中国内地和香港地区

普华永道

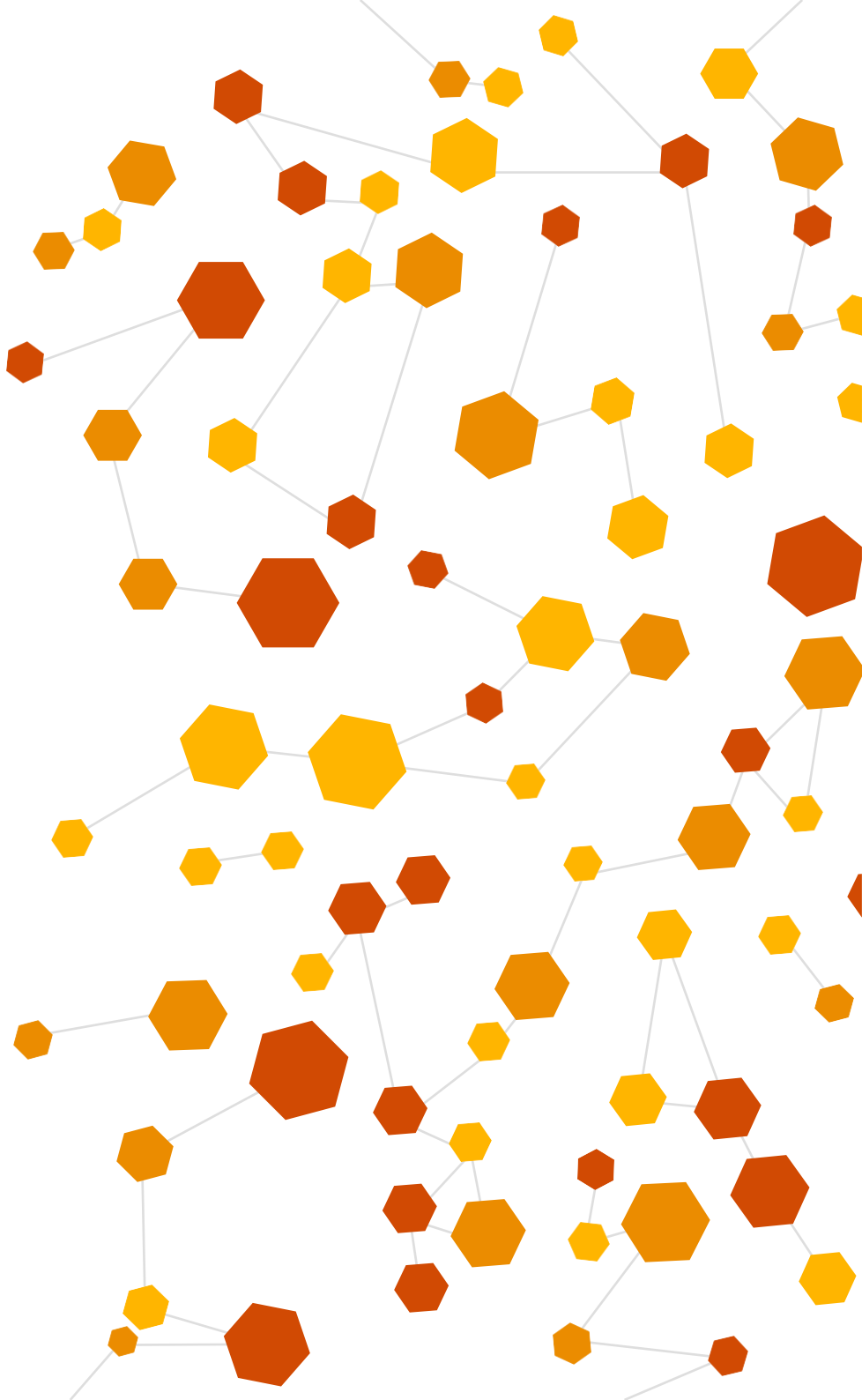
管理层指南 — 治理为先

生成式人工智能的风险和机遇管理

2023年10月

搭建可信赖的人工智能

坚持负责任的人工智能



目录

引言

2022年11月，我们迎来了一场真正的革命。一夜之间，即便是仅会使用聊天软件的网民都感受到了人工智能的魔力。

在短短一周内，便有上百万用户通过ChatGPT撰写短文、编写计算机代码、创作艺术、把长文精炼为更有文采、更简洁的短文。

与此同时，一些心怀不轨的群体则试图利用生成式人工智能编写恶意软件、散播更具迷惑性的钓鱼邮件和伪造更逼真的虚假身份。这似乎预示着大规模欺诈、隐私泄露、虚假信息和网络攻击正向我们袭来。

在ChatGPT亮相仅数月后，生成式人工智能便深入地融入我们的生活和业务。活跃消费用户增长之快前所未有。从OpenAI的GPT3到GPT4，人工智能的功能实现了飞跃，在代码编写和中级专业写作方面的成绩显著。