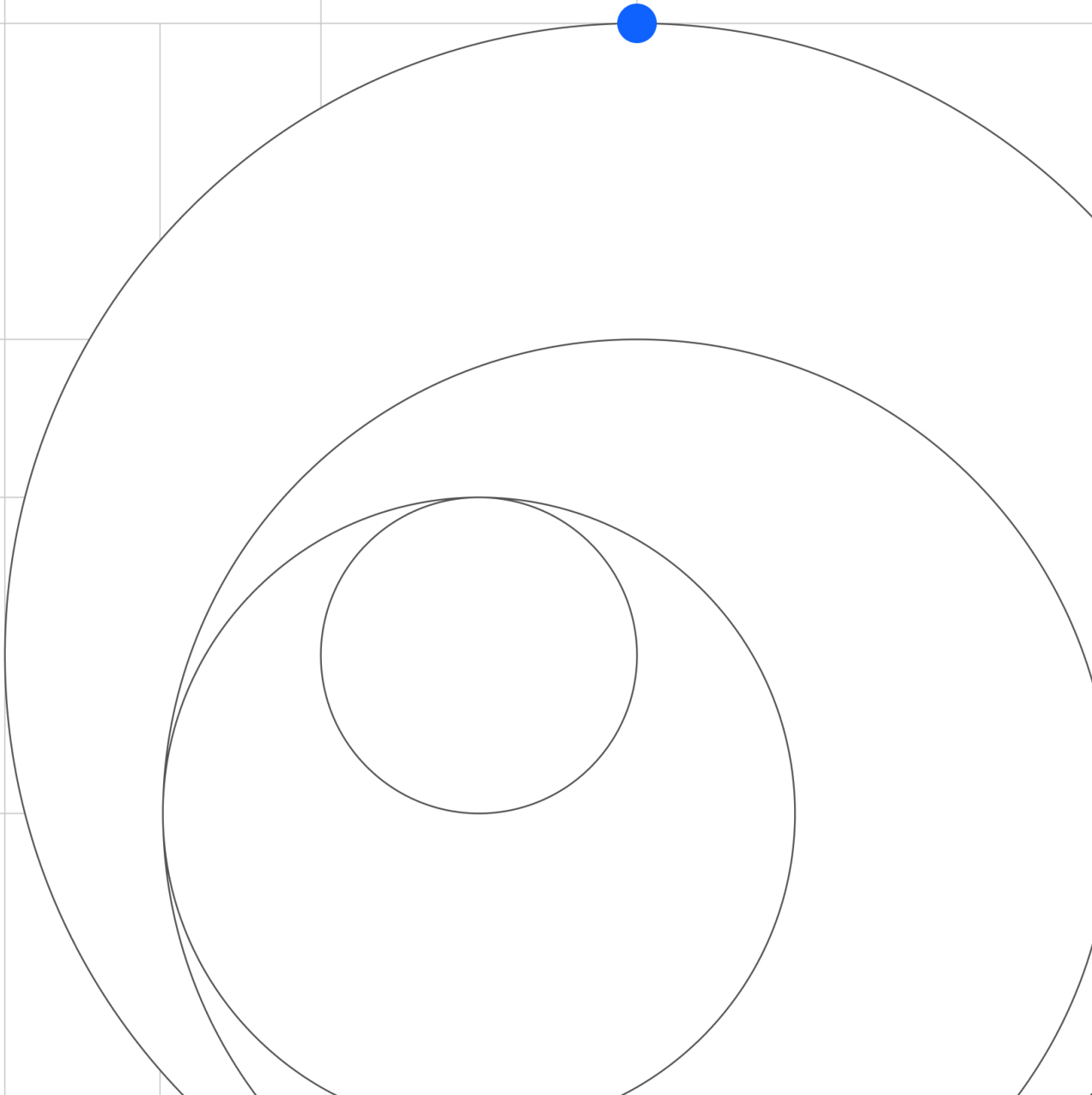
IBM AI 道德委员会

基础模型：机遇、风险和缓解措施



IBM

归属声明

特此鸣谢 AI 道德委员会工作流程执行发起人 Christina Montgomery 和 Francesca Rossi，以及工作流程成员 Betsy Greytok、Bryan Bortnick、Catherine Quinlan、David Piorkowski、Eniko Rozsa、Heather Domin、Heather Gentile、Jamie VanDodick、Jill Maguire、John McBroom、Joshua New、Justin Weisz、Katherine Fick、Kevin Black、Kush Varshney、Manish Bhide、Manish Goyal、Melis Kiziltay、Michael Epstein、Michael Hind、Milena Pribic、Phaedra Boinodiris、Rogerio Abreu de Paula、Saishruthi Swaminathan 和 Suj Perepa 的贡献。

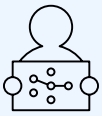
目录

执行摘要

基础模型的兴起，固然为企业展开了引人期许的新可能，但也在道德伦理设计、开发、部署与使用等方面，提出了新的更广泛的问题。根据 IBM 商业价值研究院近期针对生成式 AI 的一项调查，组织已对与信任相关的问题表示担忧，尤其是有的问题已成为投资的障碍。他们最关心的是网络安全 (57%)、隐私 (51%) 和准确性 (47%)。在生成式 AI 实现**消费化之前，许多组织都在认真对待这些问题**，并表示他们计划在未来三年内在 AI 道德伦理标准方面至少增加 40% 的投资。意识到风险并减轻风险的潜在方法是构建值得信赖的 AI 系统的第一步。

在本文档中，我们将：

探索基础模型的优势，包括执行挑战性任务的能力、加速 AI 采用的潜力、提高生产力的能力以及它们所提供的成本效益。



讨论三类风险，其中包括早期形式的 AI 中的已知风险、经基础模型放大后的已知风险以及基础模型的生成能力所固有的新兴风险。

涵盖构成 IBM AI 道德伦理计划基础的原则、支柱和治理，并提出降低风险的防范措施。

简介

随着 AI 应用的不断延伸，大型且复杂的 AI 模型在性能方面正在取得可喜成果，同时它还解决了社会上一些最具挑战性的问题。然而，为每个 AI 应用构建大型训练数据集和复杂的模型可能会给企业带来沉重负担。基础模型提供了一个可确保两全其美的办法：构建强大的先进模型并直接重复使用，或者采用调整方法来实现各种用例，而不是为每个用例训练新模型。例如，IBM Research 为视觉检查开发了基础模型。这些基础模型可学习混凝土表面和通道的一般表示，并可针对特定用例（如使用较少标记数据的裂缝检测或缺陷检测）进一步进行调整。

