WORLD GOVERNMENTS SUMMIT 2025

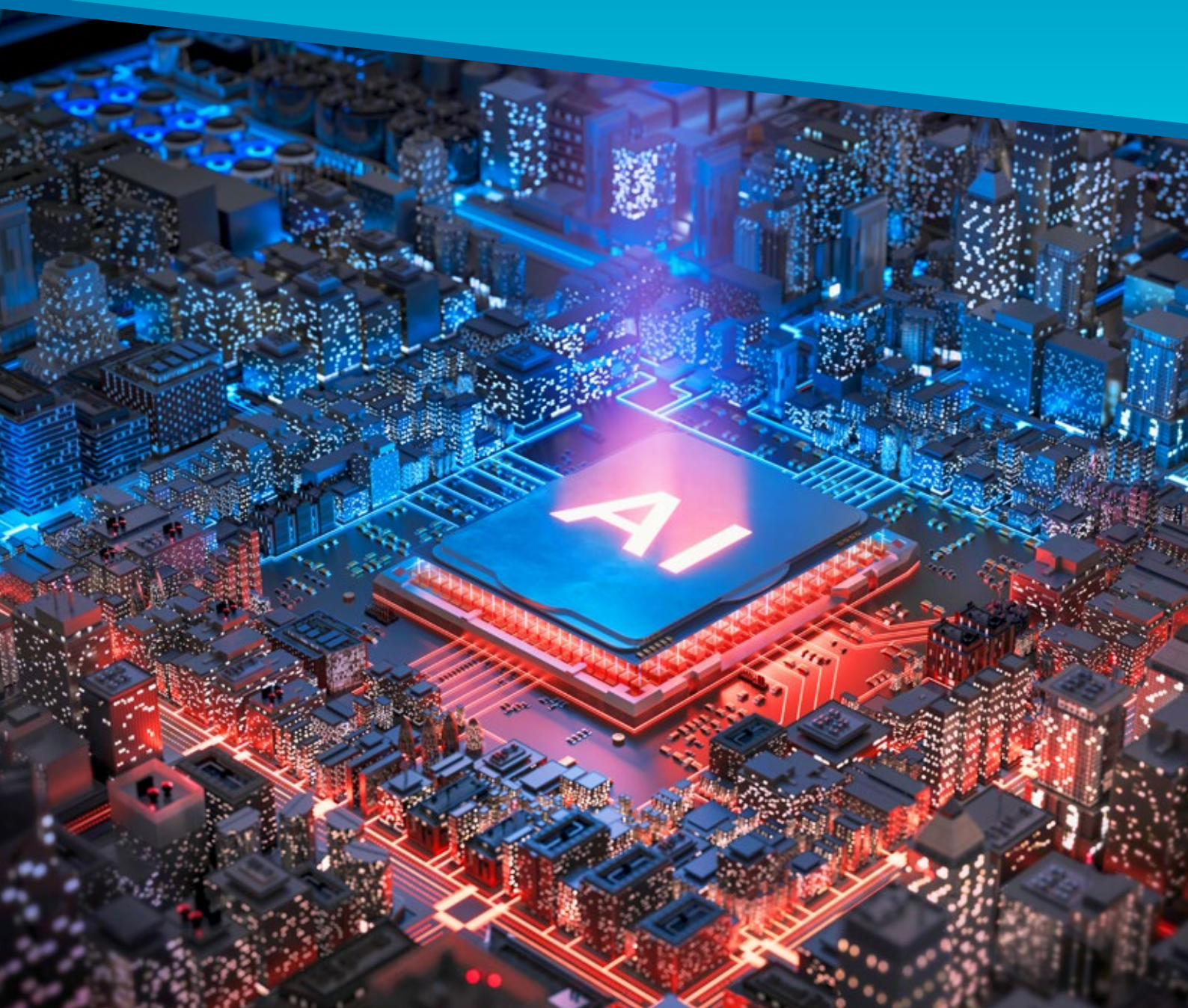
合作

KPMG

报告

人工智能治理的未来

阿联酋首相署与全球顾问



目录

主题

前言

我们认识到，一套清晰、可执行的AI原则构成了符合道德和负责任的AI发展的基石。这些原则不仅对于建立公众信任和确保组织问责至关重要，而且也有助于促进惠及公民、企业和政府的包容性创新。

随着全球监管框架的演变，例如2024年通过的欧盟人工智能法案，基于欧洲委员会的可信赖人工智能伦理指南，基于原则的治理已成为人工智能监管的基础方法。阿联酋通过其人工智能2031战略以及发布于2024年7月旨在发展和使用人工智能的阿联酋人工智能宪章，展现了区域和全球领导力，该宪章阐述了十二项关键原则，以确保人工智能安全、公平和透明地部署。

这份白皮书详细解读了阿联酋人工智能宪章的12项原则，针对人工智能全生命周期的实施提供了可操作的建议，并对照KPMG可信人工智能框架进行映射，为人工智能治理、风险管理以及监管合规提供了实用见解，同时为建设与阿联酋国家优先事项相一致、具有韧性和以人为本的人工智能系统提供了蓝图。

阿联酋宪章特别强调人类监督、包容性、安全性和法律合规性——这些价值观与经合组织、联合国教科文组织以及欧盟等全球人工智能伦理标准相契合。随着人工智能监管日益严格，主动遵循这些原则的组织将更有能力负责任地引领发展，降低风险，并充分释放人工智能创新的全部潜力。