各大科技公司也不甘落后，纷纷推出/更新竞争产品；创业公司发布了定制应用程序的模型；包括普华永道在内的各大公司已宣布大规模投资自家的“CompanyGPT”，以供内部使用和对外提供新服务。

但担忧也随之而来，有人告诫“具有人类竞争智能的人工智能系统可以对社会和人类构成深远的风险”，社会大众和行业专家对此都深感忧虑。该项技术的头部供应商也承认存在类似风险。

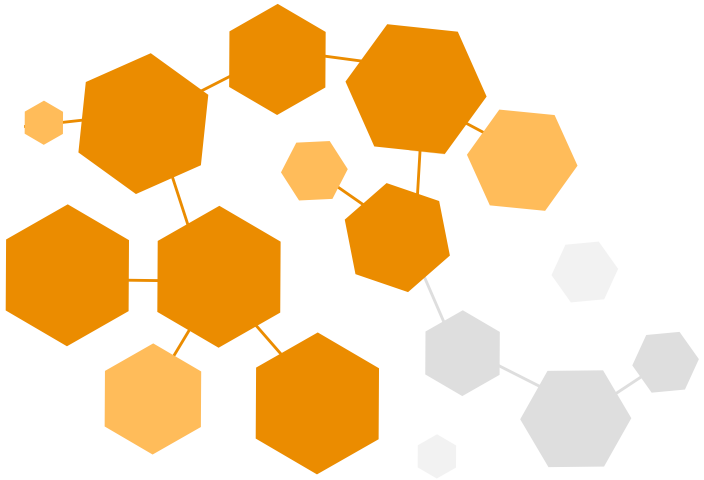
风险管理关乎成败。如果公司希望成功实施生成式人工智能计划并取得竞争优势，则需要评估该技术可能给全公司带来的风险。为此，公司需要一套风险管理框架，以便更好地抓住机遇。

针对生成式人工智能，采用以风险为导向的方法将确保公司在数字化道路上，始终符合监管机构、消费者和其他利益相关者的期望。公司只有在采用生成式人工智能的同时建立信任，才能快速、充分地利用这项改变世界的技术带来的红利。

