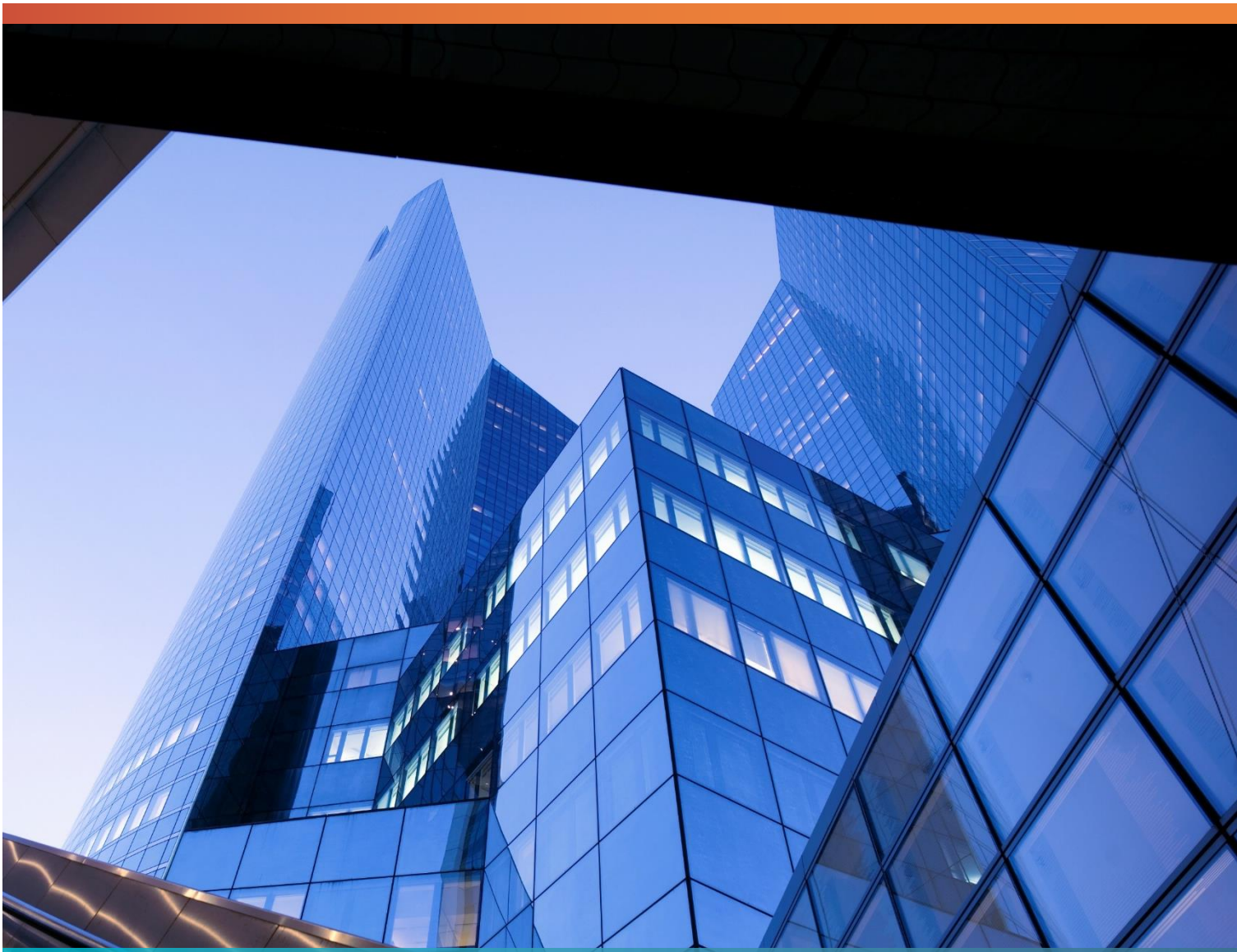
onetrust

protiviti®
Global Business Consulting
甫瀚



建立可扩展的人工智能治理框架：引领成功的关键路线和技术

目录

风险、合规与创新的平衡

在建立人工智能（AI）治理的运营模式（包括政策、开发标准和风险管理方法）时，组织常常面临一个经典的先有鸡还是先有蛋的两难困境。是应该在建立治理架构之前，首先识别和了解自身拥有的AI技术和用例，还是应该先建立治理机制，然后依据该架构确立AI用例呢?

