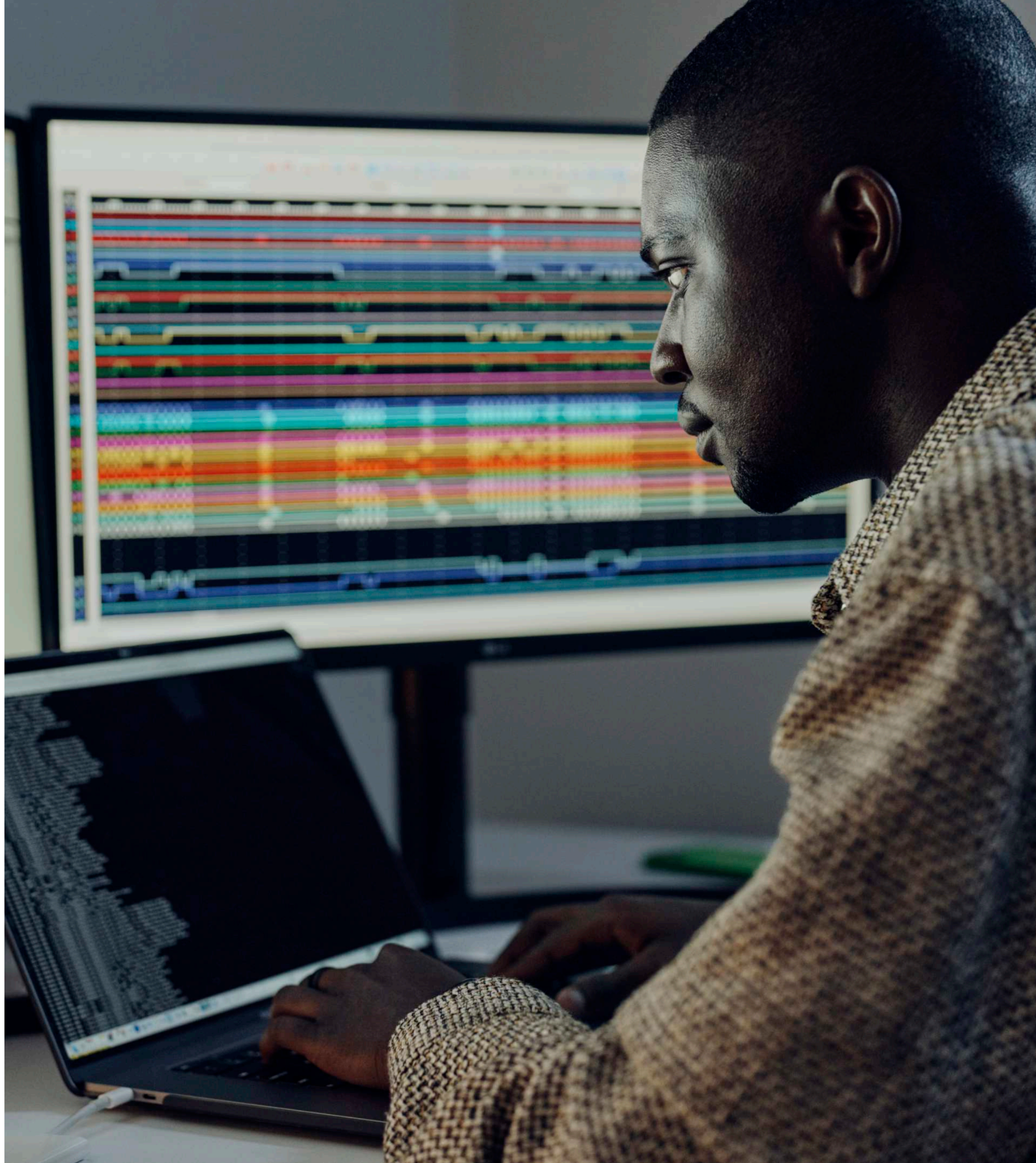
watsonx.governance™

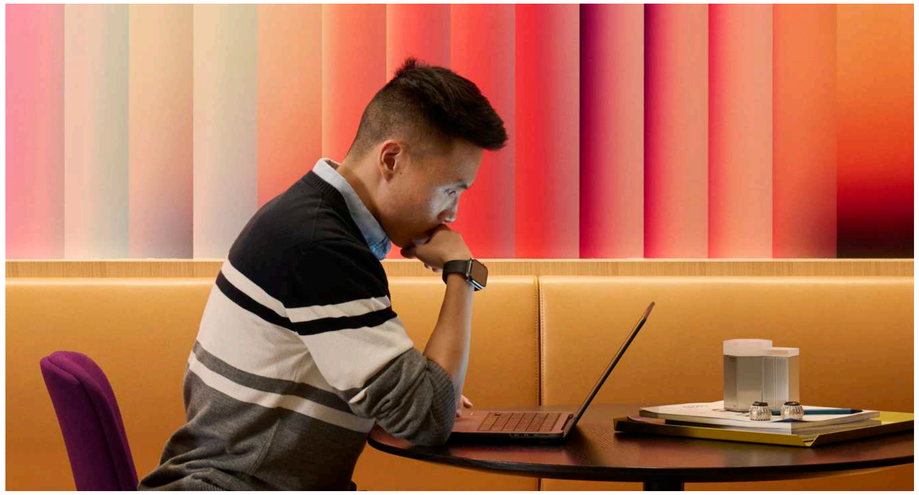
商用 AI：

通过更智能的治理、最大化 AI 投资回报率



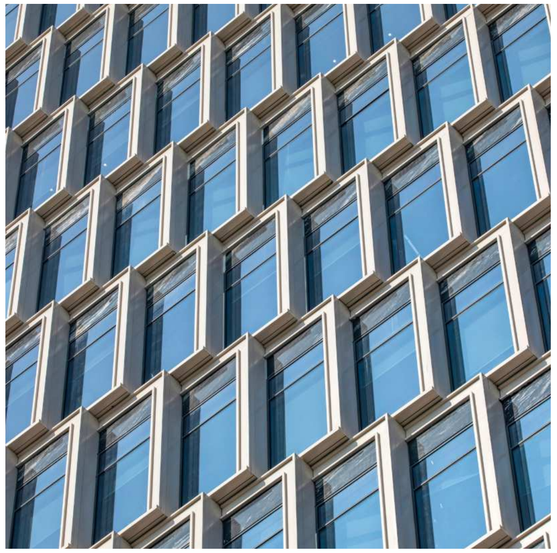
IBM

目录



01
简介

人工智能治理对于实现可扩展性至关重要



随着生成式 AI 成为新的现实、企业正通过人工智能驱动的创新来抢占先机。但其中的关键问题仍在于：您的 AI 是否得到了充分治理？

要回答这个问题、就需要将安全性和弹性通过设计融入组织的 DNA 中、而不仅仅是在政策中加以说明。为此、需要提供持续、可证明的证据、以证明控制措施正按预期运行、而不是仅仅依赖年度合规性检查。此类持续保证对于满足当今的监管要求以及应对不断变化的风险态势均至关重要。

治理是确保企业所有基于 AI 的创新理念都未偏离正轨、且符合全球道德与监管标准的关键所在。

有了治理作为安全网、就没有理由在充分发挥 AI 潜力方面退缩。

本电子书将帮助您了解更多有关人工智能治理的原则、并让您的企业走上发展的快车道。

阅读完整案例或免费试用 watsonx.governance →

与 AI 相关的风险正在上升——合规与监管问题、数据偏见与可靠性问题、以及当用户不了解 AI 模型如何运作或管理时导致的日益严重的信任缺失。

AI 智能体是未来的发展方向

随着数字化转型步伐的加快、企业正转而采用 AI 智能体、将其作为智能自动化的下一演进方向。根据 Gartner 的一项研究、83%[1] 的受访者预计 AI 智能体将在 2026 年之前提高流程效率和产出。而 IBM 的一项研究则表明、71%[2] 的人群认为 AI 智能体将自主适应不断变化的工作流程。

Gartner 预测、到 2028 年、至少 15%[3] 的日常工作决策将由智能体 AI 自主做出、而这一比例在 2024 年为 0%。

高级管理层的高管们承认、他们的组织需要做得更好。

60%

60% 的 CEO 表示、他们正在强制实施额外的 AI 策略以降低风险。[4]

63%

虽然 63% 的首席风险官和首席财务官表示他们关注监管与合规风险、但仅有 29% 的人群认为这些风险已得到充分解决。[4]