公司是否会冒险以信任换速度？

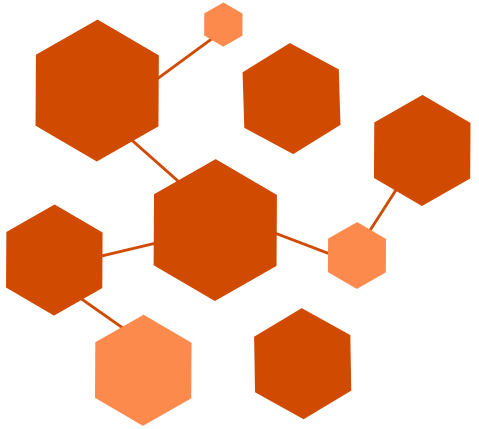
46%中国内地及68%中国香港

企业正在投资包括人工智能在内的新兴技术。
（资料来源：普华永道，2023年）



78%

中国受访者认为，产品和服务采用人工智能的利大于弊
（资料来源：益普索，2022年）



只有 48%

员工认为，加速采用人工智能有助于事业发展
（资料来源：益普索，2022年）

企业面临的风险

生成式人工智能是人工智能家族中的一个强大分支，对企业具有颠覆性影响。它能够支持企业运营的几乎所有方面（包括客户服务、软件开发和数据分析）实现自动化和提升。

借助人工智能，公司能针对不同客户定制个性化互动，改善互动方式，从而提升客户关系。人工智能可以自动完成批量任务，例如处理保险索赔和沟通或执行某些软件开发任务。

通过人工智能，团队能更轻松地了解非结构化数据，包括合同、发票、客户反馈、保单、保险理赔员告知单、绩效评估、医疗记录等。

员工生产力可以大幅提升。根据OpenAI的估计，每10名员工中约有8名*可通过生成式人工智能自动完成其至少10%的工作任务，而这仅仅是起步。生成式人工智能工具可以自动完成例行任务，使员工得以解放，从而投身更具创造性的工作，或以更创新、更全面的方式理解复杂主题和任务，提升批判性思维。

随着技术需求的持续增长，生成式人工智能的功能也在不断进步。仅仅不到四个月，人工智能语言系统就在复杂度和功能性两方面取得了显着进步，发展劲头势不可挡。

公司想要搭上人工智能的快车，实现持续发展，则需尽早招募风险专业人员。如此方能树立公司信心，顺利实施和推进生成式人工智能项目。

公司的风险管理人员必须管理好新增和加剧的风险，以及在业务、法律和监管方面的一系列挑战。白宫、美国国会、美国联邦贸易委员会、中国国家互联网信息办公室和欧盟相继采取行动，对生成式人工智能进行监管。与此同时，针对生成式人工智能违反数据保护法，在未经同意的情况下，收集、使用和披露个人信息的行为，部分国家（意大利、加拿大、西班牙、法国、德国）展开了调查，以此响应社会各界的不满或担忧。

随着生成式人工智能技术日益普及，中国国家互联网信息办公室（网信办）、国家发展和改革委员会（发改委）、教育部、科学技术部（科技部）、工业和信息化部（工信部）、公安部和国家广播电视总局（广电总局）于2023年7月10日联合发布了《生成式人工智能服务管理暂行办法》，并于2023年8月15日生效。2023年10月18日，中央网信办更进一步发布《全球人工智能治理倡议》。

习主席在中共中央政治局会议上指出，“要重视通用人工智能发展，营造创新生态，重视防范风险”。

*资料来源于美国某项研究

生成式人工智能对企业的影响

企业在涉足生成式人工智能时需重点关注以下方面。

职能转型：重构运营。

为快速获得投资回报，部署生成式人工智能的“最佳着力点”很明确，即于运营各方面（例如营销、财务、供应链和税务合规）全面部署，以实现自动化和提升。凭借生成式人工智能，公司可最大限度地利用现有资源、强化决策能力、提升客户和员工体验。生成式人工智能可以优化数据集和文档集的整理，简化人工研究工作，还可用于起草财务、风险和合规性报告、定制个性化客户服务方案、识别人工报告中违规点。然而，为确保结果可靠，企业应依托负责任的人工智能框架施加强有力的监管。

负责任的人工智能：建立信任和管理风险。

企业想要通过生成式人工智能实现彻底革新，信任是首要前提。这需要企业负责任地使用人工智能，制定精细的实施方案，确保使用过程中的诚信和道德标准。通过整合技术、流程和技能，负责任的人工智能框架可以解决生成式人工智能的相关风险，例如网络威胁、隐私问题、法律影响、性能问题、偏见和知识产权风险。为了有效实施负责任的人工智能，最好的方法是将信任融入设计，从起步阶段便将信任纳入公司体系，并根据经验教训不断完善。