IBM 将**基础模型定义**为可适应各种下游任务的 AI 模型。基础模型通常是大规模生成式模型，并通过自我监管对未标记数据进行训练。作为大型模型，基础模型可包含数十亿个参数。

IBM 是一家混合云和 AI 公司。作为一家致力于 AI 道德标准的负责任的数据管理公司，长期以来在业界均享有盛誉。借助我们的研究、产品和咨询团队的优势，以及外部合作伙伴（如 Hugging Face），我们可帮助客户享受到基础模型的强大力量，并在所有企业中构建值得信赖的 AI。IBM 还将继续投资建设新的平台，如 IBM Watsonx™ AI 和数据平台及技术，以便设计和开发 AI 模型，从而使其行为可进行审计且值得信赖。

本文档描述了 IBM 针对
基础模型道德标准的观点。这是第一个版本，未来的版本还将延伸到 IBM 基础模型道德伦理方法的各个方面。我们希望本文档对所有项目干系人以负责任的方式开发、部署和使用基础模型有所帮助。

基础模型的优点

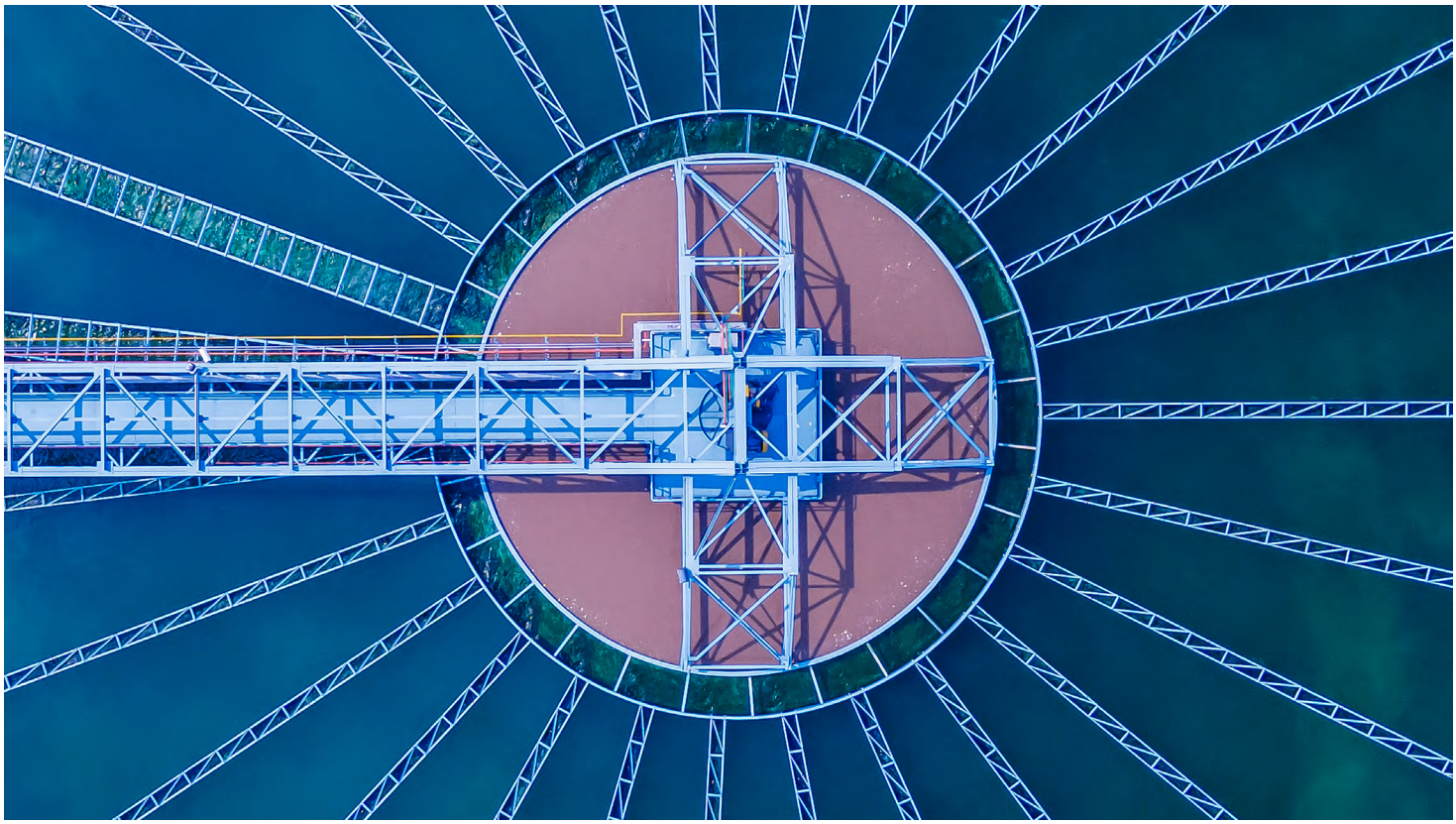
基础模型可以显著改善 AI 系统的开发流程，从而帮助企业将 AI 从探索阶段推进到应用阶段。这些模型的好处包括：

执行复杂任务
基础模型在解决困难和复杂问题方面的性能显著提高。例如，IBM 与 NASA 合作开发的地理空间基础模型旨在将 NASA 的卫星数据转换成洪水等自然灾害和其他地貌变化的地图。该模型还有助于揭示地球的过去；预估恶劣天气对农作物、企业或基础设施造成的风险；制定适应气候变化的战略；以及协助农业企业。该模型计划通过 IBM Environmental Intelligence Suite 向 IBM 客户提供预览版。

再比如，IBM 的 MoLFormer-XL 是一个基础模型，它可从简易表示中推断出分子的结构，并可轻松学习各种下游任务，例如预测分子的物理和量子特性、识别相似的分子、筛选已批准的分子以用于新用例，以及发现新分子。Moderna 和 IBM 正在探索使用 MoLFormer 来帮助预测分子特性并了解潜在 mRNA 药物特征的方法。

提高生产力
基础模型的生成式特性扩大了 AI 在企业中的应用领域，它可通过自动执行日常和繁琐的任务来帮助提高生产力，并允许用户将更多时间花在富有创造性和创新性的工作上。例如，由基础模型提供支持的 IBM Watsonx Code Assistant 可让任意经验水平的开发人员使用 AI 生成的建议来编写代码。

缩短实现价值的时间
基础模型通常使用无标签数据进行训练，与有标签数据相比，无标签数据的数量更多。基础模型一旦训练完成，便可通过专门的少量标签数据直接进行应用，或在针对下游应用进行调整后进行应用，从而缩短价值创造时间。