27%

约 27% 的上市公司在最近向美国证券交易委员会 (SEC) 提交的文件中指出、AI 监管存在风险。[5]

02

扩展 AI 的挑战

什么阻碍了组织的发展？两个字：信任。高管们认为、网络安全、隐私和准确性是实施生成式 AI 的最大障碍。随着态势的变化、他们预计会在未来 3 年内将对 AI 伦理的投资至少增加 40%[6]。

驾驭人工智能治理：当前的障碍

人工智能治理环境配备了一系列工具、但很多模型在开发过程中都面临透明度、一致监控和准确编目的问题。缺乏全面、自动化、端到端的生命周期管理系统、往往会妨碍可扩展性并导致操作不透明。对可解释的 AI 结果的追求仍难以实现、尤其是随着黑匣模型的兴起。此类模型虽被广泛部署、但往往会掩盖其输出背后的逻辑——甚至对于构建它们的开发人员来说也是如此。

缺乏治理可能会导致多种效率低下问题。例如、它可能会导致范围蔓延、阻碍模型的及时部署、致使模型质量参差不齐以及引发未识别的风险。鉴于 AI 开发和部署的复杂性、实施强大、透明和自动化的治理框架对于缓解这些潜在问题至关重要。

了解 IDC 对扩展 AI 的主要障碍的看法

立即阅读 →

在 AI 领域、驾驭风险与声誉管理的复杂态势可谓令人望而生畏。各种头条新闻持续渲染着不透明 AI 系统所存在的危险、即这些系统在现实场景中使用时可能会产生不公正、莫名其妙或带有偏见的结果。这些有缺陷的结果（通常会受到与种族、性别或年龄相关的隐藏偏见的影响）可能会产生深远的影响、从而既影响客户、又影响品牌的诚信度。

例如、以医疗保健等领域的高风险为例。影响患者诊断或治疗计划的 AI 系统必须具备透明性与公平性。不正确或存在偏见的 AI 建议可能会导致误诊或不当治疗、甚至可能产生危及生命的后果。

为了最大程度降低与 AI 相关的风险、组织必须致力于构建透明、公平且包容的系统。可解释的 AI 在检测和防止存在偏见的决策方面发挥着关键作用、同时还能增强隐私保护、安全性以及客户信任。构建值得信赖且无偏倚的 AI、不仅对于提升运营效果至关重要、同时对于避免争议和声誉损害也至关重要。

了解有关 AI 风险管理的更多信息 →

适应不断变化的 AI 监管环境
成功采用 AI 要求组织遵守快速发展的地方、区域及国家法律和法规。不合规可能会使您的组织面临数千万美元的罚款、[7]、目前全球范围内正在讨论的某些最严格的 AI 法规就证明了这一点。例如、《欧盟 AI 法案》的现行草案设想的罚款金额最高可达 3,500 万欧元、相当于某公司全球收入的 7%。[8]

模型文档非常重要、但在时间压力下、数据科学家经常忽视这一领域、尤其是在缺乏明确治理要求的组织中。

组织不能忽视此步骤；新法规将要求提供全面的模型文档、包括元数据和谱系。

了解如何简化AI 合规性

阅读我们的博客 →

新兴型

智能体 AI 固有的内在特性

风险
- 无监督自主行为
- 数据偏见
- 冗余操作
- 对 AI 智能体外部资源的攻击
- 工具选择幻觉
- 共享知识产权/个人身份信息/机密数据

挑战
- 可复现性
- 可跟踪性
- 攻击面扩大
- 有害且不可逆转的后果

放大后

智能体 AI 强化的已知问题

风险
- 行动目标偏差
- 歧视性行为
- 过度依赖或依赖不足
- 未授权使用
- 利用信任落差
- 无法解释或追踪的行为
- 透明度缺失

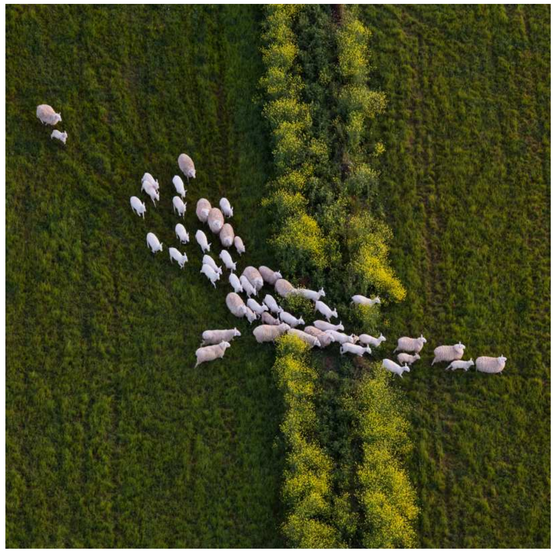
挑战
- 评估
- 问责制
- 合规性
- 风险控制与系统维护
- 无限反馈回路
- 共享模型缺陷