员工：培养技能，适应新工作方式。

生成式人工智能可以为知识型员工赋能，助力员工在更短的时间内取得更大的成果。然而，为了充分利用人工智能的潜力，公司应当为员工提供必要的技能培训，适应新工具和新工作方式。企业必须深谙运用之道，切勿因生成式人工智能看似简单而掉以轻心。随着人工智能的采用，新的岗位也将应运而生，如提示工程师和模型专家。这些人才的组合好比“人工智能工厂”，为企业人工智能系统的实施和维护提供支持。

云和数据：奠定增长基础。

生成式人工智能可挖掘非结构化数据的潜力，以此支持决策优化、收入增长和业务扩展。在推进数据和应用程序现代化时，公司应当充分考虑生成式人工智能，进而从根本上改变云应用的构建和运作方式。

新商业模式：数据货币化和行业重塑。

如果企业能自主地将非结构化数据转换为可实现的构想或新软件代码、产品、服务，或为每位客户提供真正意义上的定制体验，未来将会如何？生成式人工智能提供了这种可能，它既能建立新商业模式，也能颠覆基本价值链。企业在利用生成式人工智能提升运营效率的同时，不妨思考这项技术在近期将如何实现公司业务和所在行业的颠覆。

识别生成式人工智能的应用场景

生成式人工智能可在企业运营的大多数方面（从客户服务到软件开发和数据分析）实现自动化和提升。企业可以从以下角度识别应用场景。

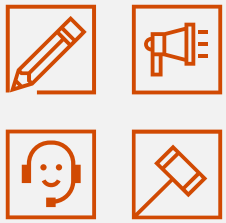
自上而下

生成式人工智能在各行业和业务职能方面的用途

行业

生成式人工智能技术能够解决不同行业（例如医疗保健、科学研究和金融）的众多问题。

哪些生成式人工智能技术可用于企业所在的行业？生成功能如何助力企业立于不败之地？



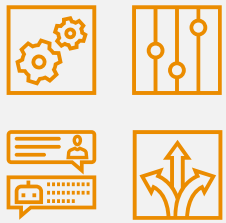
业务职能

生成式人工智能技术可用于研发、营销、销售、客户支持、运营、法律和后台流程。

生成式应用程序如何提供尽善尽美的服务？

自下而上

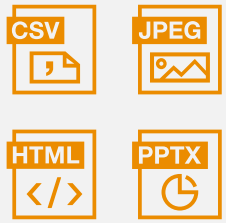
生成式人工智能优异的工作表现对员工的益处



任务类型

某些类别的任务（如汇总、转换和问答）最适合生成模型的技术机制。

这些功能如何融入企业的业务流程？



输出数据形式

生成技术可生成全新的文本、代码、图像、音频、视频等。

团队和员工个人每天产出何种类型的内容？

企业的IT和风险专业人员可以帮助企业加速实施负责任的生成式人工智能。他们可以确保生成式人工智能的隐私保护到位、公平（不良偏见得到有效管理）、有效可靠、负责透明、可解释和可说明。

换言之，即是赢得信任。

管理新增和加剧的风险

我们认为，公司需要了解和管理该技术的以下四大固有风险：

数据风险

错误信息传播、知识产权或合同问题（因未经批准使用数据），或因采用低质量数据训练生成式人工智能模型而产生的误导性和有害内容。

模型和偏见风险

在语言模型开发时，违反道德和负责任的人工智能原则，导致输出歧视性或不公平的内容。

提示或输入风险

向人工智能模型提供粗浅提示或问题而生成的误导性、不准确或不良回答。

用户风险

用户在不知情的情况下采纳错误信息和其他不良内容而导致的意外后果。例如，用户可能会将人工智能产生的虚假信息（即错误或荒谬的回答）当作事实。

因生成式人工智能的使用情况不同，公司还可能面临其他风险，尤其是在公司计划创建与基础模型关联的专有模型并添加专有或第三方数据的情况下。

公司的风险专业人员有助于获取各利益相关方对生成式人工智能的信任。因此，他们应将信任融入设计，以信任为价值主张，而非一味追求速度，以此赢得客户、投资者、业务合作伙伴、员工和社会的认可。

