建立既促进创新又降低风险的明确有效监管至关重要。

- **解决社会和伦理问题：**随着人工智能（AI）/机器学习（ML）模型的日益普及，持续解决围绕偏见、公平和问责制的潜在社会和伦理问题非常关键。将这些考虑因素集成到MRM框架中尤为重要。

- **注重人机协作：**随着人工智能（AI）模型越来越多地融入决策过程中，重点将转向人机协作。风险管理策略必须不断发展，以考虑可能影响模型输出的人为错误或偏见。

通过积极采用框架的方法来管理模型风险，我们可以充分发挥人工智能（AI）/机器学习（ML）模型的潜力，并确保它们安全、负责地融入未来的创新中。

## 参考文献

- McKinsey & Company. (2023). *The state of AI in 2023: Generative AI's breakout year*. McKinsey & Company. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-AIs-breakout-year>
- IBM. (n.d.). *Watsonx AI*. IBM. <https://www.ibm.com/products/watsonx-ai>
- CVEdetails.com. (2024). *Microsoft Exchange Server Remote Code Execution Vulnerability (CVE-2024-26198)*. CVE details. <https://www.cvedetails.com/cve/CVE-2024-26198/>
- Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing language model deployment with risk cards*. arXiv.

- <https://doi.org/10.48550/arXiv.2303.18190>
- Derczynski, L. (n.d.). *Language model risk cards: Starter set*. GitHub.  
[https://github.com/leondz/lm\\_risk\\_cards](https://github.com/leondz/lm_risk_cards)
- AI Model Cards 101: An Introduction to the Key Concepts and Terminology:  
<https://www.nocode.ai/ai-model-cards-101-an-introduction-to-the-key-concepts-and-terminology/>
- Template for Model Cards:  
<https://github.com/fau-masters-collected-works-cgarbin/model-card-template?tab=readme-ov-file>
- Model Cards for Model Reporting: <https://arxiv.org/abs/1810.03993>
- Google Cloud Model Cards: [modelcards.withgoogle.com](https://modelcards.withgoogle.com)
- GPT-4 System Card by OpenAI: [gpt-4-system-card.pdf \(openai.com\)](https://openai.com/research/gpt-4-system-card)
- Gemma Model Card: [Gemma Model Card | Google AI for Developers](https://ai.google.dev/gemma-model-card)
- Model Card for Claude 3 family of models: [Model Card Claude 3.pdf \(anthropic.com\)](https://ai.google.dev/vertexai/docs/generative-ai/model_cards/claude-3)
- Model Card for VAE (dVAE) that was used to train DALL·E:  
[https://github.com/openai/DALL-E/blob/master/model\\_card.md](https://github.com/openai/DALL-E/blob/master/model_card.md)
- Example Model Cards: <https://modelcards.withgoogle.com/model-reports>
- Meta, Model Cards & Prompt formats  
<https://llama.meta.com/docs/model-cards-and-prompt-formats/#model-cards-&-prompt-formats>
- WWT CISO 2024: Secure Your Future: A CISO's Guide to AI, World Wide Technology, 2024,  
<https://www.wwt.com/wwt-research/cisos-guide-to-ai>
- CNBC 2024: The biggest risk corporations see in gen AI usage isn't hallucinations, CNBC, 2024-05-16, <https://www.cnbc.com/amp/2024/05/16/the-no-1-risk-companies-see-in-gen-ai-usage-isnt-hallucinations.html>
- GRC-based Model Risk Management Technology Solutions: A tech-enabled service,  
<https://www.pwc.com/us/en/industries/financial-services/regulatory-services/model-risk-management-technology-solutions.html>
- Understand model risk management for AI and machine learning,  
[https://www.ey.com/en\\_us/insights/banking-capital-markets/understand-model-risk-management-for-ai-and-machine-learning](https://www.ey.com/en_us/insights/banking-capital-markets/understand-model-risk-management-for-ai-and-machine-learning)
- A FAIR Artificial Intelligence (AI) Cyber Risk Playbook,  
<https://www.fairinstitute.org/blog/fair-artificial-intelligence-ai-cyber-risk-playbook>

# 附录 1: 人工智能框架、法规和指南

本节列出了各种框架、法规和指导文件，这些文件有助于推动负责任的人工智能开发与实施。这些资源建立了最佳实践，概述了风险管理方法，并在人工智能的整个生命周期中促进道德考量。

## 1. 美国国家标准与技术研究院(NIST)网络安全框架(CSF) v2.0

**定义：** NIST网络安全框架 (CSF) 是一个自愿的、基于风险的框架，旨在指导组织改善其网络安全态势。它概述了五个核心功能：识别、保护、检测、响应和恢复。

**与人工智能的相关性：** 尽管NIST CSF并非专为人工智能设计，但其原则可以适应于管理人工智能系统相关的网络安全风险。这些风险可能包括数据泄露、人工智能模型篡改以及人工智能赋能应用程序中的漏洞。

**与模型风险管理 (MRM) 的关系：** NIST CSF通过为人工智能模型中使用的底层基础架构和数据提供安全保障的基础，来补充模型风险管理 (MRM)。人工智能中的有效风险管理需要强大的网络安全实践，而NIST CSF有助于建立这些实践。