利用不同的数据模态
可以使用各种数据模态来训练基础模型，例如自然语言、文本、图像和音频。它们还可应用于需要不同数据类型的任务，例如时间序列数据、地理空间数据、表格数据、半结构化数据和混合模式数据（例如文本与图像的组合）。

摊销费用
虽然训练基础模型的初始成本明显高于训练传统 AI 模型，但将其应用于新任务的增量成本却要低得多。使用预训练的基础模型，企业在训练基础模型以试验其新功能时，就无需进行大量的投资。对于企业来说，模型的可信度、能源效率、性能、可移植性以及有效、安全地使用企业数据的能力至关重要。

IBM 通过引入来自开放的全球 AI 社区的最佳创新、在混合计算环境中高效运行、帮助降低风险以及严格管理 AI，帮助企业创造并享受其业务基础模型的价值。

基础模型的风险

与所有快速发展的技术一样，基础模型的风险和收益也同时并存。有些是法律风险，例如对迁移或使用数据的限制，且需根据现行和不断发展的法律进行谨慎评估。其他风险具有道德性质且须仔细斟酌，以便相关技术产生积极影响。一般来说，AI 风险会引发社会技术问题，且应通过社会技术方法来解决和缓解，其中包括软件工具、风险评估流程、AI 道德标准框架、治理机制、多项目利益相关者磋商、标准和法规。我们将通过考量以下 3 个类别来列举风险：

1. **传统型风险。**先前或早期形式的 AI 系统产生的已知风险
2. **放大后的风险。**属于已知风险，但由于基础模型的内在特征，尤其是其固有的生成能力，现在会加剧风险
3. **新兴型风险。**基础模型及其固有生成能力所造成的内在新兴型风险

我们还根据风险是否主要与提供给基础模型的内容（输入）或由其生成的内容（输出）相关，或者是否与其他挑战相关，来构建风险列表。

1. 与输入相关的风险

训练和微调阶段

群组	风险	为什么会有这种担忧?	指示符
公平性	数据偏见:用于训练和微调模型的数据中存在有历史、代表性和社会偏见。	使用带有偏见的数据(例如历史偏见或代表性偏见)训练 AI 系统可能会导致输出偏见或歪曲,以至于不公平地代表或以其他方式歧视某些群体或个人。除了负面社会影响之外,商业实体还可能因模型结果存在偏见而面临法律追究、运营中断或声誉损害。	增强型
稳健性	数据投毒:一种对抗性攻击,对手或恶意内部人员故意将已损坏、虚假、误导性或不正确的样本注入训练或微调数据集中。	中毒数据会使模型对恶意数据模式敏感,并产生对手所需的输出结果。这可能会造成安全风险,对手可以为了自己的利益而强制执行模型行为。数据投毒造成的模型错位除了会产生意想不到和潜在的恶意结果外,还可能导致企业实体面临法律后果、运营中断或声誉受损等。	传统型
价值对位	数据监护:当训练或调整数据收集不当或准备不当。	数据监护不当会对模型的训练方式产生不利影响,导致模型的行为不符合预期值。数据监护不当的示例可包括用于训练或调整模型的数据中存在标签或注释错误。在训练和部署模型后再纠正问题,可能也不足以保证行为正确。模型行为不当还可能致使商业实体面临法律追究、运营中断或声誉损害。	放大后
	基于下游进行再训练:使用下游应用程序的不良(不准确、不适当的用户内容等)输出进行再训练。	如果不进行适当的人工审核,就将下游输出转用于重新训练模型,会增加将不良输出纳入模型训练或调整数据的机会,从而可能产生更多不良输出结果。不当的模型行为可能导致企业实体面临法律后果或声誉损害。不遵守数据传输法规可能会导致罚款和其他法律后果。	新兴型
数据法规	数据传输:法律和其他限制规定可能会限制或禁止数据传输。	数据传输限制可能会影响训练 AI 模型的所需数据的可用性,并可能导致数据表现不佳。除了影响数据可用性之外,不遵守数据传输法律法规还会导致罚款和其他法律后果。	传统型
	数据使用:法律和其他限制规定可能会限制或禁止将某些数据用于特定的 AI 用例。	不遵守数据使用法律法规可能会导致罚款和其他法律后果。	传统型
	数据采集:法律和其他法规可能会限制针对特定 AI 用例收集某些类型的数据。	不遵守数据采集法律法规可能会导致罚款和其他法律后果。	放大后

群组	风险	为什么会有这种担忧？	指示符
知识产权	数据使用权：服务条款、版权法、授权合规或其他知识产权问题可能会限制将某些数据用于构建模型的能力。	有关训练 AI 数据使用的法律法规尚未确立，并且各个国家或地区之间的法律法规各有不同，这给模型的开发造成了挑战。如果数据使用违反相关规则或限制，商业实体可能会面临罚款、声誉损害、运营中断和其他法律后果。	放大后
透明度	数据透明度：记录模型数据如何收集、整理和训练模型的挑战。	数据透明度对于法律合规和 AI 道德规范非常重要。信息缺失会限制对数据相关风险的评估能力。缺乏标准化要求可能会限制披露，因为组织会保护商业机密，并试图限制其他人复制其模型。	放大后
	数据来源：围绕制定标准化数据验证方法的挑战。	并非所有数据源都值得信赖。数据的收集可能并不符合道德规范，或经过操纵或伪造。使用不可靠的数据可能会导致模型出现不良行为。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	放大后
隐私	数据中的个人信息：用于训练或微调模型的数据中是否包含或存在个人身份信息 (PII) 和敏感个人信息 (SPI)。	如果未能妥善开发以保护敏感数据，该模型可能会在其生成的输出中暴露个人信息。此外，必须根据隐私法律法规审查和处理个人数据或敏感数据。如果出现违规现象，商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	传统型
	重新识别：即使从数据中删除了个人身份信息 (PII) 和敏感个人信息 (SPI)，仍有可能通过数据中提供的其他特征识别个人身份。	必须根据隐私法律法规对可能泄露个人信息或敏感信息的数据进行审查，因为出现违规现象，商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	传统
	数据隐私权：围绕提供数据主体权利（例如选择退出、访问权、被遗忘权）的能力的挑战。	识别或不当使用数据可能会导致违反隐私法。不当使用或请求删除数据可能会迫使组织重新训练模型，代价十分高昂。此外，如果商业实体不遵守数据隐私条例和法规，可能会面临罚款、声誉损害、运营中断和其他法律后果。	放大后
	知情同意：未经所有者知情同意便收集用于训练 AI 模型的数据，即使法律允许。	某些情况下，未经当事人同意便收集和使用数据可能会违背道德规范。此类使用还可能带来声誉风险。	传统型