要超越美好的愿望，组织必须将这些原则嵌入到运营现实中。这意味着要演变现有的治理模式，以支持人工智能的独特需求——例如数据来源追踪、模型问责制、可解释性、偏见审查和人工监督。治理框架必须从静态政策转变为与快速发展的AI生命周期保持一致的动态控制机制。

将阿联酋人工智能宪章融入企业治理也提供了战略优势。它标志着为未来合规做好准备，能够实现风险意识创新，并确保人工智能部署不仅合法，而且与公众期望和社会价值观相一致。较早将这些原则付诸运营的组织将更善于管理伦理困境，回应监管机构的询问，并与用户、监管机构和更广泛的社会建立持久信任。

主动践行这些原则不仅确保了合规准备，也带来了清晰的业务价值。尽早将负责任的AI实践嵌入到组织中的，可以自信地加速创新，降低合规成本，并提升作为值得信赖、具有远见卓识的领导者的声誉。通过构建透明、包容和以人为本的AI系统，企业可以解锁新的机遇，获得利益相关者的信任，并在日益AI驱动的经济中实现差异化。

在全球范围内，欧盟、加拿大、美国和新加坡等司法管辖区正迅速将人工智能伦理纳入具有约束力的立法和运营框架。这预示着一个全球性转变，即人工智能治理将不再是非强制性的——而是数字竞争力和企业韧性的核心组成部分。

阿联酋宪章：12项人工智能原则

1. 加强人机联系:
阿联酋旨在加强人工智能与人类之间的和谐互利关系，确保所有人工智能发展都优先考虑人类福祉和进步。

2. 安全：
阿拉伯联合酋长国高度重视安全，确保所有人工智能系统符合最高的安全标准。该国鼓励修改或移除有风险的系统。

3. 算法偏见:
阿拉伯联合酋长国旨在应对人工智能算法带来的挑战，包括算法偏见，以促进所有社区成员的公平公正环境。这推动了人工智能技术的负责任发展，使其包容且易于人人使用，支持多元化并尊重个体差异。它确保了平等的技术利益，并在无排斥或歧视的情况下提升了生活质量。

4. 数据隐私：
根据阿联酋对隐私权的立场，虽然数据对人工智能的发展至关重要，支持并促进人工智能创新，但社区成员的隐私仍然是最受重视的。

5. 透明度：
阿拉伯联合酋长国寻求对人工智能及其系统如何运作和做决策形成清晰的认识，这有助于建立信任，增强责任，并促进在使用这些技术的问责制。

6. 人工监督：
该宪章强调了人类判断和人类监督相对于人工智能的不可替代的价值，并与伦理价值观和社会标准保持一致，以纠正可能出现的任何错误或偏见。

7. 治理和问责：

阿联酋采取负责任和主动的立场，强调人工智能治理和问责制的重要性，以确保该技术得到合乎道德和透明的使用。

8. 技术卓越：

人工智能应成为创新的灯塔，反映阿联酋在数字化、科技和科学卓越方面的愿景。阿联酋通过采用人工智能的科技卓越性，寻求全球领导地位，以驱动创新，增强竞争力，并通过针对复杂挑战的创新和有效解决方案提高生活质量，从而为社会整体带来可持续的进步。

9. 人类承诺：

人工智能中的人类承诺体现了阿联酋的精神，这对于确保该技术的发展服务于公共利益至关重要。它专注于提升人类的福祉和保护基本权利，强调将人类价值观置于技术革新的核心，以确保对社会产生积极而持久的影响。

10. 与人工智能和平共处：

与人工智能和平共处对于确保技术不损害人类安全或基本权利，同时增进我们社区的幸福感和进步至关重要。

11. 推动人工智能意识，共创包容未来：

创建一个包容的未来至关重要，确保每个人都能从人工智能进步中受益，并保证社会各阶层都能公平地获得这项技术和其优势。

12. 对条约和适用法律的承诺：

阿联酋强调在人工智能的发展和使用中遵守国际条约和地方法律的重要性。

KPMG 可信赖人工智能框架

随着人工智能日益成为关键决策和日常运营的有机组成部分，普华永道开发了其可信人工智能框架，以帮助组织应对这一不断变化的格局。该框架为人工智能生命周期带来了结构、问责制和清晰度，确保人工智能系统从战略到部署都符合伦理道德、透明度，并与人类价值观保持一致。

基于安永在各个行业的全球经验，该框架基于十个核心原则。这些原则包括公平性和透明度，确保人工智能系统具有包容性和可理解性；可解释性和问责制，促进人类监督和责任；以及隐私、安全和安全，保护个人和系统。此外，该框架强调数据完整性和可靠性，以确保人工智能的持续性能，以及可持续性，以确保人工智能的进步有助于更广泛的社会和环境目标。

阿联酋人工智能宪章反映了对负责任人工智能发展的类似承诺，通过十二条指导原则表达了国家愿景，这些原则与 KPMG 可信赖人工智能框架中的原则紧密一致。下表说明了阿联酋人工智能原则如何映射到 KPMG 的一个或多个可信赖人工智能原则：

	阿联酋人工智能原则	对齐的KPMG全球可信人工智能原则
1.	强化人机联系	可解释性，公平性，问责制
2.	安全	安全，可靠，安全
3.	算法偏见	公平性、透明度、数据完整性
4.	数据隐私	隐私，数据完整性
5.	透明度	透明度，可解释性
6.	人工监督	责任透明，可解释性
7.	治理与问责	责任追究，透明度，数据完整性
8.	技术卓越	可靠性，可持续性
9.	人类承诺	公平性、可持续性、问责制
10.	与人工智能和平共存	安全，安全，公平
11.	促进人工智能意识，共创包容未来	公平性，可解释性
12.	履行条约和适用法律	责任、隐私、数据完整性