风险领域专家应考虑隐私、网络安全、合规、第三方管理、法律义务、知识产权方面的全部风险，并相互协作管理好整体企业风险。

同时，公司应与人才/人力资源负责人合作，制定各级培训计划，让每位员工熟知生成式人工智能的风险与回报。安排经验丰富的人员检查生成式人工智能输出的“初稿”。

监控员工表现，防止随时间推移出现“技能萎缩”、自满、质量下降的情况。

各种成熟的框架为设计和部署值得信任的人工智能应用程序提供了良好开端，其中包括香港政府资讯科技总监办公室（OGCIO）发布的《人工智能道德框架》、香港个人资料私隐专员公署（PCPD）发布的《开发及使用人工智能道德标准指引》、中国银行保险监督管理委员会发布的《关于银行业保险业数字化转型的指导意见》以及《IS0/IEC 23053:2022 使用机器学习（ML）的人工智能（AI）系统框架》。

对于公司而言，制定有效的人工智能治理策略至关重要。除风险管理专业人员外，公司内外部人员均可能影响公司负责任地使用生成式人工智能的能力，包括数据科学家、数据工程师、数据提供方、领域专家、社会与文化分析师、多样性、公平性、包容性与可接取性、社区影响领域的专家、用户体验设计师、治理专家、系统出资人、产品经理、第三方实体、评估人员以及法律和隐私保护专业人员。

打造值得信任的人工智能的关键治理要素

1.模型设计和开发

人工智能模型规划、设计、采购和/或开发流程的关键控制。

- 法务与监管合规性
- 影响分析和风险评估
- 战略调整和业务关怀
- 模型设计、方法和假设
- 范围和要求明确
- 限制、防范措施和超驰控制
- 方案分析
- 培训数据

2.模型验证

采用以风险为导向的测试方法，包括设计评估指标和验收标准。

- 数据安全
- 有效性和可靠性
- 安全性和稳健性
- 透明度和问责制
- 可说明性和可解释性
- 隐私保护
- 公平（有害偏见管理）
- 性能和商业价值

3.模型管理

在生产环境中人工智能模型部署、操作和监控的关键控制。

- 性能监控（即漂移）
- 效益管理
- 变更和发布管理
- 问题管理
- 配置和版本管理
- 人工监督
- 培训和技能提升
- 反馈机制

关键输入项和依赖关系

- 岗位和职责清晰
- 多元包容的团队
- 领导层参与
- 员工队伍和技能规划
- 政策和程序明确
- 风险管理框架
- 风险容忍度明确
- 第三方风险管理

生成式人工智能所带来的关键风险以及风险管理人员可采取的措施，请参阅后续各章节。

董事会和董事成员可采取的措施

首先，董事会要提高董事对人工智能和生成式人工智能的认识，利用管理层和外部资源，紧跟技术发展潮流，与时俱进，了解新的用例、商业模式的变化、相关风险并坚持负责任的实施原则。

董事需要从业务角度考虑人工智能和生成式人工智能技术及其使用。董事会行使监督职能，负责向管理层提出问题（本报告中分享了一些可供借鉴的示例），并适时向管理层提出质疑。董事会应考虑采用人工智能和生成式人工智能是否需要额外的技能，或是否依赖管理层或第三方。

复核人工智能的应用成本和效益

- 董事会应当与管理层讨论应用人工智能的总体效益和成本。
- 从效益角度出发，企业需要重新构想完成工作的方式，包括员工的工作方式、与客户的互动方式、销售的产品和竞争方式。
- 管理层希望营造一种鼓励员工学习新技能、激励员工快速创新以抓住新机遇的企业文化，因此，公司可能需要投入成本，为全体员工提供人工智能和生成式人工智能应用方面的知识和技能培训。

考虑与利益相关方的沟通

- 董事会还应考虑公司如何向内外部利益相关方讲好人工智能故事；
- 管理层注重战略变革，以期在当今瞬息万变的商业环境中保持适应能力和竞争力
- 确保采取防范措施以保护员工、客户、