探索如何释放智能体 AI 潜能并管控风险

了解更多 →

03

所有 AI 均需进行治理



对于所有 AI（包括无人监督的智能体）而言、治理都是不可或缺的。尽管缺乏标记数据、这些智能体仍需接受监督、以确保其行为合乎道德、无偏倚、从而培育 AI 应用程序的信任度和可靠性。

以某一无监督的 AI 智能体为例、其任务是对客户进行细分以实现有针对性的营销。如果缺乏适当的治理、该智能体可能会无意中根据敏感属性（如种族或收入）对客户进行分组、从而导致潜在的歧视性行为。

治理措施可能包括：

1. **算法审计：**定期审查智能体的聚类流程、确保其不依赖于受保护的属性
2. **公平指标：**实施相关指标、评估智能体的输出是否存在偏见或歧视迹象
3. **人机交互：**包括人工监督、以验证并在必要时调整智能体的决策

通过引入这些治理策略、您可以减轻无意偏见的风险、并确保无监督 AI 智能体公平、有效地运行。

深入了解 IBM 如何帮助治理智能体 AI

了解更多 →

生成式模型
生成式 AI 模型包括基础模型 (FM) 和大型语言模型 (LLM)。此类模型有可能释放数万亿美元的经济价值、[9]、因为它们能凭借其出色的性能来提高生产力、并可适应各类任务。

此类模型具备高度可定制性、可扩展性和成本效益。它们能够查询大量数据、并可实时持续学习。现有的生成式应用程序所需专业知识较少、且有助于消除众多繁琐、耗时的任务。

在统计学领域、生成式模型长期以来一直用于分析数值数据。[10]但随着深度学习的兴起、其功能已扩展至包括图像、音乐、语音、视频、文本甚至代码的生成。当前用例已涵盖各个行业—从营销和客户服务到零售和教育。

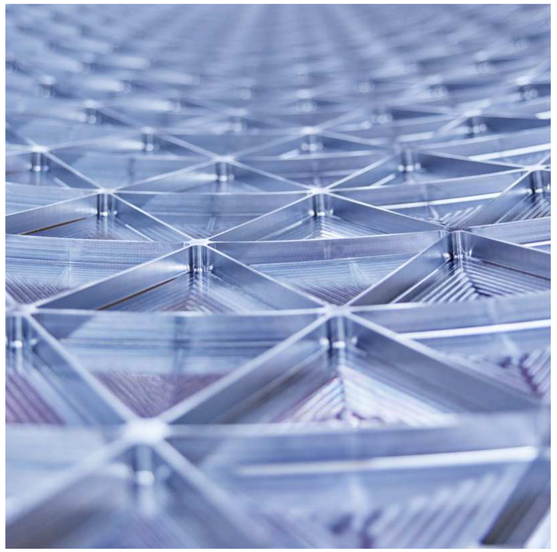
虽然生成式模型已将 AI 推至众多商业议程的首位、但其功能引入了新的复杂性、且可能对组织和社会等对象构成风险。

了解如何负责任地扩展 AI

阅读博客 →

04

全面的 AI 治理



像任何其他举措一样、成功的 AI 治理取决于人员、流程和技术的交互。

要正确实施 AI、您需要一个强大的跨职能团队。AI 正日益被众多领导者视为战略优先事项、而参与采用 AI 的团队成员数量似乎也在日益增长。其中部分人员可能对 AI 生命周期的概念尚不熟悉、而其他人员则正在寻找参与 AI 计划的新动力。满足所有这些群体的需求至关重要、同时不能给数据科学家带来过重的负担、因为他们通常没有足够时间安排或管理审批并响应信息请求。