阿联酋人工智能宪章与KPMG可信人工智能框架之间的紧密契合为行动提供了坚实的基础。可信人工智能原则已经通过定义的方法论在人工智能生命周期中得到实施——涵盖战略和设计、数据赋能、模型开发、测试和评估，以及部署和监控。在此经过验证的基础之上，本白皮书应用了相同结构化的方法处理阿联酋的十二条人工智能原则。对于每一条，都概述了实际步骤，以帮助组织嵌入符合伦理、以人为本的人工智能实践，并将原则转化为具体成果。

原则1

强化人机联系

阿联酋旨在加强人工智能与人类之间的和谐互利关系，确保所有人工智能发展都优先考虑人类福祉和进步。

用现实世界的术语理解原理

这一原则旨在确保人工智能系统增强和扩展人类能力，通过创造更智能、更具包容性的解决方案，赋予人类超越自身潜能的力量。人工智能应与伦理原则保持一致，尊重人的尊严、权利和价值观。最终，阿联酋的目标是营造一个人工智能与人类协作的环境，以改善生活质量、提高生产力和推动社会进步。

现实世界中的例子：

- **医疗人工智能：** 医疗保健中使用的AI系统，辅助医生更准确、高效地诊断疾病，最终改善患者结果。

- **智慧城市：** 将人工智能驱动技术融入城市规划以改善基础设施、优化交通流量并提升居民生活质量。

将此原则嵌入人工智能治理

为加强人机联系，应重点开发能够增强人类能力、提升福祉并为员工、顾客和社会带来积极影响的AI系统。确保您的AI计划与最佳实践保持一致，并在开发与实施的每个阶段审慎考虑其对人类价值观、尊严和权利的更广泛影响，同时体现阿联酋的文化和社会价值观。考虑将人机协同决策机制纳入其中，以进一步加强人机关系，确保有效的监督、信任和问责。

最佳实践与方法论

策略与设计

• **以人为本设计:** 设计人工智能系统，通过持续收集多样性反馈并将其用于改进和增强人工智能的影响力，以增强人类能力。

• **人工智能伦理目标：** 为人工智能发展制定明确的伦理准则，优先考虑人的福祉并解决潜在的负面影响，确保与全球标准及阿联酋的文化价值观保持一致。

• **人机回路集成：** 应尽早嵌入人机协同机制，以强化决策能力，确保问责制，并使人工智能系统与核心人类价值观保持一致。

数据赋能

• **数据敏感性：** 通过确保数据收集和处理尊重人类隐私和尊严，建立信任，确保个人和敏感数据的负责任使用。
• **幸福指标：** 在评估用于训练人工智能模型的训练数据时，应考虑诸如福祉、安全性和用户体验等因素。
• **丰富多样的数据表示：** 使用反映多样化人类体验的数据集，确保人工智能系统能够服务于广泛的社会需求。

模型开发

• **人机协作功能：** 开发人工智能系统，通过提供洞察力和支持来增强人类能力，而不取代人类决策。

• **透明算法：** 构建模型，使人类能够轻松地理解、信任并与人工智能系统协作。透明度有助于确保人类监督得以维持。

• **消减偏差：** 确保人工智能模型不带可能损害人类进步的偏见，包括确保跨不同群体间的公平对待，并维护社会价值观。

测试与评估

• **人类影响评估：** 评估人工智能系统对人类福祉的潜在积极和消极影响，确保结果与预期效益保持一致。
• **用户反馈:** 结合用户反馈来优化人工智能系统，确保它们与人类需求和进步相关。

部署和监控

• **持续协作：** 与人类用户和利益相关者保持积极合作，以确保已部署的 AI 系统能够持续带来益处并促进人类进步。

- **监测人工智能对人类的影响：** 追踪人工智能对社会产生的长期影响，确保人工智能系统继续优先考虑和增强人类福祉。
- **适应人类需求：** 持续适应人工智能技术以满足不断变化的需求人类用户的价值，尤其是作为社会上下文会变化。

将原理扩展到智能代理系统

智能代理系统必须设计为补充而非取代人类角色。它们应通过情境感知和反馈机制增强人类决策力和生产力。确保直观的人类交互和透明度将有助于维持信任。组织必须优先考虑代理-人工智能界面中的用户体验。用户在情感和认知层面的影响应被持续监控并改进。

关键工具、技术与进一步阅读

工具和技术：

• 以人为本的人工智能设计框架 • 人机协作工具包（例如：微软小慧、Salesforce Einstein）• 用户体验研究和认知负荷测试工具（例如：Optimal Workshop）• 普华永道可信人工智能框架 • 普华永道可信人工智能风险和控制矩阵（RCM）

拓展阅读：

• 斯坦福 HAI：人机协作研究 • 哈佛伯克曼克莱恩中心：增强伦理 • 微软：计算的未来——AI 与人类价值观

原则2

安全

阿联酋高度重视安全，确保所有人工智能系统符合最高安全标准。该国家鼓励修改或移除有风险的系统。

用现实世界的术语理解原理

人工智能安全是指确保人工智能系统按预期运行，不对个人、企业或社会造成危害。这包括技术稳健性、风险缓解以及纳入安全措施以防止或应对意料之外的结果或故障。欧盟人工智能法案通过强制对高风险人工智能应用设定严格的安全标准， exemplifies this approach。优先考虑安全对于最大限度地降低风险、保持运营连续性和阿联酋公众信任至关重要。