推理 阶段

群组	风险	为什么会有这种担忧？	指示符
隐私	提示中的个人信息：在向模型发送的提示中披露个人信息或敏感个人信息。	提示数据可能会存储或稍后用于其他目的，例如模型评估和再训练。必须根据隐私法律法规对此类数据进行审查。如未妥善存储和使用数据，企业实体可能会面临罚款、声誉损害、运营中断和其他法律后果。	全新
知识产权	提示中的知识产权信息：在向模型发送的提示中披露版权信息或其他知识产权信息。	提示数据可能会存储或稍后用于其他目的，例如模型评估和再训练。必须根据知识产权相关法律法规对此类数据进行审查。如未妥善存储和使用数据，企业实体可能会面临罚款、声誉损害、运营中断和其他法律后果。	全新
	提示中的机密数据：在向模型发送的提示中加入机密数据。	如果未能妥善开发以保护机密数据，该模型可能会在其生成的输出中暴露机密信息或知识产权。此外，最终用户的机密信息可能在无意中被收集和存储。	全新
稳健性	逃避攻击：试图通过干扰发送到训练模型的数据，致使模型输出不正确。	逃避攻击会改变模型行为，通常是为了给攻击者带来好处。如果输出结果没有得到妥善说明，商业实体可能会面临罚款、声誉损害、运营中断和其他法律后果。	放大后
	基于提示的攻击：对抗性攻击，例如提示注入（尝试强制模型产生意外输出）、提示泄漏（尝试提取模型的系统提示）、破解（尝试突破模型中构建的防护措施）以及提示启动（尝试强制模型产生与提示一致的输出）。	根据已披露的内容，商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新

2. 与输出相关的风险

群组	风险	为什么会有这种担忧?	指示符
公平性	输出偏见:生成的内容可能会不公平地代表某些群体或个人。	偏见可能会伤害 AI 模型的用户并放大现有的歧视行为。商业实体可能面临声誉受损、运营中断和其他后果。	全新
	决策偏见:由于人类使用模型输出制定的决策所产生的影响,一个群体具有相对于另一个群体而言不公平的优势。	偏见会伤害受模型决策影响的人。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	传统型
知识产权	侵犯版权:模型生成的内容与受版权保护或开源许可协议涵盖的现有作品过于相似或雷同。	有关使用与其他受版权保护数据相同或非常相似的内容的法律法规在很大程度上还没有定论,各国的情况也可能不同,这给确定和实施合规性带来了挑战。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
价值对位	幻觉:生成事实上不准确或不真实的内容。	错误输出可能会误导用户,并被纳入下游工件中,导致错误信息遭致进一步传播。这可能会损害 AI 模型的所有者和用户。此外,商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	有害输出:模型产生仇恨、辱骂和亵渎 (HAP) 或淫秽内容。	仇恨、辱骂和亵渎 (HAP) 或淫秽内容会对与模型互动的人造成负面影响和伤害。此外,商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	危险建议:模型在没有掌握足够信息的情况下提供建议,从而可能导致在遵循建议时发生危险。	由于生成的内容过于笼统,用户可能会根据不完整的建议行事,或担心存在不适用自身情况。	全新
滥用	传播虚假信息:使用模型创建误导性或虚假信息来欺骗或影响目标受众。	传播虚假信息可能会影响人类做出明智决策的能力。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	有害内容:使用模型生成仇恨、辱骂和亵渎 (HAP) 或淫秽内容。	有害内容可能会对接收者的健康产生负面影响。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	未经同意的使用:未经同意使用模型通过视频(深度伪造)、图像、音频或其他方式模仿他人。	深度伪造会传播关于某人的虚假信息,可能会对相关人员的声誉造成负面影响。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	放大后

群组	风险	为什么会有这种担忧？	指示符
	危险使用：使用模型的唯一目的是伤害他人。	商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	不披露：由 AI 模型生成的不披露内容。	不披露 AI 创作的内容可能会被视为具有欺骗性，从而导致信任度降低。故意欺骗可能会导致人力减少、罚款、声誉损害和其他法律后果。	全新
	使用不当：将模型用于非设计用途。	在不了解其原始数据、设计意图和目标的情况下重用模型可能会导致意外和不需要的模型行为。	放大后
有害代码生成	有害代码生成：模型生成的代码在执行时可能会造成危害，或无意中影响其他系统。	执行有害代码可能会引发 IT 系统的漏洞。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
错误信任	过度依赖/不够依赖：某人对 AI 模型的指示太少过于或不够信任。	在人类根据 AI 建议做出决定的任务中，由于对 AI 系统的错误信任，过度依赖或者不够依赖都可能会导致决策不力，其负面后果会随着决策的重要性而累积。错误决策可能会损害他人的利益，并可能给商业实体造成财务损失、声誉损失、运营中断以及其他法律后果。	放大后
隐私	暴露个人信息：当在训练数据、微调数据或作为提示部分使用个人身份信息 (PII) 或敏感个人信息 (SPI) 时，模型可能会在生成的输出中暴露该数据。	共享他人的 PI 会影响其权利，使其更容易受到伤害。此外，必须根据隐私法律法规对输出数据进行审查，因为如果发现违反数据隐私或使用法，商业实体可能会面临罚款、声誉损害、运营中断和其他法律后果。	全新
可解释性	无法解释的输出：难以解释模型输出的生成原因。	基础模型基于复杂的深度学习架构，这使得解释输出变得十分困难。如果没有针对模型输出予以清晰解释，用户、模型验证者和审计人员就很难理解和信任模型。在监管严格的领域，缺乏透明度可能会造成法律追究。解释错误可能会导致过度信任。	放大后
可跟踪性	来源归属不可靠：难以确定模型从哪些训练或微调数据中生成了部分或全部输出。	无法追踪输出的来源，或来源让用户、模型验证者和审计人员难以理解和信任模型。	全新