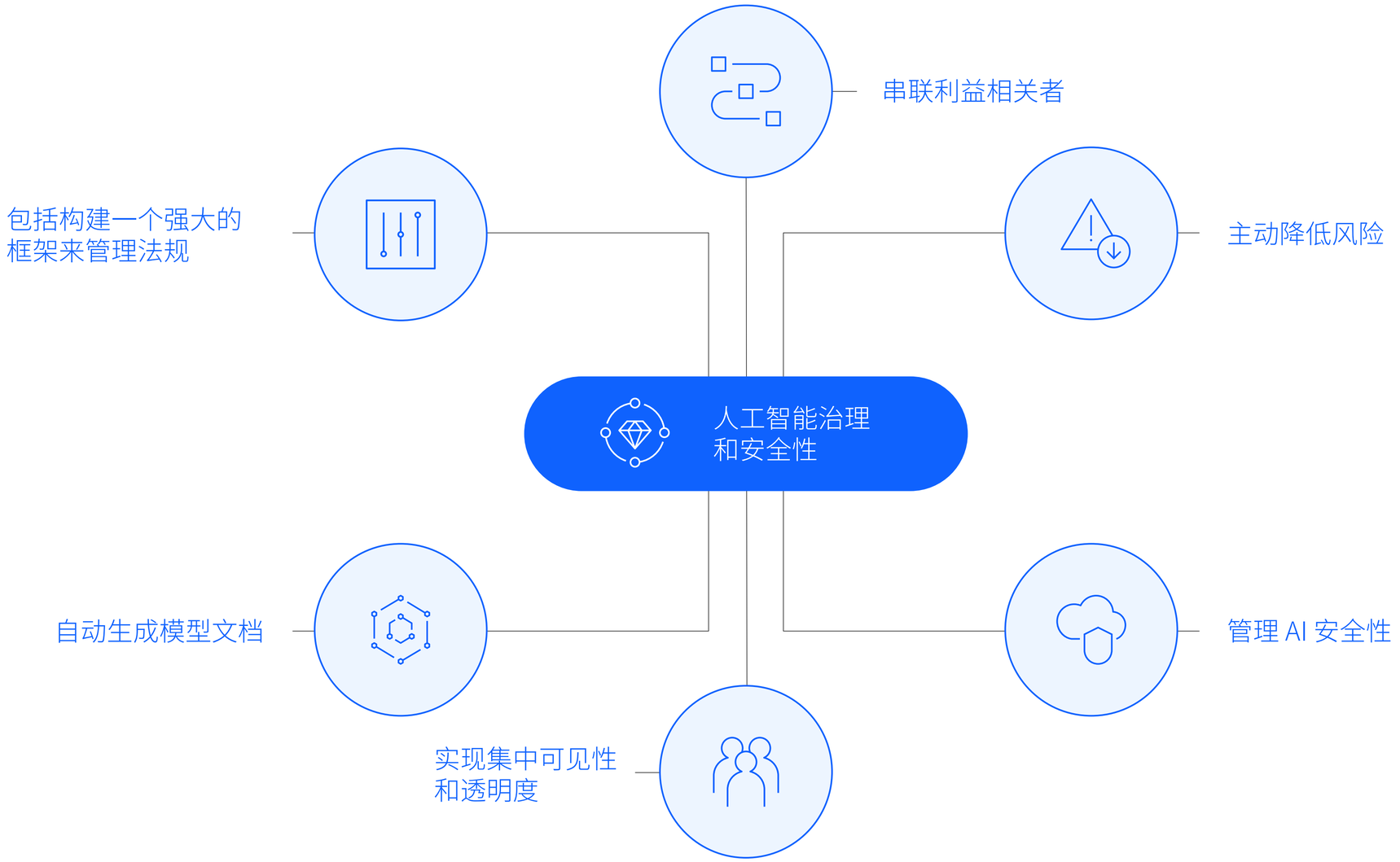
首先、应协调利益相关者并确保关键方的支持。然后、让他们参与构思并就成果和负责任的 AI 采用达成共识。