现实世界中的例子：

• **自动驾驶汽车:** 人工智能驱动的汽车在低能见度条件下无法识别行人，导致事故和监管审查。

• **医疗人工智能：** 诊断人工智能错误解释医学影像，导致错误的治疗和潜在的法律责任风险。

将此原则嵌入人工智能治理

确保人工智能安全需要结构化方法——从风险评估到持续测试和失效安全机制。像KPMG的可信人工智能风险框架这样的工具通过提供结构化方法论来支持这一过程，以识别、评估和减轻与人工智能相关的风险，包括与安全相关的风险，并符合ISO 42001和欧盟人工智能法案等标准。结合强大的人工智能治理框架，这些工具有助于确保安全措施保持有效、透明，并与本地及全球最佳实践保持一致。通过将安全最佳实践嵌入到开发的每个阶段，组织可以提高可靠性、维持合规性并建立信任。

最佳实践与方法论

策略与设计

• **安全目标与指标：** 为人工智能计划建立明确的安全目标和指标，重点关注可靠性、弹性、透明度和安全性。像KPMG的人工智能指标这样的工具可以衡量绩效并确保符合道德规范。
• **风险识别：** 定义与人工智能系统相关的安全风险，并建立风险缓解协议。
• **利益相关者咨询：** 在开发前，就安全问题与监管机构、行业专家和最终用户进行沟通。
• **安全设计** 确保人工智能系统的设计具有明确的覆盖机制，以防故障时造成危害。包含降级机制、监控和人工参与。

数据赋能

• **数据完整性校验：** 验证训练数据的准确性、完整性和一致性，以防止人工智能故障。
• **偏差和异常检测:** 识别可能导致不安全人工智能行为的偏见，例如医疗保健或自主系统中的错误分类。
• **仿真和压力测试数据：** 在多种场景上训练人工智能模型，包括边缘情况，以确保在实际应用中的鲁棒性。

模型开发

• **安全意识算法：** 实现优先考虑安全的算法，包含护栏和约束以防止有害决策。

• **安全保障机制：** 在最终设计中嵌入容错机制，如人工介入和日志记录。

• **可解释性和透明度：** 确保人工智能决策可以被理解和审计，以便在部署前识别潜在的安全风险。

测试与评估

• **对抗测试：** 通过针对潜在故障点进行测试来识别 AI 系统的薄弱环节，并在部署前建立纠正措施。

• **测试边界情况：** 评估人工智能在极端条件下的性能（例如自动驾驶汽车的能见度低或金融市场的不可预测波动）。

- **安全基准测试：** 定义和测量行业标准安全性。

部署和监控

• **持续安全监测：** 定期审计部署后的ai系统以检测异常或故障。
• **事件响应计划：** 建立清晰的AI故障升级协议，以确保快速补救。
• **监管合规报告：** 在公司内部记录和沟通安全措施，以证明遵守安全标准。

将原理扩展到智能代理系统

自主性人工智能引入动态决策，这需要实时风险检测和缓解能力。安全协议不仅必须嵌入代码中，还必须嵌入智能体与系统和人员互动的方式中。安全措施和升级到人类监督者的路径是必不可少的。必须优先进行针对对抗性或不预期智能体行为的模拟测试。组织应跟踪智能体的行为以确保问责制。