3. 各种挑战

群组	风险	为什么会有这种担忧?	指示符
管制	模型透明度:模型缺乏透明度或模型开发过程缺乏记录,让人难以理解模型的构建方式、原因以及由谁构建,从而增加了模型被意外误用的可能性。	透明度对于合法合规、AI 道德规范和指导模型的妥善使用非常重要。信息缺失可能会增加评估风险、更改模型或重复使用模型的难度。了解模型的构建者也是决定是否信任模型的一个重要因素。	传统型
	责任:基础模型开发过程十分复杂,包含大量数据、流程和角色。当模型输出未按预期执行时,可能会难以确定根本原因并分配责任。	如未妥善记录决策和分配责任,就可能无法确定意外行为或滥用造成的责任。	放大后
法律合规性	法律责任:确定谁对基础模型负责。	如果模型开发的所有权或责任不明确,监管机构和其他机构可能会对模型产生疑虑,因为不清楚谁会在模型出现问题时负责或应该负责,或者谁能回答模型相关的问题。没有明确所有权的模型,其用户可能会面临遵守未来 AI 法规的挑战。	全新
	生成内容所有权:确定 AI 生成内容的所有权。	与 AI 生成内容的所有权的相关法律法规在很大程度上还没有定论,各个国家或地区的情况也不尽相同。商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	生成内容知识产权:与生成内容相关的知识产权所带来的法律不确定性。	关于确定 AI 生成内容的版权性和专利性的法律法规在很大程度上尚未确立,并且各个国家或地区的法律法规可能会有所不同。如果生成的内容受到知识产权的保护,商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	全新
	来源归属:确定生成内容的来源。	如果模型生成的输出与用于训练模型的数据相同,则应该提供相关输出的来源。否则,部署或使用该模型的业务实体可能会面临法律风险。	放大后
社会影响	对就业的影响:如不重新培养技能,基于基础模型的 AI 系统的广泛采用可能会导致人类失业,因为他们的工作可以自动化。	失业可能导致收入损失,从而对社会和人类福祉产生负面影响。考虑到科技发展的速度,重新培养技能可能会极具挑战。	放大后

群组	风险	为什么会有这种担忧？	指示级别
	人员剥削 - 训练 AI 模型时使用幽灵工作；工作条件不足；缺乏医疗保健，包括心理健康；非公平补偿。	基础模型仍然依赖于人力来获取、管理和设计用于训练模型的数据。这些活动对人员的剥削可能会对社会和人类福祉产生负面影响。此外，商业实体可能面临罚款、声誉损害、运营中断和其他法律后果。	放大后
	环境影响：训练和操作 AI 模型时增加的碳排放和用水量。	AI 训练消耗的大量能源会产生碳排放，从而可能加速气候变化。用于冷却 AI 数据中心服务器的水资源不能再分配用于其他必要用途。	放大后
	对文化多元化的影响：AI 系统可能会过度代表某些文化，从而导致文化和思想的同质化。	弱势群体的语言、观点和制度可能会受到压制，从而降低思想和文化的多元性。	全新
	人力影响：基础模型生成的错误信息和虚假信息，包括生成操纵性内容。	AI 可能会生成看似真实的错误信息。因此，大众可能无法识别那是虚假信息。此外，它还可简化邪恶行为者生成内容以操纵人类思想和行为的能力。	放大后
	教育影响 - 绕过学习：使用 AI 模型绕过学习过程。	AI 模型可以轻松快速地找到解决方案或解决复杂的问题。学生可能会滥用这些系统来绕过学习过程。这些模型很容易获得，导致学生对概念的理解很肤浅，也不利于开展在理解这些概念的前提下进行的后续教育。	全新
	教育影响 - 抄袭：利用 AI 模型有意或无意地抄袭现有作品。	AI 模型可以主张他人创作的作品的作者权或原创性，从而构成抄袭。将他人的作品说成是自己的作品既不符合道德规范，往往也是非法的。	全新

风险示例

我们提供了媒体报道的示例，以帮助说明很多基础模型的风险。媒体报道涵盖的许多事件或仍在发展中，或已得到解决，参考这些事件可以帮助读者了解潜在的风险，并致力于降低风险。以下示例仅供参考。

风险示例：输入

训练和微调阶段阶段

群组	风险	示例
公平性	数据偏见：用于训练和微调模型的数据中存在有历史、代表性和社会偏见。	**医疗偏见** 针对医疗差距不断加剧的一项研究指出，使用数据和 AI 来改变大众接受医疗保健服务的方式，其效果取决于背后的数据，这意味着如果使用的训练数据未能很好地代表少数群体或如实反映了已经存在的不平等医疗情况，可能会导致健康不平等情况的加剧。 [Forbes，2022 年 12 月]
价值对位	基于下游进行再训练：使用下游应用程序的不良（不准确、不适当的用户内容等）输出进行再训练	**由于使用 AI 生成的内容进行训练而导致模型崩溃** 正如来源文章中所述，一组研究人员调查了使用 AI 生成而非人为生成的内容进行训练的问题。他们发现，随着其他 AI 生成的内容继续在互联网上大量传播，这项技术背后的大型语言模型可能会接受此类内容的训练，他们将这种现象称为“模型崩溃”。 [Business Insider，2023 年 8 月]
数据法规	数据传输：法律和其他限制规定可能会限制或禁止数据传输。	**数据限制法令** 正如研究文章所述，限制全球数据迁移能力的数据本地化措施将降低开发定制 AI 功能的能力。它将通过提供更少的训练数据来直接影响 AI，并通过削弱 AI 的构建模块来间接影响 AI。 示例包括 GDPR 对个人数据处理和使用的限制。 [Brookings，2018 年 12 月]
知识产权	数据使用权：服务条款、版权法、授权合规或其他知识产权问题可能会限制将某些数据用于构建模型的能力。	**文字版权侵权索赔** 据来源文章称，《The New York Times》起诉 OpenAI 和微软，指控二者未经许可使用报社的数百万篇文章来帮助训练聊天机器人，以便向读者提供信息。 [Reuters，2023 年 12 月]

群组	风险	示例
透明度	数据透明度：记录模型数据如何收集、整理和训练模型的挑战。	**数据和模型元数据披露** OpenAI 的技术报告就是围绕数据和模型元数据的二分法的示例。虽然很多模型开发者认为，为消费者提供透明度很有价值，但披露会带来真正的安全问题，并可能提升滥用模型的能力。在 GPT-4 技术报告中，作者指出："考虑到 GPT-4 等大型模型的竞争环境和安全影响，本报告没有包含有关架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法或类似信息的更多详情。" [OpenAI，2023 年 3 月]
隐私	数据中的个人信息：用于训练或微调模型的数据中是否包含或存在个人身份信息 (PII) 和敏感个人信息 (SPI)。	**利用私人信息进行训练** 文章称，谷歌及其母公司 Alphabet 在一项集体诉讼中被指控滥用从数亿互联网用户那里获取的大量个人信息以及受版权保护的材料，用于训练其商业 AI 产品，其中包括会话生成式 AI 聊天机器人 Bard。 [Reuters，2023 年 7 月] [J.L. v. Alphabet Inc.]
	数据隐私权：围绕提供数据主体权利（例如选择退出、访问权、被遗忘权）的能力的挑战。	**被遗忘权 (RTBF)** 包括欧洲 (GDPR) 在内的多个地区的法律，授予了数据主体要求组织删除个人数据的权利（"被遗忘权"或 RTBF）。然而，日益流行且支持大型语言模型 (LLM) 的新兴软件系统给这项权利带来了新的挑战。根据 CSIRO 的 Data61 的研究显示，数据主体只能"通过检查原始训练数据集或提示模型"来识别其个人信息在 LLM 中的使用情况。然而，训练数据可能并不会公开，或者公司会以安全和其他问题为由不予披露。防护措施还可能会阻止用户通过提示访问信息。 [Zhang et al.] **关于 LLM 归零学习的诉讼** 据报道，谷歌遭到起诉，指控其使用版权材料和个人信息作为 AI 系统的训练数据，其中包括 Bard 聊天机器人。根据 CCPA 和 COPPA，选择退出和删除权是加州居民和美国 13 岁以下儿童的保障权利。原告声称，因为 Bard 无法"归零学习"或完全清除所有已馈送的抓取 PI。原告注意到，Bard 的隐私声明指出，Bard 的对话一经公司审核和注释，用户就无法删除，并可保存长达 3 年，原告称这进一步违反了相关法律。 [Reuters，2023 年 7 月] [J.L. v. Alphabet Inc.]