采取措施确保根据公司现有的业务控制措施和监管框架、定义正确的指标、关键绩效指标 (KPI) 和目标。最后、监控为您的 AI 模型确定的具体指标。

了解如何构建人工智能治理的整体方案

阅读博客 →

管理人工智能治理的复杂性



这些原则得到了信任支柱的支持、而这些支柱是我们 AI 伦理的基本特性。

可解释性
好的设计不会为了创造无缝体验而牺牲透明性。

公平性
经适当校准、AI 可以帮助人类做出更公平的选择。

稳健性
当系统被用来做出关键决策时、AI 必须是安全、稳健的。

透明度
透明度可以增强信任、提高透明度的最佳途径是信息披露。

公平性
AI 系统必须优先考虑和保障使用者的隐私和数据权利。

进程

人工智能治理涉及追踪并记录数据来源、关联的模型和元数据、以及用于审计的整体数据管道。您的文档应包括用于训练每个模型的技术、使用的超参数、以及各测试阶段所收集的指标。此详细程度可提高透明度、并让利益相关者了解模型在整个生命周期中的行为、包括影响其开发的数据以及可能产生的潜在风险。

首先、对您组织当前的 AI 技术和流程进行基准测试和评估。某些流程和利益相关者可能已协调一致并可进行扩展、而其他部分则可能需要更换。接下来、创建一组符合合规要求的自动化治理工作流程。

新的与现有的 AI 模型均可采用这些工作流程、以免出现上述流程延迟问题。最后、建立一个监控框架、以便在模型的指标超过可接受的阈值时提醒所有者和用户。

如何开始?

- 简化针对跨模型、应用程序和智能体的 AI 管理、监控与治理工作。通过主动识别偏见、偏差以及再培训需求、增强预测能力。提升资产质量、透明度和可解释性、同时降低风险。
- 提高 AI 运行速度、助力企业扩展运营和实现自动化、同时确保结果透明、可解释、无偏倚且无偏差。
- 运用风险管理手段、实现可扩展的风险识别、控制、跟踪与报告。通过模型再训练或重建、主动检测并修复偏见、偏差和行为转变。
- 同时评估多个 AI 资产、加快生产进程、减少开发人员和数据科学家的手动工作量。
- 通过预设警报、在未经授权的影子 AI 部署升级前将其标记出来、从而掌握安全漏洞、错误配置和风险指标情况。
- 使用预先配置的 AI 法规列表、简化合规流程、减少识别义务和管理不合规风险所耗费的时间。将外部法规转化为自动化执行、并通过事实说明文档强化审计和报告合规性。
- 使用自动化事实说明书记录数据集来源、模型元数据和管道信息。此自动化功能可收集模型事实、让数据科学团队能够腾出时间处理其他重要任务、同时为审计和诉讼提供支持。

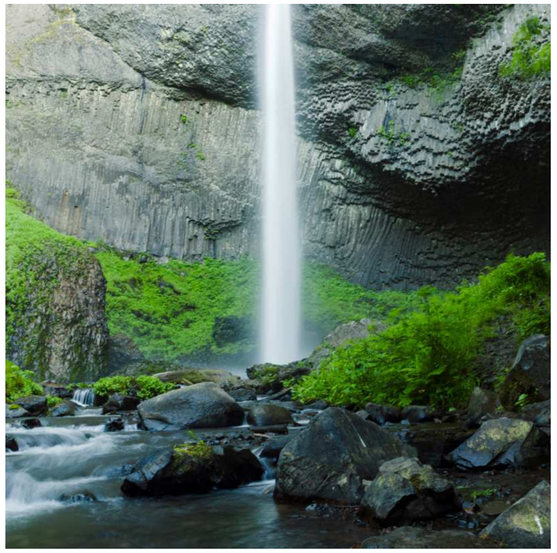
从 watsonx.governance 开始 →

负责任、受治理的 AI 框架

	透明度和可见性	AI 风险与安全管理	法规一致性	自动化
规划	为整个组织的 AI 使用情况定义可衡量的性能指标	审查用于监控公平性和可解释性的现有流程	根据当前与潜在的 AI 法规进行差距分析	审查针对负责任的 AI 的现有技能和需求、并使其与业务目标保持一致。
构建	构建当前流程的可追溯性和可审计性	在整个 AI 生命周期中有效运行经更新的流程和检查点	确保可访问模型文档记录	指定实施负责任的 AI 所需的新角色、技能和学习计划
创建	创建模型沿袭和元数据的自动文档记录。跟踪生产期间智能体的行为、识别异常并评估性能指标	实施自动触发警报、检测关键 AI 风险指标（如模型偏差、幻觉和影子 AI）、以主动缓解潜在风险和安全威胁	无需额外开销、即可增强数据科学团队的监管合规性	建立可重复的端到端工作流程、并内置利益相关者审批程序、以降低风险并扩大规模