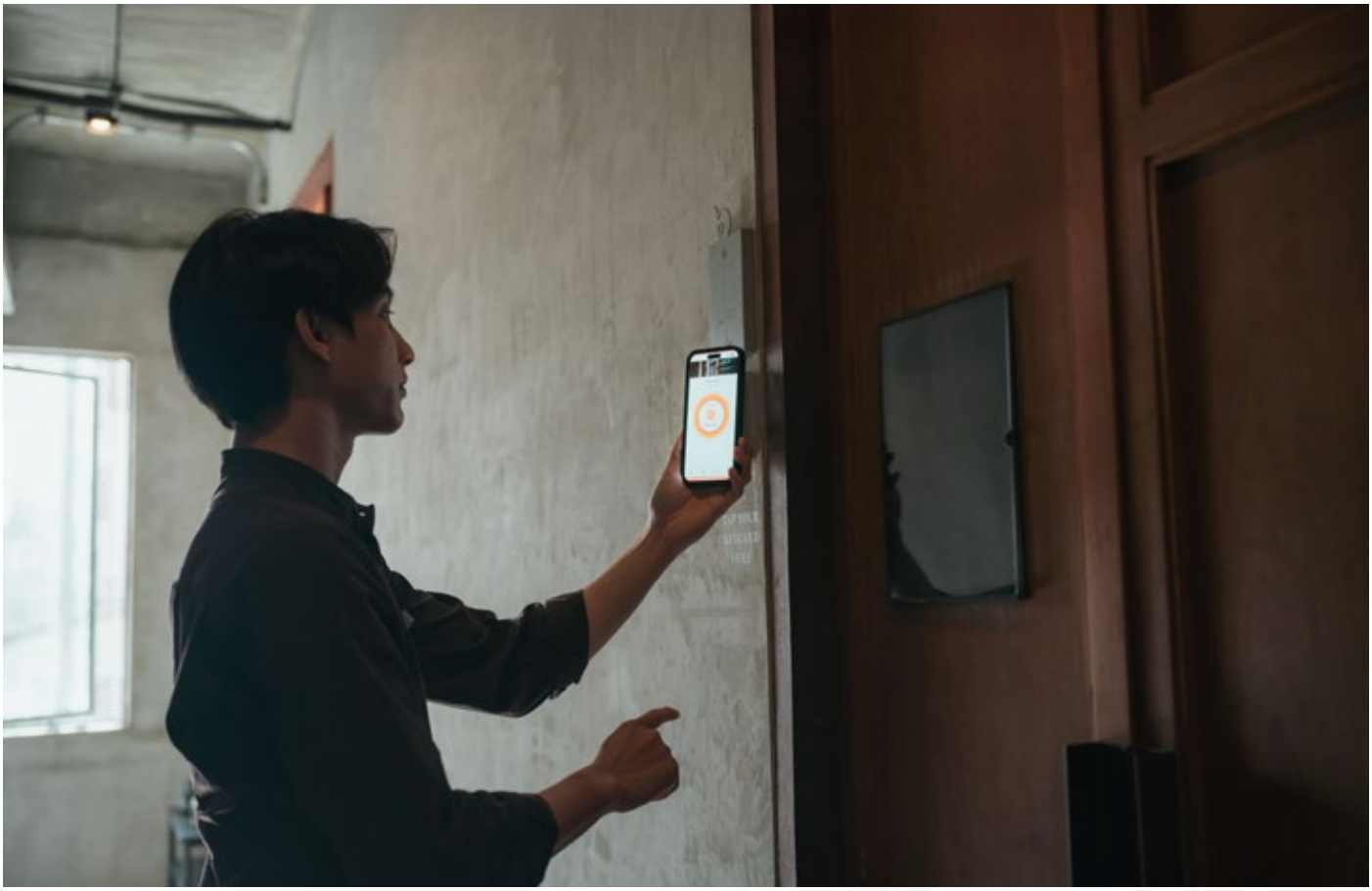
关键工具、技术与进一步阅读

工具和技术：

• 对抗测试框架（例如CleverHans、Foolbox）• 形式化验证工具（例如TLA+、Z3）• 贝叶斯网络、蒙特卡洛dropout • 红队和仿真实验室 • KPMG可信人工智能框架 • KPMG可信人工智能风险和控制矩阵（RCM）

拓展阅读：

• NIST人工智能风险管理框架 • OpenAI的系统安全实践 • 欧盟人工智能法案：高风险系统的安全规定



原则三

算法偏见

阿拉伯联合酋长国旨在应对人工智能算法带来的挑战，包括算法偏见，以促进所有社区成员的公平公正环境。这推动了人工智能技术的负责任发展，使其包容且易于人人使用，支持多元化并尊重个体差异。它确保了平等的技术利益，并在无排斥或歧视的情况下提升了生活质量。

用现实世界的术语理解原理

在实践中，算法偏见发生时，人工智能系统会基于性别、种族、年龄或社会经济地位等因素，无意中偏向或歧视某些群体。这可能是由于有偏见的数据训练、模型假设存在缺陷，或在开发过程中缺乏多样化的代表性。解决偏见对于建立信任、确保公平以及减轻财务、法律和声誉风险至关重要。

现实世界中的例子：

- **招聘系统：** 人工智能系统基于男性主导行业的有偏见的训练数据拒绝女性候选人，导致公司面临歧视指控或监管处罚。

- **贷款批准：** 基于人工智能的信贷系统根据历史上的歧视性做法拒绝贷款申请人，可能违反公平放贷法律。

将此原则嵌入人工智能治理

要有效解决您AI系统中的算法偏见，应将公平性嵌入开发的每个阶段。这包括采取主动的设计选择，使用具有代表性和平衡性的数据，并进行持续测试以确保公平的结果。有效的AI治理，在强大的框架支持下，应贯穿每个阶段，以确保问责制、透明度，并符合道德和地方监管标准。通过这样做，组织可以将公平性原则转化为可执行、具有影响力的步骤，从而推动符合阿联酋要求的负责任AI发展。

最佳实践与方法论

策略与设计

• **定义公平目标：** 清晰明确地制定人工智能计划的目标，为每个人工智能计划设定衡量指标，识别潜在的偏见，并确保代表不同利益相关者群体的需求。这应受您组织的人工智能治理框架的指导和协调。

• **包容式设计：** 在设计阶段吸纳多元化团队——涵盖伦理学家到受影响的社区成员——以识别并解决潜在的偏见来源，在它们出现之前。

• **采用可解释人工智能 (XAI):** 确保人工智能模型决策过程的透明性，通过实施可解释人工智能框架，使利益相关者能够理解并审计人工智能输出。

数据赋能

•**确保代表性：** 确保数据集能够准确反映多样的群体特征（年龄、性别、民族、社会经济地位等），特别是在招聘、医疗保健和金融等领域。
• **偏见审计：** 定期对训练数据集进行审计，以识别和减轻历史和社会偏见，包括直接偏见、代理偏见、采样偏见和测量偏见。
• **数据标注：** 提高数据标注的质量和准确性，以防止主观或带有偏见的标注。确保数据标签反映所代表人群的多样性。