答案取决于组织的成熟度和文化。无论如何，遵循本文概述的路线，将帮助组织明确识别、管理和衡量AI解决方案及其有效性的必要措施，以确保其在整体治理框架下持续发挥效能且风险可控。

第一部分：建立框架

如果组织决定首先建立治理框架，可以通过自问几个关键问题来启动这一流程。即使没有选择将建立治理框架作为第一步，无论在AI生命周期的哪个阶段需要正式制定治理框架，这些考虑因素都适用。

组织如何开始构建其AI资产清单？

创建AI系统和用例清单对于组织制定有效的AI治理策略和风险管理至关重要。这个过程使组织能够清楚地了解其现有的AI技术、正在进行的AI计划及未来规划。

一旦组织设计并实施了新的AI应用准入流程，并确定了其已拥有且需管理的AI技术，AI资产清单便可帮助其理解AI计划的规模、复杂性、预期成果及潜在风险。

以下是组织开始构建AI清单的步骤：

定义AI组件的术语及分类方式

目前业界对AI的定义缺乏共识，因此组织必须明确定义与AI相关的术语。这些定义需适应其独特的业务和风险环境，从而支持对AI技术（包括代表其使用AI的第三方）的一致性识别，并为风险决策提供依据。清晰的术语有助于系统化、统一地识别组织内所有AI应用，并纳入第三方风险管理。

在制定治理策略和方法时，组织需考虑其使命与目标、行业领域、风险容忍度、监管环境、技术现状等因素。明确AI使用范围并识别影响风险管理的AI属性，包括AI功能应用、数据来源、开源与闭源、第三方模型、AI类型（如生成式AI、机器学习、自然语言处理等），是建立AI资产清单的关键第一步。

识别目前及计划实施的AI系统和用例

术语定义完成后，需全面梳理组织内已使用、评估中或计划实施的AI实例。此过程可涵盖含AI组件的特定平台/应用、生成式AI模型及算法。

为此，需与组织内各部门（如IT、运营、销售、市场营销、人力资源等）的利益相关者协作，确保识别所有使用AI或受AI影响的业务职能。自动扫描工具和与机器学习技术栈的集成可辅助识别开发环境中的模型，并发现组织生态中的AI应用。

识别组织内正在使用的AI，还有一部分工作包括降低员工和合作伙伴使用“影子AI”的风险。具体而言，影子AI指那些未被IT和风险管理职能知晓、追踪及管理的AI。

终端用户或许在寻找和使用AI解决方案方面有较强的主观能动性，但这可能会引发组织知识产权泄露、AI及其输出结果滥用，进而引发法律及监管风险。

第一部分：建立框架

记录每个用例的关键信息

对于发现的每一项AI应用，组织需根据已制定的分类框架收集必要细节，包括但不限于：用途、能力、处理方式（是否使用生成式AI、自动化决策等）、输入数据、解决方案特性、输出结果（做出的决策或执行的任务）以及系统的训练方法。

组织还需收集开发者信息（内部团队或外部供应商）、用户信息（内部员工、客户、合作伙伴或其他外部利益相关者）、部署环境（本地/云端/混合）以及与AI资产清单对应的监管合规要求。