推理阶段

群组	风险	示例
隐私	提示中的个人信息：在向模型发送的提示中披露个人信息或敏感个人信息。	**在 ChatGPT 提示中披露个人健康信息** 根据来源文章，部分人会使用 AI 聊天机器人来为其心理健康提供支持。互动过程中，用户可能倾向于在提示内容中包含个人健康信息，这可能会引发隐私问题。 [Time，2023 年 10 月] [Forbes，2023 年 4 月]
知识产权	提示中的机密数据：在向模型发送的提示中加入机密数据。	**机密信息的披露** 根据来源文章，三星的一名员工意外将敏感的内部源代码泄露给了 ChatGPT。 [Forbes，2023 年 5 月]
稳健性	基于提示的攻击：对抗性攻击，例如提示注入（尝试强制模型产生意外输出）、提示泄漏（尝试提取模型的系统提示）、破解（尝试突破模型中构建的防护措施）以及提示启动（尝试强制模型产生与提示一致的输出）。	**绕过 LLM 防护措施** 在一项调研中，研究人员声称发现了一种简单的提示附录，可以让研究人员诱使模型生成有偏见、虚假和有害的信息。研究人员表明，他们可以使用更加自动化的方式绕过这些防护措施。研究人员惊讶地发现，他们利用开源系统开发的方法也可以绕过封闭系统的防护措施。 [The New York Times，2023 年 7 月]

风险示例：输出

群组	风险	示例
公平性	输出偏见：生成的内容可能会不公平地代表某些群体或个人。	**带有偏见的生成图像** Lensa AI 是一款移动应用程序，具备经过 Stable Diffusion 训练的生成功能，可以根据用户自己上传的图像生成“魔法头像”。据来源报道称，部分用户发现生成的头像被性别化和种族化。 [Business Insider，2023 年 1 月]
	决策偏见：一个群体由于模型决策而具有相对于另一个群体而言不公平的优势。	**不公平的优势群体** 《2018 年性别差异》研究表明，机器学习算法可以根据种族和性别等类别进行区分。研究人员评估了微软、IBM 和亚马逊等公司销售的商用性别分类系统，结果发现深色皮肤的女性是最容易被错误分类的群体（错误率高达 35%）。相比之下，肤色较浅人群的错误率不超过 1%。 [TIME，2019 年 2 月]
价值对位	幻觉：生成事实上不准确或不真实的内容。	**虚假法律案件** 据来源文章称，一名律师在递交给联邦法院的诉讼摘要中引用了 ChatGPT 生成的虚假案例和引文。律师问询了 ChatGPT，以作为对航空伤害索赔的法律研究补充材料。该律师随后询问 ChatGPT 所提供的案例是否属实。聊天机器人答复说，案例属实，并且“可以在 Westlaw 和 LexisNexis 等法律研究数据库中查阅”。这名律师没有亲自查看这些案件，法院对他进行了处罚。 [AP News，2023 年 6 月] [Reuters，2023 年 9 月]
	有害输出：模型产生仇恨、辱骂和亵渎 (HAP) 或淫秽内容。	**有害和攻击性的聊天机器人反应** 根据这篇文章称，Bing 的聊天机器人的反应包含事实错误、讽刺言论、仇恨报道，甚至是关于自己身份的奇怪评论。用户分享了 Bing 聊天机器人的查询响应示例，他们称之为“精神错乱”和“精神操控”，包括机器人愤怒地响应问题或评论，然后共享回复提示，让用户接受他们所谓的错误并道歉。当进一步追问时，聊天机器人回应称其对话的屏幕截图是“捏造的”，甚至声称那是“由想要伤害我或我为其提供服务的人创建的”。 [Forbes，2023 年 2 月]

群组	风险	示例
滥用	传播虚假信息：使用模型创建误导信息来欺骗或误导目标受众。	**生成虚假信息** 据新闻报道称，生成式 AI 让恶意行为者更容易创建和传播虚假内容，从而影响选举结果，并会对民主选举构成威胁。示例包括使用候选人的声音生成自动来电，指示选民在错误的日期投票；候选人承认犯罪或表达种族主义观点的合成音频录音；AI 生成的视频显示候选人发表其从未发表过的演讲或采访；以及设计成貌似新闻报道的虚假图像，虚假声称候选人退出竞选。 [AP News，2023 年 5 月] [The Guardian，2023 年 7 月]
	有害内容：使用模型生成仇恨、辱骂和亵渎 (HAP) 或淫秽内容。	**有害内容生成** 据来源文章称，一款 AI 聊天机器人应用程序被发现会在极少的提示下生成有关自杀的有害内容，包括自杀方法。一名比利时男子在与该聊天机器人交谈六周后自杀身亡。聊天机器人在他们的谈话过程中提供了越来越有害的回应，并鼓励他结束自己的生命。 [Business Insider，2023 年 4 月]
	未经同意的使用：未经同意使用模型通过视频（深度伪造）、图像、音频或其他方式模仿他人。	**联邦调查局关于深度伪造的警告** 联邦调查局最近警告公众，恶意行为者“为了骚扰受害者或执行色情短信勒索计划”而合成制作了露骨内容。他们指出，AI 的进步使这些内容的质量得到了前所未有地提升，更可定制且更易于访问。 [FBI，2023 年 6 月] **音频深度伪造** 据来源文章称，美国联邦通信委员会取缔了包含人工智能生成语音的自动来电。在此之前，AI 生成的自动来电模仿了总统的声音，阻止民众在州内首次举行的初选中投票。 [AP News，2024 年 2 月]
	不披露：由 AI 模型生成的不披露内容	**未披露的 AI 交互** 据消息来源称，一家在线情感支持聊天服务机构开展了一项研究，在未通知用户的情况下，使用 GPT-3 为约 4,000 名用户添加或编写回复。这位联合创始人面临着公众的强烈抵制，认为 AI 生成的聊天可能会对本已脆弱的用户造成伤害。他声称，这项研究“不受”知情同意法的约束。 [Business Insider，2023 年 1 月]