## 2. 美国国家标准与技术研究院 (NIST) 的人工智能风险管理框架(AI RMF)(提案)

**定义：** 人工智能风险管理框架 (AI RMF) 是NIST提出一个框架，专门设计用于管理人工智能系统相关的风险。该框架仍在开发中，但旨在提供一种全面的方法来识别、评估、减轻和监控人工智能风险。

**与人工智能的相关性：** 人工智能风险管理框架 (AI RMF) 解决了在人工智能开发、部署和使用中风险管理的挑战。它为组织提供了一个结构化的方法，以确保其人工智能系统是安全、可靠和可信赖的。

**与模型风险管理 (MRM) 的关系：** 一旦人工智能风险管理框架 (AI RMF) 最终确定，它可能会成为人工智能模型风险管理 (AI MRM) 实践的基石。该框架建立在现有风险管理框架 (如NIST CSF) 的基础上，并针对AI系统的特定需求进行了定制。

## 3. ISO 27001:2022信息安全、网络安全和隐私保护信息安全管理体系要求

**定义：** ISO 27001是国际信息安全管理体系 (ISMS) 标准。它概述了建立、实施、维护和持续改进ISMS以管理信息安全风险的要求。

**与人工智能的相关性：** 与NIST CSF类似，ISO 27001为保护信息资产提供了基础，这对于依赖大型数据集的人工智能系统至关重要。通过实施ISO 27001控制措施，组织可以

保护用于训练和运行人工智能模型的敏感数据。

与模型风险管理(MRM)的关系:通过ISO 27001建立的强大国际信息安全管理体(ISMS)有助于减轻模型风险管理(MRM)中的数据相关风险。安全的数据处理实践对于防止数据泄露、未经授权的访问和人工智能模型中数据的篡改至关重要。

#### 4. ISO 42001:2023人工智能管理体系

定义: ISO 42001是一个相对较新的国际标准,旨在增强组织的韧性。它指导组织对破坏性事件进行识别、评估、理解、准备、响应以及从破坏性事件中恢复。

与人工智能的相关性:人工智能系统可能容易受到硬件或软件故障、网络攻击或运营环境意外变化等引起的中断影响。ISO 42001帮助组织建立弹性,以抵御中断并确保人工智能系统的安全和可靠运行。

与模型风险管理(MRM)的关系:通过弹性考虑纳入其中,ISO 42001通过确保框架能够适应可能影响人工智能系统的不可预见情况,进而加强模型风险管理(MRM)。

#### 5. 美国注册会计师协会(AICPA)的系统和组织控制SOC 2

定义: SOC 2是针对存储和处理客户数据的服务组织的一套审计程序。它专注于与安全性、可用性、完整性、保密性和隐私性相关的控制措施。

与人工智能的相关性:许多组织依赖基于云的人工智能服务。SOC 2报告确保这些服务提供商已实施适当的控制措施以保护客户数据。

与模型风险管理(MRM)的关系: SOC 2报告通过为第三方人工智能服务提供商采取的数据安全控制措施提供独立验证,有助于模型风险管理(MRM)。这种独立验证有助于组织评估这些服务的可信度,并减轻与数据共享相关的风险。

#### 6. 欧盟人工智能法案(2024年6月生效)

定义: 欧盟人工智能法案(AIA)是欧盟制定的一项法规,旨在解决与人工智能系统相关的风险,并为其开发、部署和使用建立法律框架。该法案根据风险级别对人工智能系统进行分类,并对高风险人工智能应用提出具体要求。

与人工智能的相关性: 欧盟人工智能法案(AIA)特别关注确保人工智能系统的安全性、透明度和可追责性,这对于在各个行业建立对人工智能技术的信任和信心至关重要。

与模型风险管理(MRM)的关系: 欧盟人工智能法案(AIA)为管理人工智能风险提供了监管基础,通过引入强制执行的风险评估、减轻和合规性的法律义务,来补充现有的模型风险管理(MRM)框架。

## 7. 经济合作与发展组织（OECD）的人工智能原则

定义：经济合作与发展组织(OECD)的人工智能原则是由40多个国家认可的国际标准。该原则旨在促进在社会和经济中负责任地管理可信赖的人工智能。它们关注创新和可信赖的人工智能，同时尊重人权和民主价值观。

与人工智能的相关性：该原则倡导设计尊重法治、人权、民主价值观和多样性的人工智能系统，并鼓励在人工智能系统中实现透明度和责任的披露。

与模型风险管理（MRM）的关系：经济合作与发展组织(OECD)人工智能原则，支持将伦理、社会和法律考量纳入人工智能系统的生命周期中。它通过指导组织应对更广泛的社会风险，并确保人工智能开发符合全球标准和价值观，进而增强模型风险管理（MRM）实践。

## 8. 公平的人工智能(AI)网络风险操作手册(FAIR- AIR方法手册)

定义：信息风险因素分析(FAIR™)一个国际标准的定量风险分析模型，用于信息安全和运营风险分析。FAIR-AIR可帮助您识别与人工智能相关的损失暴露，并在网络风险管理中，针对这一新类别做出基于风险的决策。

与人工智能的相关性：对人工智能模型或基于人工智能的系统进行定量风险评估是具有挑战性的。FAIR-AIR可以帮助应对这一新类别中具有挑战性的网络风险量化任务。

与模型风险管理（MRM）的关系：模型风险评估除了定性风险评估外，还可以采用定量方法。定量分析可以提供模型，以理解财务方面的风险，并与业务部门进行更好的沟通。



Cloud Security Alliance Greater China Region



扫码获取更多报告