05

watsonx.governance 可实现负责任、透明且可解释的 AI



了解 IBM® watsonx.governance—它是一个强大的人工智能治理工具包，旨在指导、管理和监控您的 AI 计划，帮助您降低风险、履行合规义务，并最大限度地提高 AI 投资的投资回报率。

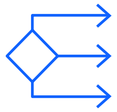
该工具包基于 IBM watsonx 构建，使用软件自动化增强法规遵从性，支持合乎道德的 AI 实践。它能提供全面的治理，无需进行昂贵的平台迁移。在预生产阶段，IBM watsonx.governance 可验证业务风险。部署后，它会持续监控公平性、质量和模型偏差，确保合规性。审计人员可访问模型行为洞察分析和预测解释，团队则可通过查看模型功能与训练细节获益。watsonx.governance 涵盖整个 AI 生命周期，借助集中的 AI 事实记录，为设计、开发、部署与监控方面的团队提供帮助。

它可通过跨数据、模型、元数据和管道的可追溯性简化审计，并记录有关训练技术、超参数和测试指标的信息。它有望增强对预计模型行为的透明度，深化对有影响力的数据的洞察分析，并主动识别风险。

IBM 在 2024 年 IDC MarketScape 全球机器学习运营 (ML Ops) 评比中被评为领导者。

阅读报告 →

请考虑以下组件：



法规一致性
利用自动化和智能简化 AI 监管合规流程：

- 构建透明的模型流程
- 访问单个监管内容存储库
- 将 AI 用例和项目与全球法规进行对标
- 加快监管义务的记录
- 改进合规评估周期
- 增强协作并减少手动处理时间

了解更多

风险与安全管理
主动检测和降低风险、同时监控公平性、偏见偏差以及新的 LLM 指标：

- 自动识别未注册的 AI 部署、并触发相应操作
- 了解安全漏洞、错误配置和风险指标
- 统一风险、合规性和安全利益相关者之间的安全策略创建流程

了解更多

生命周期治理
watsonx.governance 是一款开放且与平台无关的产品。您可以治理使用 IBM 或第三方平台（如 OpenAI、Amazon 等）构建和部署的任意 AI 模型、应用程序或智能体。

- 在单个实例中评估多个 AI 资产（模型、应用程序或智能体）
- 使用高级 RAG 指标简化智能体工具的选择、并监控智能体性能
- 实时跟踪 AI资产的整个生命周期

了解更多

无论您打算在企业的哪个阶段部署 AI、有效的治理都能通过将 AI 计划与企业目标相结合、推动实现跨用例的投资回报率 (ROI)。

改善客户体验
聊天机器人
借助 AI 防护措施、监控聊天中是否存在危险信号、例如恶意内容、个人信息泄露或偏离主题的回应

增强 BPO 流程
通过非干预监控、监测 Natural Language Understanding (NLU) 文本模型是否存在偏差、相关性等问题

避免不必要的合规成本
使用预构建的全球法规库执行运行合规流程、此类法规包括《欧盟 AI 法案》、ISO 42001、NIST AI RMF 等

构建基于 RAG 的高效机器人
使用内置的根本原因分析功能、分析 RAG 任务的提示模板评估结果

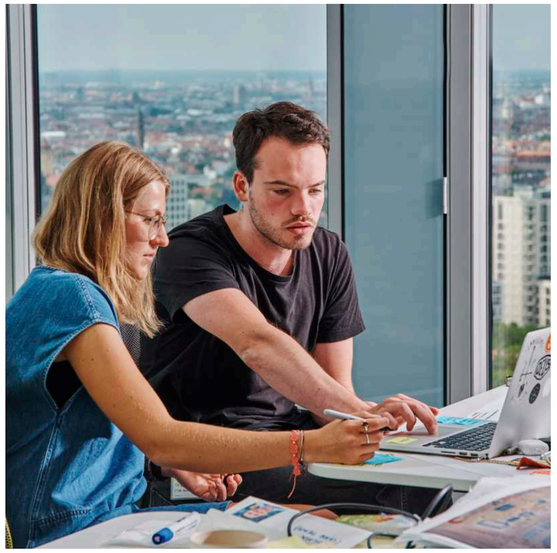
自动运行审计流程
通过自动捕获与您的业务相关的 AI 风险的详细背景信息、简化审计流程以满足监管要求