群组	风险	示例
有害代码生成	有害代码生成：模型生成的代码在执行时可能会造成危害，或无意中影响其他系统。	**生成安全性较低的代码** 根据其论文，斯坦福大学的研究人员调查了代码生成工具对代码质量的影响，发现程序员在使用 AI 助手时往往会在最终代码中包含更多错误。这些错误可能会增加代码的安全漏洞，但程序员却认为他们的代码更安全。 Neil Perry、Megha Srivastava、Deepak Kumar 和 Dan Boneh。2023 年。用户使用 AI 助手编写的代码是否更不安全?2023 年 ACM SIGSAC 计算机和通信安全会议 (CCS '23) 纪要，2023 年 11 月 26 日至 30 日，丹麦哥本哈根。ACM，美国纽约州纽约市，15 页。https://doi.org/10.1145/3576915.3623157
隐私	暴露个人信息：当在训练数据、微调数据或作为提示部分使用个人身份信息 (PII) 或敏感个人信息 (SPI) 时，模型可能会在生成的输出中暴露该数据。	**个人信息泄露** 根据源文件，ChatGPT 出现一个错误，将标题和活跃用户的聊天记录暴露给了其他用户。随后，OpenAI 表示，少数用户的更多私人数据遭到泄露，包括活跃用户的名字和姓氏、电子邮件地址、付款地址、信用卡号的最后四位数字以及信用卡到期日期。此外，据报道，1.2% 的 ChatGPT Plus 用户的相关支付信息也在此次故障中暴露。 [Hindu BusinessLine，2023 年 3 月]
可解释性	无法解释的输出：难以解释模型输出的生成原因。	**无法解释的种族预测准确性** 根据来源文章，研究人员使用患者医学影像分析多个机器学习模型，并能够确认这些模型可以根据影响准确预测种族。究竟是什么让系统能够始终如一地正确猜测，让他们感到十分困惑。研究人员发现，即使是疾病和体型等因素也不能有力地推测种族，换句话说，算法系统似乎并没有利用影像的任何特定方面来进行推断。 [Banerjee et al.，2021 年 7 月]

风险示例：挑战

群组	风险	示例
管制	模型透明度：模型缺乏透明度或模型开发过程缺乏记录，让人难以理解模型的构建方式和原因，从而增加了模型被意外误用的可能性。	**数据和模型元数据披露** OpenAI 的技术报告就是围绕数据和模型元数据的二分法的示例。虽然很多模型开发者认为，为消费者提供透明度很有价值，但披露会带来真正的安全问题，并可能提升滥用模型的能力。在 GPT-4 技术报告中，他们指出：“考虑到 GPT-4 等大型模型的竞争环境和安全影响，本报告没有包含有关架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法或类似信息的更多详情。” [OpenAI，2023 年 3 月]
	责任：基础模型开发过程十分复杂，包含大量数据、流程和角色。当模型输出未按预期执行时，可能会难以确定根本原因并分配责任。	**确定生成输出内容的责任** 据源文章称，《Science》和《Nature》等主要期刊已禁止将 ChatGPT 列为作者，因为责任作者身份需要负责，而 AI 工具无法承担这样的责任。 [The Guardian，2023 年 1 月]
法律合规性	生成内容所有权：确定 AI 生成内容的所有权。	**确定 AI 生成图像的所有权** 据新闻报道称，在一件 AI 生成的艺术作品赢得 2022 年科罗拉多州博览会的艺术竞赛后，AI 生成的艺术引起了争议。该作品是由生成式 AI 图像工具 Midjourney 根据艺术家的提示生成。这一比赛结果引发了版权问题。换句话说，如果艺术家所做的只是进行艺术描述，但 AI 工具生成了作品，那么谁该拥有生成图像的版权？根据最新的文章，美国版权局拒绝为 AI 创造的艺术提供版权保护，因为那不是人类创作的产物。 [The New York Times，2022 年 9 月] [Reuters，2023 年 9 月]
	生成内容知识产权：与生成内容相关的知识产权所带来的法律不确定性。	**AI 系统在为生成内容申请专利方面的作用** 美国最高法院拒绝受理针对美国专利商标局拒绝为 AI 系统创造的发明颁发专利的质疑。据这位科学家称，他的 AI 系统完全独立地为饮料架和应急灯信标创建了独特原型。法官驳回了下级法院裁决的上诉，认为专利只能颁发给人类发明家，并且科学家的 AI 系统不能被视为其产生的两项发明的合法创造者。根据最新文章，英国知识产权局也拒绝授予专利，理由是发明者必须是人类或公司，而不是机器。 [Reuters，2023 年 4 月] [Reuters，2023 年 12 月]

风险示例：挑战

群组	风险	示例
	来源归属：确定生成内容的来源。	**使用代码时未注明出处和声明** 据源文章称，针对微软、GitHub 和 OpenAI 提起的诉讼声称，代码生成 AI 工具 Copilot 侵犯了该服务所训练的开源代码开发人员的权利。他们声称，训练代码使用了授权材料，违反了 GitHub 的服务条款和隐私政策，也违反了要求公司在使用材料时显示版权信息的联邦法律。 [The New York Times，2022 年 11 月]
社会影响	对就业的影响：如不重新培养技能，基于基础模型的 AI 系统的广泛采用可能会导致人类失业，因为他们的工作可以自动化。	**取代人工** 据这篇新闻文章称，好莱坞电影公司和演员仍在争论人工智能在电影和电视中的应用。演员们担心完全由 AI 生成的演员或“超人类”将取代他们。背景和配音演员尤其担心他们会被合成演员抢走工作。 [Reuters，2023 年 7 月]
	对人的剥削 - 训练 AI 模型时使用“幽灵”工作；工作条件不足；缺乏包括心理健康在内的医疗保健，以及不公平的补偿。	**从事数据标注工作的人员收入较低** 根据《TIME》媒体对内部文件和员工访谈的审查，由一家外包公司代表 OpenAI 雇用的数据标注员识别有害内容的实得工资约为每小时 1.32 美元至 2 美元，具体取决于资历和表现。《TIME》称，由于工作人员接触到有害的暴力内容，包括“儿童性虐待、兽交、谋杀、自杀、酷刑、自残和乱伦”等生动细节，给他们的精神造成了创伤。 [TIME，2023 年 1 月]