制定问责治理模型

- 公司需要建立问责治理模型，首先要明确公司内部人工智能治理的责任。
- 有效的治理模型使公司能够评估与特定技术和单个用例相关的独特效益和风险权衡。

监督人工智能监管计划以衡量项目成果

- 管理层需要确定人工智能的发展方向及优先使用场景。当公司确定了这一发展道路，可能需要进行大规模的数字化转型，同时需要大量投资。
- 为实现公司转型而进行重大投资时，董事会应了解人工智能的数字化转型战略和计划，以及如何与业务战略保持一致。

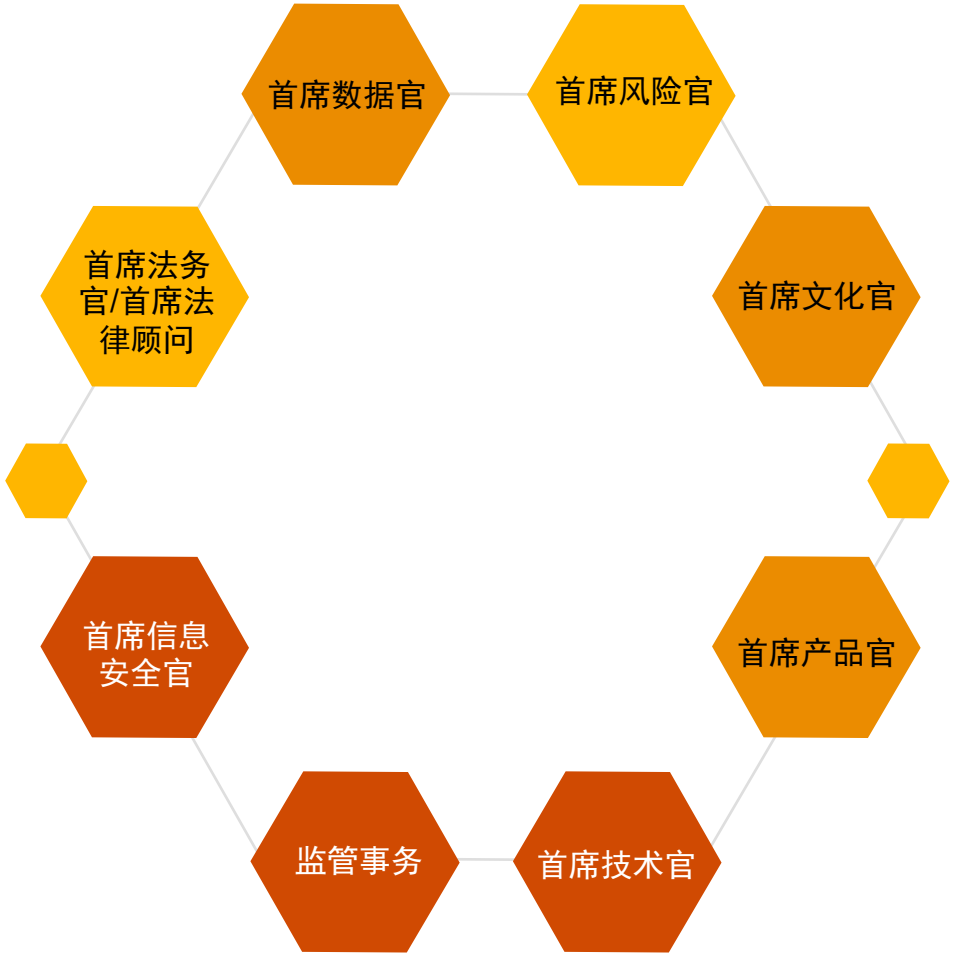
高级管理层可采取的措施

我们提供了如下两个示例，其中高管及其团队可协作管理风险，同时企业也可以通过有效的治理强化风险管理。

示例1：风险开发生成式人工智能医疗咨询聊天机器人所带来的机遇与风险

某医疗服务提供商考虑使用生成式人工智能提供医疗建议，以此代替临床工作人员为患者远程提供健康咨询服务。提供商收集多年的患者数据、症状、诊断和治疗方法，用于训练模型。