此外，对于每项AI应用，至少应描述其解决的问题，或是为组织创造的价值。若是较大规模的投资项目，还需有完备的商业计划书。

基于风险评级对AI用例进行分类

在记录AI用例及其业务案例后，建议组织根据风险评级对其进行分类。关键是尽早建立风险标准，以识别高风险用例，或可能使组织暴露于超出其风险偏好的潜在用例。

风险标准通常需法律、风险和合规部门直接参与制定，随后可通过以下因素确定风险等级：业务价值、固有操作风险、实施控制后的剩余风险、风险发生可能性、问题发生对业务运营的潜在影响、所用数据的敏感性、法律影响或声誉风险等。组织完成风险排序后，可优先对高风险用例进行审查和补救。

除了关注适用于自身的AI特定监管风险外，各组织还应确保其AI风险承受能力和风险分级系统与整体企业风险管理职能深度融合，这将有助于优先针对高风险用例开展治理工作。

维护动态清单

AI资产清单并非一次性任务，而应随着新用例的引入、现有用例的迭代或淘汰，以及已部署AI解决方案的调整持续更新。

这种方式能确保清单始终反映最新动态，并能够可靠的支撑持续治理与风险管理。每当发生实质性变更时，组织可启动TEVV（测试、评估、验证与确认）流程，从而深入了解特定AI用例的影响与风险，并将新信息纳入清单。

触发实质性变更的条件可预先定义，例如：AI开发团队提出的重大功能调整，或风险与合规部门识别的重大监管更新。此外，若某AI用例计划用于处理敏感信息，则应自动触发数据保护影响评估（DPIA）。

与现有技术清单整合

将AI资产清单与组织内现有技术清单（如核心IT资产清单、记录个人信息处理活动的业务流程清单）整合，以全局视角审视AI在技术生态中的定位。

构建清单往往需要大量跨职能协作，因此领导层的支持是成功的关键。

第一部分：建立框架

在制定治理原则时需考虑哪些因素，AI政策的核心组成部分有哪些？

为确保AI技术的部署和使用符合负责任和道德原则、法律与监管要求以及开发、部署和使用的领先实践（包括将风险缓解措施嵌入设计过程），组织需将若干关键考量纳入AI政策中。

AI政策的核心组成部分可归纳为以下要素，这些要素将为企业负责任地使用AI奠定坚实基础，这些要素不仅符合现行法规，还能通过践行道德承诺增强组织内外部利益相关者的信任：

AI原则

明确组织已采用的治理框架（如隐私、安全、数据治理）以及计划参考的AI治理框架（如《欧盟人工智能法案》、NIST人工智能风险管理框架（AI RMF）、微软负责任人工智能原则等）。

结合监管环境和行业标准，定义并记录与组织整体目标和倡议一致的AI原则。这些原则可能包括公平性、透明性与可解释性、问责制、安全性、尊重人权等。

政策声明

阐明组织对负责任AI原则的承诺，说明政策目标，并概述应对AI伦理、法律和社会影响的整体措施。

领导与治理架构

明确如何将AI原则落实到组织运营中，并详细规定相关角色与职责（包括各道防线如何协同识别和缓解风险）。

涵盖监督委员会或董事会的职能、全员及特定岗位（如IT部门、第三方）的教育与意识提升计划，及其在AI治理中的位置与作用。企业可选择将AI相关内容整合至现有政策中，或制定独立的AI专项政策。

合规与风险管理策略

制定合规和风险管理策略，以满足当前法规要求，识别并修复AI风险，保护输入/输出数据安全，并实施性能与合规监控机制。该策略还可用于开展组织内的AI培训、管理利益相关者参与，以及实现持续迭代。

纳入围绕数据隐私、安全、质量控制和同意机制的稳健实践，以确保组织从数据收集到处置的整个数据生命周期内的妥善管理。清晰描述组织确保遵守AI使用相关法律法规及行业标准的流程，包括定期合规检查或审计。

说明如何识别和缓解AI开发与部署中的潜在风险，讨论管理这些风险的策略（如技术保障措施、政策执行机制）。

以上要素相辅相成，共同构建一套全面的政策框架，既能指导内部运营，又能向外部利益相关者（如客户、监管机构）传递组织对负责任使用AI的承诺。

组织应该进行哪些测试和风险评估?