模型开发

• **模型设计时的偏差检测：** 用于检测潜在代理偏差的屏幕AI模型——例如邮政编码或收入水平等变量，可能会无意中反映种族或性别等敏感属性。

• **公平感知算法：** 使用能够通过调整模型预测或结果的失衡来识别和纠正偏见的公平性增强算法。

• **文档模型选择：** 记录模型开发过程中做出的关键决策背后的理由，包括特征和算法的选择，以确保透明度和问责制。

测试与评估

• **阈值设置：** 在部署之前，定义与构思阶段设定的公平性目标和指标相一致的、可接受的公平性阈值。
• **冲击试验：** 测试完全训练好的模型是否对公平性阈值存在偏差，评估人工智能系统结果在不同人群中的差异，并根据需要进行调整以确保公平的结果。

部署和监控

• **持续监测:** 持续监控部署后的人工智能系统的性能，以确保公平性并符合治理框架。确保系统能够适应不断发展的社会规范和期望。

• **反馈机制：** 建立机制，允许用户和利益相关者报告有偏见的结果或问题。

• **定期报告：** 定期发布偏见影响测试报告，让内部和外部利益相关者都能对组织进行问责。

关键工具、技术与进一步阅读

工具和技术：

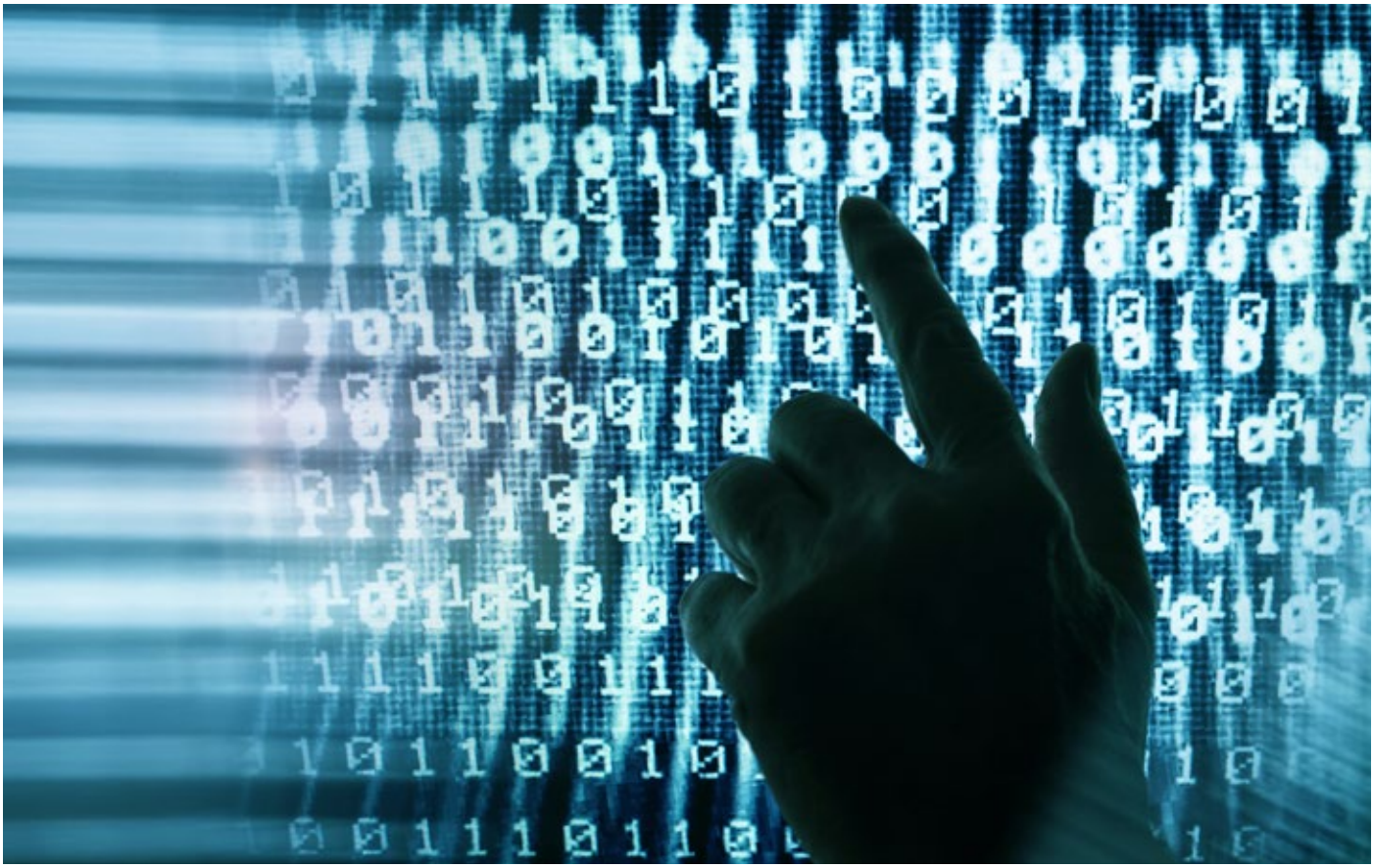
• IBM 公平性 360 • 微软公平学习 • 谷歌假设工具 • SHAP、LIME 用于可解释公平性分析 • 普华永道可信人工智能框架 • 普华永道可信人工智能风险与控制矩阵 (RCM)

将原理扩展到智能代理系统

在自主AI中的偏见可以通过自主循环和自适应行为而加剧。随着智能体进化他们的决策规则，必须持续评估公平性。需要多样化的训练环境和动态的偏见审计。智能体必须被限制通过反馈循环来强化系统性偏见。干预措施必须通过伦理审查委员会进行记录和管理。