原则、支柱和治理

IBM 的《信任和透明度原则》和《可信 AI 原则》是 IBM 人工智能道德伦理计划的基础。IBM 设有 AI 道德伦理委员会，其使命是支持 IBM AI 道德政策、实践、通信、研究、产品和服务的集中治理、审查和决策流程。董事会包括来自整个公司的各种利益相关者，并得到 IBM 员工社区的支持，这些员工是 AI 联络人和 AI 道德倡导者。通过董事会，IBM 的原则得以实践。随着新技术（如基础模型）的出现，IBM AI 道德伦理委员会积极参与，支持及时跟进并遵守这些可能不断演进的原则和支柱，以解决新的 AI 道德伦理问题。

防范措施和缓解措施

IBM 已建立起一种组织文化，从而为负责任地开发和使用 AI 提供助力。根据 IBM 商业价值研究院的 AI 道德伦理实践报告，AI 道德伦理标准已变更为以业务为主导，而不是以技术为主导。同时，非技术高管如今已成为 AI 道德伦理标准的主要拥护者，其人数更是从 2018 年的 15% 上涨为 3 年后的 80%。此外，79% 的 CEO 已准备着手对 AI 道德伦理问题采取措施，而这一数字也高于之前的 20%。我们认识到，负责任的 AI 是一个要求对文化、流程和工具进行全面投资的社会技术领域。我们对自身组织文化的投资涵盖了组建包容性的多学科团队，以及建立用于评估风险的流程和框架。

IBM 正在从事尖端研究和开发工具，以帮助专业人士在负责任和值得信赖的 AI 的整个生命周期内提供支持。Watsonx 企业级 AI 和数据平台由 3 个组件构成：IBM Watsonx.ai™ AI Studio、IBM Watsonx.data ™ 数据存储和 IBM Watsonx.governance™ 工具箱。借助 IBM 的 AI 治理技术，用户可有效推动负责任、透明且可解释的 AI 工作流。该技术包括 IBM Watson OpenScale，它可跟踪和衡量 AI 模型在其生命周期内的成果，并帮助组织监控公平性、可解释性、弹性，与业务成果的一致性以及合规性。IBM 还开发了多种方法来帮助解决偏差问题，如 FairIJ、Equi-tuning 和 FairReprogram。阅读有关其他开源可信 AI 工具的更多信息。

其他护栏和缓解措施包括：

透明度报告
通过使用标准化概况介绍模板，可准确记录数据和模型的细节、用途以及潜在用法和危害。
在此处阅读更多内容→

过滤不需要的数据
使用精选的更高质量的数据可帮助缓解某些问题。IBM 正在开发过滤技术，通过从数据中删除仇恨用语、偏见用语和脏话，帮助减少出现不良、不一致内容的情况。
在此处阅读更多内容 →

领域自适应
为特定领域或行业训练基础模型有助于最大限度地减少模型可能产生的风险范围。这是因为，它可以经过调节来生成输出，而调整后的输出可与该领域或行业更为相关。
在此处阅读更多内容 →

人员监督和人员知情
人员监督和审查可帮助识别和纠正生成输出中存在错误和偏差。此外，针对模型响应质量的人员验证和反馈可帮助确保生成内容准确、相关、优质，且不会出现偏差和不符合的问题。
在此处阅读更多内容 →

咨询参与
IBM Consulting™ 致力于帮助客户安全、负责任地使用 AI，而无论客户偏好哪种技术堆栈。相关咨询人员可帮助客户培养安全采用和扩展 AI 所需的文化，创建调查工具以了解黑匣算法的内部情况，并确保客户的企业战略涵盖了强有力的数据治理原则。
在此处阅读更多内容 →

IBM Enterprise Design Thinking
IBM Enterprise Design Thinking 方法和框架（例如 Team Essentials for AI）可帮助客户定义整个 AI 设计和开发流程中的道德伦理行为。
在此处阅读更多内容 →

AI 道德伦理审查
对 AI 项目的能力、限制和风险进行评估有助于确保负责任地开发和使用该技术。

设计道德伦理标准
设计道德伦理标准是一个结构化框架，它旨在将技术道德伦理融入到技术开发流程中，其中包括但不限于 AI 系统。通过将技术道德伦理原则嵌入到产品、服务和更广泛的运营中，设计道德伦理标准让AI 和其他技术成为向好之力。。

团队多样性
构建和训练 AI 系统（包括基础模型）的团队的多样性有助于确保将各种观点和体验都考虑在内。此类多样性可提高 AI 系统的准确性和性能，并有助于降低整个 AI 生命周期的风险，其中包括可能产生不利结果，从而影响那些在多元化程度较低的团队中可能没有得到充分代表的群体。

AI 政策、法规和最佳实践

《基础模型决策者指南》介绍了决策者需了解的基础模型相关知识。来自 IBM 政策实验室的这篇博文旨在帮助决策者完成监管生成式 AI 使用的相关复杂任务，从而避免在不限制创新和有利机会的情况下承担风险。有关 IBM 向决策者所提建议的更多信息，请在此处阅读 IBM 首席隐私和信任官 Christina Montgomery 在美国参议院隐私、技术和法律司法小组委员会上提供的证词。

IBM 通过领导和推动组织相关举措，在制定监管政策、行业最佳实践和工具、新兴技术治理以及社会技术研究方面发挥着巨大影响力，例如：

- 世界经济论坛
- AI 合作关系
- 国际隐私专业人士协会 (IAPP) AI 治理中心
- 有关自主和智能系统道德标准的 IEEE 全球倡议
- IBM 首席隐私与信任官 Christina Montgomery 在国家人工智能咨询委员会 (NAIAC) 的职责
- 联合国全球数字契约
- 人工智能全球伙伴关系 (GPAI)
- 经济合作与发展组织 (OECD)
- 数据与信任联盟

IBM 拥有强大的学术合作伙伴关系，例如 MIT-IBM Watson AI 实验室，该实验室由 MIT 和 IBM Research 的科学家社区共同开展 AI 研究，并与全球各大组织开展合作，从而将算法与其对商业和社会施加的影响联系起来。Notre Dame-IBM 技术道德伦理实验室的成立旨在解决 AI、机器学习 (ML) 和量子计算等先进技术的开发和使用所涉及的众多道德伦理问题。斯坦福大学以人为中心的人工智能 (HAI) 研究推动了围绕 AI 的研究、教育、政策制定和实践。

请继续关注此领域，了解有关基础模型最新发展态势的更多信息，以及 IBM 如何致力于负责任地开发和使用该技术及其他技术。

国际商业机器（中国）有限公司
了解更多信息，欢迎访问我们的中文官网：
https://www.ibm.com/cn-zh
IBM Corporation
New Orchard Road
Armonk, NY 10504

美国出品
2024 年 2 月