为了解自身的AI风险状况和风险水平，组织应该进行多种测试和风险评估。以下是一些关键的评估手段：

数据质量检查

实施强有力的数据治理实践，包括数据质量检查。不完整、过时、错误或存在偏见的数据可能导致AI系统做出错误决策，产生不准确、不完整或不恰当的响应。

AI风险评估

开展全面的风险评估，重点关注使用AI可能造成的潜在危害，并确保其与既定的AI原则保持一致。其中包括针对每个用例识别和评估相关风险，例如数据隐私或合规性问题、算法偏见导致不公平结果的可能性、安全漏洞等。

此外，该类风险不仅应从运营层面评估，还需考虑战略层面，例如因AI使用不道德或未遵守新兴法规而导致的财务或声誉损失。

更重要的是持续更新的风险评估，不仅针对特定用例（在立项、测试验证和部署后考虑可能发生的变化），还要从整个组织的风险偏好角度进行评估，也就是说，明确组织愿意承担的风险程度，以及期望通过设计去规避的风险类型。

这有助于组织在业务线选择与企业整体风险策略之间取得平衡，包括保护品牌声誉、股价及合规性。

隐私影响分析和数据保护影响评估

隐私影响评估（PIA）可帮助处理个人信息的组织确保遵守数据保护法规，同时识别和缓解与处理个人信息相关的风险。

类似地，数据保护影响评估（DPIA）可用于分析、识别并将与保护组织所处理的数据相关的风险最小化，无论这些数据是否包含个人信息。

AI红队测试

AI红队测试是一种合乎道德的黑客手段，通过模拟潜在对抗性攻击来评估AI系统的稳健性和安全性。测试需在受控环境中由网络安全和AI专家执行。此实践可帮助组织发现AI漏洞，从而理解真实威胁并改进现有缓解措施。

性能与合规性监控

建立持续的性能监控流程，以便于企业定期评估AI功能在部署后实际场景中的表现（与开发/测试阶段相比）。此外，也能通过报告AI不准确甚至有害的输出，为受AI影响的各方提供反馈机制。

同样，需建立合规性监控机制，以跟踪监管环境的变化，并在必要时触发对项目文档和流程的更新。针对AI用例的持续性能监控与合规措施，应根据准入阶段的风险评估结果来制定。

危机模拟演练

开展危机模拟演练（如桌面推演、预案演练等），通过模拟AI相关故障、系统被攻破或数据泄露等假设场景，测试组织有效管理此类事件的能力。此类演练可与企业现有的业务连续性和灾难恢复计划相结合。

组织的AI政策该如何落地?

将组织的AI政策付诸实践需要多个步骤，以确保政策不仅仅是纸上谈兵，而是真正体现在组织的行动和实践中。以下是实现这一目标的方法：

领导层承诺

任何政策要发挥作用，首要的是高层领导（C级别高管）必须做出明确承诺。这意味着领导者需要公开支持政策，对政策的实施设定期望，并确保自身及他人对政策负责且遵守。

治理机构和决策体系

组织通常设有执行委员会或风险委员会，负责在批准重大投资前对其进行评估。对于IT部门可自主决定的小规模投资，可由IT风险委员会等小型IT治理机构决策。此外，还可设立包含业务代表和风险管理职能在内的投资治理机构。

在IT职能之外，组织还需考虑其他控制职能的现有治理结构（如隐私、安全和数据治理）。许多组织选择成立AI委员会或AI工作组作为AI相关决策的核心团队。无论纳入哪些团队，治理架构的设计必须包含清晰的决策流程，以促进而非阻碍负责任的AI应用快速落地。

角色与职责

在确定相关治理机构后，组织应明确定义并记录清晰的RACI责任分配矩阵（谁负责、谁批准、咨询谁、通知谁）和决策流程，明确AI解决方案（包括第三方/及其使用的解决方案）的发起方、所有方、管理方和监督方的职责。