拓展阅读：

• 世界经济论坛：公平工具包 • 人工智能伙伴关系：管理人工智能中的偏见 • 人工智能现在研究所：算法问责报告



原则4

数据隐私

根据阿联酋对隐私权的立场，虽然数据对人工智能的发展至关重要，支持并促进人工智能创新，但社区成员的隐私仍然是最受重视的。

用现实世界的术语理解原理

人工智能中的数据隐私是指在整个人工智能生命周期中对个人和敏感信息的保护——从数据收集和存储到模型训练和部署。由于人工智能模型依赖于庞大的数据集，组织必须确保个人信息保持机密，并负责任地使用，同时保持透明度和获得同意。未能保护隐私可能导致监管处罚、声誉损害和公众信任的削弱，所有这些都可能破坏长期创新努力。

现实世界中的例子：

- **零售AI：** 使用客户购买历史进行个性化营销而不征得明确同意，导致客户投诉和数据保护违规。

- **医疗人工智能：** 在未经适当同意的情况下，医院与第三方共享患者数据进行训练的医学AI模型，导致保密性泄露和法律诉讼。

将此原则嵌入人工智能治理

组织应该在人工智能的每个生命周期阶段都将数据隐私纳入考量。这意味着限制数据收集范围，获取明确同意，确保安全存储，并将隐私设计嵌入系统架构。问责结构应清晰明确，明确隐私监督的负责人。安全的数据处理、明确的同意以及个人信息保护措施等因素在欧盟人工智能法案中也得到强调——这进一步强化了组织应将隐私视为核心设计原则而非事后考虑的需求。这样做能够降低监管风险，建立信任，并使组织成为负责任的领导者。

最佳实践与方法论

策略与设计

- **设置隐私目标：** 为每个人工智能系统定义隐私标准，与法律和组织指南保持一致，并定义指标以在整个人工智能生命周期中监控合规情况。
- **隐私影响评估 (PIA):** 在早期规划阶段开展隐私影响评估，以预见和减轻隐私风险。
- **隐私性和机密性设计：**
将数据最小化、匿名化、退出选择权等设计原则融入系统架构和用户体验中。
- **第三方模型评估：** 在使用第三方模型时，评估供应商的数据隐私和保密协议，以确保它们符合所需标准。

模型开发

- **保护隐私技术：** 结合差分隐私、联邦学习等技术，在模型训练过程中保护个人数据。

- **模型可解释性：** 提供清晰文档说明数据如何使用，以及模型如何利用该数据做出决策。

测试与评估

- **数据泄露测试：** 验证模型不会无意中泄露敏感数据。
- **隐私压力测试：** 模拟潜在的隐私泄露场景，并验证安全防护措施的有效性。
- **风险阈值：** 定义与隐私相关的可接受风险阈值，并在部署前确保模型符合要求。

数据赋能

- **数据最小化和匿名化：** 限制收集个人数据，并在可能的情况下对数据集进行匿名化。
- **同意与使用清晰：** 确保所有数据均基于知情同意、明确授权收集，维护审计追踪，并提供清晰的数据使用条款和退出机制。
- **访问和分类控制：** 仅限授权人员访问数据，记录所有活动，并实施标签分类法以标记受保护数据和使用受限属性。
- **负责聚合和第三方使用**
聚合数据源时保护敏感信息，并验证第三方数据符合隐私和合同标准。

部署和监控

- **实时监控：** 持续监控人工智能系统，确保隐私合规问题或异常情况。
- **事件响应协议：** 建立针对潜在数据泄露或滥用事件的升级程序。
- **隐私报告：** 定期发布隐私控制和事件报告，以促进利益相关者的信任。

将原理扩展到智能代理系统

智能体式人工智能系统通常依赖于持续的数据摄取和上下文感知能力，从而引发复杂的隐私问题。同意机制必须实时适应，并且数据使用必须可审计。智能体应在数据最小化和按需知晓原则下运行。联邦学习或边缘处理可以减少中心数据暴露。隐私风险评估应随着每个新智能体能力的出现而更新。

关键工具、技术与进一步阅读

工具和技术：

• 差分隐私库（谷歌DP，开放DP）• TensorFlow隐私 • 数据掩码和匿名化工具（例如ARX，Privitar）• 同意管理平台（例如OneTrust）• KPMG可信人工智能框架 • KPMG可信人工智能风险和控制矩阵（RCM）

拓展阅读：

• 阿联酋联邦数据保护法 • 欧盟人工智能指南 • NIST隐私框架

原则五

透明度

阿联酋寻求对人工智能及其系统如何运行和做决策形成清晰的认识，这有助于建立信任，提高这些技术使用中的责任和问责制。

用现实世界的术语理解原理

人工智能透明度通过提供清晰的洞察，确保人工智能系统在整个生命周期中的运作方式，从而向利益相关者进行负责任的披露。它确保利益相关者了解何时在与人工智能交互、决策是如何做出的，以及涉及哪些数据或限制。透明的系统支持问责制，实现知情监督，并通过使人工智能对用户、监管机构和受影响方易于理解和解释来建立信任。没有透明度，人工智能可能会成为一个“黑箱”，增加出现有偏见、有害或无法挑战的结果的风险。

现实世界中的例子：

- **贷款批准：** 人工智能系统未经解释拒绝客户贷款，导致客户不满且无法申诉。

- **医疗人工智能：** 患者不明白为什么人工智能系统推荐某一种治疗方法而不是另一种，导致对系统的信任度降低，并可能对医疗决策构成挑战。

将此原则嵌入人工智能治理

为确保人工智能透明度，企业应采用可解释的人工智能模型，维护清晰的文档，并在整个人工智能生命周期中嵌入负责任的披露机制。通过在强大的人工智能治理框架内整合最佳实践，组织可以与阿联酋的透明度原则保持一致，同时培养对人工智能驱动决策的信任和信心。