- **首席数据官：**确保数据准确无误，数据已清洗，不偏倚某些人群、年龄组等。
- **首席合规官：**确定数据的使用满足当地健康记录法规和《隐私法》的合规性义务，例如：《健康保险法》、《个人健康档案法》、《老年护理法》和当地非健康个人隐私保护法规。
- **首席产品官：**配合采取隐私保护融入设计的方法，并让用户了解其输入信息将如何被使用，哪些数据将被保留。
- **首席技术官：**为该使用场景设计一个专用实例，避免无意中将数据与其他运行中的生成式人工智能工具混用。
- **首席法务官/首席法律顾问：**负责合规性，特别是遵守健康和数据相关法律，对知识产权和数据、与过失建议相关的风险进行潜在法律分析，与生成式人工智能平台协商合同条文，确保将患者数据与人工智能模型的公共实例分离。



- **首席信息安全官：**将该应用程序和数据存储区确定为“一级核心资产”，并按照数据的最高敏感级别提供充分的保护。

- **内部审计：**围绕拟实施的系统和模型制定审计风险评估计划，考虑各地区健康记录法规和《隐私法》的法律和合规风险。例如：《健康保险法》、《个人健康档案法》、《老年护理法》和当地的非健康个人隐私保护法，并对系统和模型的可靠性和性能进行评估。

- **首席风险官：**与首席合规官协商制定政策、培训、测试和控制措施，以确认人工智能生成的医疗建议准确且符合国家医疗委员会的标准。

示例2：高效验证信用分析并树立风险意识

一家银行考虑使用生成式人工智能自动化流程对交易对手信用评估中的所有交易对手进行年度信用检查，并根据市场事件和其他触发因素对高风险客户进行季度检查。

- **首席数据官：**确保数据准确无误，数据已清洗，并且不存在对某些群体的固有偏见。设置专用沙盒测试用例以支持产品。

- **首席合规官：**更新流程图和合规材料，展示如何使用该技术达成决策，并证明其符合《隐私法》、《隐私（信用报告）法则》、《竞争与消费者法案》和《反歧视法》等相关法律。

- **监管事务：**更新报告规程。

- **首席产品官：**呼吁采用隐私保护融入设计的方法，让最终用户了解他们提供的数据将如何被使用，哪些数据将被保留。

- **首席法务官/首席法律顾问：**与信贷机构和其他数据供应商协商合同条款，允许将其数据用于生成式人工智能，并与生成式人工智能平台协商，保证客户数据不会与其他实例混用或用于训练其他实例。

- **首席财务官/财务总监：**确保内部控制环境以及相关风险和控制框架足以解决人工智能应用带来的潜在影响，提高财务报告流程完整性控制的设计和运行有效性，并符合国际财务报告准则或萨班斯法案404等法律法规要求。

- **首席信息安全官：**将该应用程序和数据存储区确定为“一级核心资产”，并按照数据的最高敏感级别为其提供保护。

底线

生成式人工智能的合理应用能够为公司节省时间和成本，优化产品和服务质量，甚至提高声誉。但该方法应以人为主导、科技为辅助，切勿颠倒了主次。

公司想要从这一突破性技术中获得切实的最大收益，必须从其整体利益出发，管理好应用人工智能技术带来的诸多风险。同时也需要利益相关方集思广益，全面考量引入生成式人工智能解决方案带来的影响和问题。实现风险与创新回报的平衡将有助于树立公司信誉并赢得竞争优势。