通常，这些治理机构分为两大类职能群体：一类从业务高层视角审查提案，另一类从技术层面评估功能实现和风险缓解措施。无论分工如何，必须确保所有环节通过审核，以确保AI解决方案由合适的人员持续管理。

第一部分：建立框架

政策宣贯

组织应通过内部通讯、会议、培训课程等形式，向全员清晰传达AI政策。确保所有人理解在使用或开发AI系统时的行为准则。同时，组织还应建立领导层定期审查政策的机制。

培训计划

组织应制定全面的培训计划，以教育员工按照组织政策中概述的要求负责任地使用AI技术。培训应包含在日常工作中如何将伦理原则应用在AI系统的实际案例。此外，需设置常规重训的触发条件，例如距离上次培训的时间间隔过长或考核成绩过低等。

反馈渠道

组织应建立员工及其他利益相关者的反馈渠道，以便其就政策实施提供意见或提出与AI使用相关的潜在伦理问题。组织应制定响应反馈的流程，并明确责任方。

迭代改进

组织应根据收到的反馈、技术或社会期望的变化、新法规等，定期审查并更新AI政策，确保其长期有效性和适用性。另外，考虑到技术上每月都会迭代上线更新更强的功能，因而组织需评估年度审查的频率是否足够，是否需要更频繁的审查。

外部沟通与报告

组织应通过年度报告或网站声明等方式公开承诺对AI的负责任使用。此举可增强客户、用户、监管机构等外部利益相关者的信任，展现AI技术应用的透明度。

此外，组织还可定期与客户、合作伙伴、监管机构等交流，分享负责任使用AI的经验，同时汲取外部视角。此类互动有助于优化组织的AI实践。

第一部分：建立框架

法律和通用框架提供了哪些指导？

当前国内外AI法律与框架涵盖以下关键领域：

透明度和可解释性

组织应尽可能提高AI系统的透明度，以用户可理解的方式解释其决策逻辑。AI系统不应以“黑盒”的方式运行，尤其是当决策对用户产生重大影响时，用户有权了解其运作机制。

数据隐私

保护AI系统使用的个人数据至关重要。组织需密切关注法规变动，以遵守像类似欧洲《通用数据保护条例》（GDPR）的数据保护法律或特定行业法规。这包括采取充分的数据保护（安全性与隐私性）措施以防范数据泄露、获取必要的数据使用授权、以及尊重用户对其个人数据的权利。

风险管理

组织应定期开展全面风险评估，识别部署AI技术可能带来的潜在风险，包括偏见或歧视等伦理风险、系统故障或网络攻击等技术风险、以及违反法律法规的法律风险等，并且制定相应缓解策略。

持续监控与改进

对已部署的AI系统，组织应持续地进行性能与合规性监控，即便在开发阶段已进行全面测试后，仍要留意任何意外情况，并确认所有控制措施和防护机制都按预期运行。组织应对问题、事件和错误进行追踪，并向相关利益相关者报告。若在欧盟境内运营，可能还需通知监管机构。

利益相关者参与和认知

应在组织内持续与利益相关者沟通负责任使用AI的重要性。通过知识共享会议，帮助员工和客户了解AI的部署方式及其局限性，从而增强对这些技术的信任，同时也有助于在出现问题时更有效地识别和报告。

第一部分：建立框架

基于现行法规和框架的AI治理通用控制与最佳实践

数据质量控制

包括测量输入与输出数据的准确性、及时性、完整性和偏差的流程与机制。

政策执行控制

组织应建立AI政策的执行机制，确保遵循已达成一致的AI原则；即使暂无违规处罚，也应致力于维护全球相关法律法规所倡导的核心目标。

安全与隐私控制

组织应通过安全与隐私控制措施识别风险与漏洞，同时确保在AI生命周期的每个阶段，数据都得到妥善保护，并且AI的使用符合伦理规范。具体措施可能包括监控未经授权的访问、信息滥用以及针对个人或组织的潜在危害。此外，还可以开展弹性测试，以评估AI系统的恢复能力，并确定是否需停止使用特定模型或用例。