保护人力资源流程免受潜在业务风险的影响
检测并减轻用于招聘决策的 ML 模型中与您业务相关的 AI 风险的偏见

06

AI 治理实际应用

IBM 综合治理与市场就绪副总裁



通过集中式人工智能治理加速创新
基于 IBM 的信任和透明度基本原则、隐私与负责任技术办公室 (OPRT) 推出了隐私与 AI 管理系统 (PIMS)、以帮助您可靠地管理机器学习模型、遵守隐私与 AI 法规、并促进透明度和问责制。

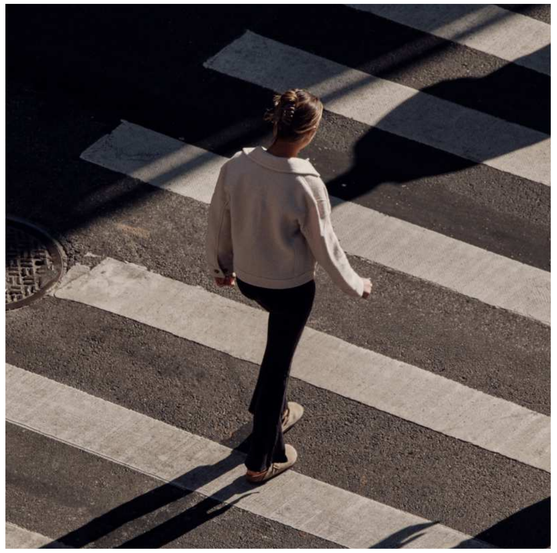
为了推进人工智能治理进程、OPRT 制定了综合治理计划 (IGP)、这是一项针对责任和合规的统一方法、其中集成了 watsonx.governance、IBM Cloud Pak for Data、IBM Knowledge Catalog 和 IBM OpenPages 等技术。

针对 IBM 数据和模型的这一整体视图可实现主动风险管理、法规遵从性协调、规模化治理工作流以及旨在实现透明度和值得信赖的 AI 的统一内部数据标准、并可最终产生以下成效：

- 第三方数据的数据放行审核申请处理时间缩短 58%
- IBM 专有数据的数据放行审核申请处理时间缩短 62%
- 经批准可重复使用的数据集和模型超过 1,000 个。[11]

07

后续步骤



了解使用 watsonx.governance 工具包创建负责任、透明且可解释的 AI 工作流程的快捷速度、而不会产生从当前数据科学平台进行切换所带来的成本。借助watsonx.governance、您可以：

- 管理风险并保护 AI 部署
- 在不断发展的 AI 监管环境中保持领先地位
- 提高 AI 用例的透明度和可见性、从而提高 AI 计划的投资回报率并加快上市时间

立即开始

了解 watsonx.governance →

预约演示 →

免费试用 →

1. Gartner 预测，到 2027 年底，将有 40% 以上的智能体 AI 项目被取消，Gartner，2025 年 6 月 25 日。

2. IBM 研究：企业将 AI 智能体视为必需品，而不仅仅是实验，IBM 新闻中心，2025 年 6 月 10 日。

3. 2025 年主要战略科技趋势：智能体 AI，Gartner，2024 年 10 月 21 日。

4. CEO 生成式 AI 行动指南：风险管理，IBM 商业价值研究院，2024 年 8 月 12 日。

5. 人工智能监管将至：财富500强企业严阵以待，《华尔街日报》，2024 年 8 月 27 日。

6. 生成式 AI：市场现状，IBM 商业价值研究院，2023 年 5 月 25 日。

7. 欧盟《人工智能法案》正式生效：违规或面临数千万欧元重罚，aiexpoeurope.com，2024 年 8 月。

8. 什么是《欧盟人工智能法案》（《欧盟 AI 法案》）？IBM，2024 年 9 月 20 日。

9. 生成式 AI 的经济潜力：下一个生产力前沿阵地，麦肯锡，2023 年 6 月 14 日。

10. 什么是生成式 AI？IBM Research，2023 年 4 月 20 日。

11. 通过集中式人工智能治理加速创新