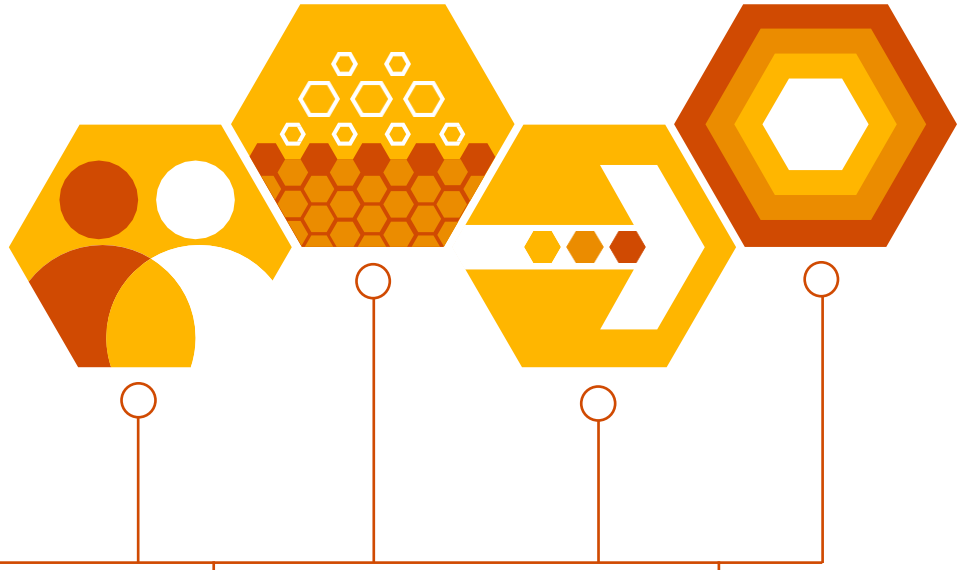
最终，生成式人工智能的进步取决于公司员工。提升员工技能，使员工了解使用生成式人工智能作为辅助或指导工具的局限性，以便员工充分利用其潜能。在制定企业风险防范措施的基础上，赋能员工运用自身经验，对生成式人工智能模型的输出结果进行批判性评估。每一位有判断力的用户都是信任的守护者。

如果企业对生成式人工智能风险有深刻理解，懂得如何设计、衡量和管理可信赖的人工智能系统，将更快地推进人工智能转型，更容易发现高价值的使用场景。

普华永道风险及控制服务部门的服务

我们的风险及控制服务部汇集了数据科学、工程、数据道德、数字法律、风险和治理领域的专家，可与公司所在行业的专家携手，协助公司以负责任的方式加快生成式人工智能的实施步伐，并能够助力公司在建立人工智能信任的发展道路上迈出关键一步。

“理解、建立、加速和保证”是我们助力企业部署人工智能的方法论，该方法论以我们25年的技术和数据治理领导者经验、屡获殊荣的人工智能咨询服务和全球技术联盟生态系统为基础。



理解

理解公司应对人工智能机遇和风险的基准线和准备情况。

- 创意构思和策略制定
- 使用场景探索与选择

建立

建立安全、快速地采用人工智能所需的基础。

- 数据成熟度审查
- 风险及控制审查
- 人工智能治理模型
- 负责任的人工智能

加速

加速创意构想、探索和原型设计。推广适用的构思并监测其有效性和安全性。

- 数据平台实施
- 生成式人工智能概念验证部署

保证

审查公司现有的人工智能框架、平台、模型和实施措施，帮助公司建立并维护在利益相关方群体中的信任。

- 控制实施
- 数据可观察性托管服务

联系我们了解更多信息

贺琪伟
亚太地区风险及控制服务主管合伙人
中国内地及香港地区风险及控制服务主管合伙人
普华永道香港
jennifer.cw.ho@hk.pwc.com

吕麟晖
中国内地及香港地区风险及控制服务 - 数字化鉴证与分析服务主管合伙人
普华永道中国
george.l.lu@cn.pwc.com

徐世达
中国内地及香港地区风险及控制服务市场主管合伙人
中国中部风险及控制服务主管合伙人
普华永道中国
jasper.xu@cn.pwc.com

毛英伟
风险及控制服务合伙人
普华永道香港
chris.yw.mo@hk.pwc.com

潘欣鹏
风险及控制服务合伙人
普华永道中国
rachel.pan@cn.pwc.com

张俊贤
网络安全及科技风险服务合伙人
风险及控制服务
普华永道中国
chun.yin.cheung@cn.pwc.com