监控控制

组织应建立AI监控机制，包括实际性能与预期性能对比、人工介入以及生成式AI的使用情况。

最重要的不仅仅是制定这些管控措施，还需确保它们在组织的各个层面——从战略规划到运营执行——都能得到有效实施。这既需要领导层坚定的决心，也需要全体员工积极参与，无论他们在公司架构中担任何种角色、履行何种职能。

第二部分：理解AI系统

一旦建立起有效的治理框架以更全面地覆盖AI应用场景，组织通常需要将这一框架转化为可落地的IT解决方案，使其具备可扩展性、将抓取的数据存储在便于维护和使用的系统中，并能实现关键流程的自动化。

AI治理项目所需的技术能力有哪些?

识别AI治理所需技术能力的计划将帮助您深入理解AI系统的技术细节，确保全面掌握所使用AI的特性并制定最佳治理方案。以下是需要构建的关键技术能力：

AI管理系统

人工智能管理系统在AI技术的开发和部署中至关重要。一个强大的AI管理系统应能支持AI系统的规划、开发、实施与部署全流程管理，包括跨多个风险领域的评估、识别、管理和追踪。以下是AI管理系统应包含的核心组件：

AI/ML风险识别与评估

通过自动化工具、规则逻辑和人工输入识别风险。评估完成后，可根据结果衡量每项AI风险的影响程度，并开始实施控制措施以降低风险。

AI/ML风险与控制知识库

该知识库包含基于全球AI法规与框架的风险清单及对应的缓解措施，可帮助组织在风险或违规事件发生时更高效地应对。

AI/ML风险与问题追踪

建立可记录的工作流程，用于分配任务并为已识别的风险收集证据。

AI/ML的可审计性与可追溯性

如前所述，透明性与可解释性是可信任AI和负责任使用AI的基础。AI管理系统需能够支持跟踪信息在系统中的流动路径，从而理解数据输入如何影响最终输出。

系统卡片

透明度的另一体现是使用简明语言描述AI系统的功能与运作方式。系统卡片以通俗易懂的语言解释AI系统的作用，使用户了解其用途并决定是否继续参与。

第二部分：理解AI系统

模型风险管理

模型风险管理指对机器学习模型的内部控制、审计、文档、政策及流程的管理。该过程需关注以下几个关键方面：

模型评估

这些评估对模型性能进行技术分析，以对照预定的性能、准确性、安全性和公平性标准。在AI生命周期中尽早且多次开展评估，有助于发现AI模型输出中存在的偏见或错误。

模型运维监控

这一过程包括监控机器学习运维（模型运维），以确保模型的性能可靠及合规。这需要对机器学习的模型进行监控，以检测模型退化、数据漂移、概念漂移等变化，并确保模型保持可接受的性能水平。监控类型包括被动监控、主动式监控、实时监控、日志监控、性能监控以及安全监控。

模型卡片

与系统卡片类似，模型卡片记录了模型的关键细节，包括用途、性能指标、训练数据、已知局限性或偏见。

对于高度依赖AI模型进行日常运营和决策的组织而言，建立有效的风险管理框架至关重要。实施这些控制措施有助于确保模型不会产生带有偏见、不准确或有害的输出。

AI应用安全

尽管AI系统对组织极具价值，但它们也确实带来了其独特的安全风险。以下是一些用于开发、测试和向应用程序添加安全功能的工具，以防止漏洞的出现：

对抗性防御

保护AI系统免受对抗性攻击的技术和方法，包括对抗训练、防御性蒸馏、梯度掩蔽、特征压缩、随机化变换、以及集成技术。

安全团队和TEVV（测试、评估、验证与确认）团队能够识别出默认情况下未知的额外漏洞，以便控制措施或风险缓解机制可以被应用、维护并监控。

例如，若您有一个系统，该系统有时可能因恶意篡改（而这种篡改难以通过自动化手段检测出来）而产生错误输出，引入人机协同控制（human-in-the-loop）可协助发现这种错误输出，并将其作为安全事件进行处理、追踪和管理。

内容异常检测

此过程指在数据集中发现明显偏离预期行为的模式或实例。尽早捕获此类异常有助于避免后续产生不良后果。

数据隐私管理

AI管理的核心在于数据管理。为此，对AI系统进行持续评估、监控和审计以确保其符合数据保护法规和伦理义务至关重要。以下是一些参考方法：

数据目录

创建数据资产清单，以便在AI应用场景中促进数据的发现、使用和保护。

数据血缘

追踪数据在其生命周期内的流动过程，包括来源、存储位置及任何形式的转换，并提供可视化展示以突出数据管道内部的上下游依赖关系。

数据政策执行

需建立机制以确保数据的机密性、完整性和可用性，并通过书面政策规范这些机制。

隐私增强技术（PETs）

隐私增强技术（PETs）是用于在存储、处理和传输过程中保护个人数据的工具，旨在降低数据使用风险。典型示例包括同态加密、联邦学习、假名化、差分隐私、以及合成数据。

构建稳健的AI治理体系不仅关乎合规性，更涉及战略性地管理风险并确保AI计划为组织创造实际价值。通过本文总结的路线（例如构建全面的AI资产清单、制定明确的治理原则、整合风险评估等），组织可建立一套兼具实用性和可扩展性的治理框架。这种方案不仅能规避潜在风险，还能充分助力组织释放AI潜力，确保AI技术部署是负责任的、透明的并与更广泛的业务目标保持一致。随着AI持续引发各行各业的变革，那些重视AI有效治理的组织将更有能力应对未来的挑战并探索新的机遇。

关于

关于OneTrust

OneTrust以负责任的方式释放数据和人工智能的全部潜力。我们的平台确保公司数据得到安全处理，赋能各组织在降低风险的同时负责任地推动创新。

OneTrust拥有涵盖数据与人工智能安全、隐私、治理、风险、伦理以及合规等方面的全面解决方案，能够实现数据团队和风险团队之间的无缝协作，以促成快速且值得信赖的创新。OneTrust被公认为信任领域的市场领导者，拥有 300 多项专利，为全球超过 14,000 家客户提供服务，客户覆盖行业巨头及小型企业。

关于甫瀚咨询

甫瀚咨询（上海）有限公司是一家具有全球视野的咨询机构。我们在中国开展业务至今已逾二十年，分别在上海、北京、深圳、成都和香港设有五个区域团队。依托甫瀚全球网络，我们能迅速汇聚甫瀚全球超过25个国家90个分支机构的资源与洞见，灵活调动最适合的专业团队为客户带来高质量的交付，并支持中国企业的海外拓展。

甫瀚咨询的业务遍及运营与财务管理绩效优化、风控与合规、内部审计、信息技术咨询、数字化转型，以及气候变化与可持续发展等领域。我们为中国各行业优秀企业、世界500强企业、全球各地资本市场的上市公司以及拟上市公司提供成熟及定制化的解决方案，亦为成长型企业提供陪伴式服务。

公司地址

北京
朝阳区建国门外大街1号
国贸写字楼1座718室
电话：(86.10) 8515 1233

上海
徐汇区虹桥路1号
港汇恒隆广场办公楼一座
2301+2310室
电话：(86.21) 5153 6900

深圳
福田区中心四路1号
嘉里建设广场1座1404室
电话：(86.755) 2598 2086

成都
锦江区红星路三段1号
国际金融中心1号办公楼25楼
电话：(86.755) 2598 2086

香港
中环 干诺道中41号
盈置大厦9楼
电话：(852) 2238 0499

protiviti®
Global Business Consulting
甫瀚

关注甫瀚咨询 获取更多资讯