最佳实践与方法论

策略与设计

• **识别利益相关者：** 记录所有用户组，并定义他们将如何与AI系统交互。
• **定义透明度目标：** 识别不同用户（例如监管者、客户、内部团队）需要何种程度的可解释性。
• **设计信息披露机制：** 在用户与人工智能交互时通知用户，获取数据使用的同意，公开数据刷新频率，并沟通已知数据或模型局限性。
• **可解释性设计：** 在构建人工智能系统时，应考虑可解释性，同时定义成功标准和指标，并考虑利益相关者的需求。

数据赋能

• **数据溯源追踪：** 维护数据来源记录，确保训练数据可追溯和可审计。
• **数据可解释性:** 确保人工智能模型使用的输入数据是可理解的，并且有详细的文档记录，以便利益相关者能够追溯数据如何影响人工智能决策。
• **用户可解释的数据输入:** 确保 AI 模型依赖于逻辑上可以解释的输入，避免使用晦涩或不可透明的变量。

模型开发

• **可解释人工智能 (XAI):** 尽可能使用可解释的模型，或实现解释复杂模型如何做出决策的技术。
• **特征重要性分析：** 清晰记录影响AI决策的关键特征，有助于阐明哪些数据点对结果影响最大。

- **开发披露机制:** 集成安全漏洞负责任披露工具和框架直接输入到模型架构中。
- **决策记录和审计追踪：** 维护用于启用人工智能决策过程日志透明度，促进审计并支持发生争议时的评论。

测试与评估

• **可解释性用户测试：** 使用最终用户测试人工智能系统，以确定解释是否易于理解和可操作。
• **公平性和可解释性指标：** 建立可量化的指标来衡量AI决策的透明度，并在必要时进行调整。

部署和监控

• **持续监测：** 持续跟踪人工智能决策，并更新治理框架以适应并确保模型保持可解释性。
• **发布系统文档：** 清晰解释模型类型、预期用途、运行条件、限制和数据来源。
• **透明度报告：** 发布定期报告，概述人工智能系统如何运行、其能力和限制，包括潜在风险和安全保障措施，以保持问责制。

将原理扩展到智能代理系统

使用智能代理，决策链可能跨越多个模型和交互。组织必须确保代理能够解释它们的选择并提供可追溯的逻辑。应设计可视或叙述性的解释以便用户理解。透明日志应包括触发器、意图和推理。可能需要基于角色的可解释性——为用户、审计员和监管者。必须定期评估代理行为与声明目的的一致性。

关键工具、技术与进一步阅读

工具和技术：

• SHAP, LIME, ELI5, Skater • 模型卡，数据集数据表 • 决策日志和审计追踪工具 • 可解释人工智能框架（XAI）• 普华永道可信人工智能框架 • 普华永道可信人工智能风险与控制矩阵（RCM）

拓展阅读：

• 谷歌PAIR指南 • 经合组织人工智能原则——透明性 • AI可解释性360 (IBM)

第六原则

人工监督

阿联酋强调人类判断和对人工智能的监督的不可替代的价值，确保人工智能系统与道德价值观和社会标准保持一致，以纠正可能出现的任何错误或偏见。

用现实世界的术语理解原理

人类监督确保人工智能系统并非决策的唯一责任者，而是人类保持控制权，特别是在高风险情况下。虽然人工智能可以协助处理大量数据并根据模式做出决策，但人类判断对于确保这些决策符合伦理、无偏见且与社会价值观一致至关重要。

现实世界中的例子：

- **医疗人工智能：** 在医疗人工智能应用中，人类医生必须监督诊断和治疗建议，确保建议符合患者健康。

- **自动驾驶汽车:** 在自动驾驶中，人类监督是必要的，以便在紧急情况下进行干预并确保安全规程得到遵守。

将此原则嵌入人工智能治理

为确保有效的人类监督，组织应在其人工智能治理中纳入能够实现人类干预和判断的机制。这包括构建允许人类监督者轻松监控、覆盖或调整人工智能决策的系统，特别是在决策可能产生重大社会或伦理影响的情况下。采用这些做法与阿联酋的人类监督原则和国际标准相符，例如欧盟人工智能法案，该法案强调人类干预的重要性，特别是在高风险人工智能应用中。确保人类监督可以帮助组织满足监管要求并避免完全自主决策带来的风险。

最佳实践与方法论

策略与设计

• **设计用于人机协作：** 研发优先考虑人类决策的AI系统，允许AI协助但不取代人类判断，特别是在高风险环境中。

• **为人工监督创建明确的角色：** 清晰界定人工监督在人工智能系统中的作用，确保在需要时存在明确定义的干预协议。

数据赋能

• **偏置检测与校正:** 实施人在回路机制，定期评估和纠正训练数据中的偏差，确保人工智能模型保持公平。

• **数据伦理审查:** 进行伦理审查以确保所使用的数据与社会价值观一致，并且不会导致不公平或带有偏见的决策。

模型开发

• **人工智能决策透明度：** 确保人工智能模型决策过程透明化，使人能够理解决策是如何做出的，并在必要时进行干预。

• **人在回路系统：** 在构建人类判断是关键组成部分的系统中，特别是在医疗保健、金融和执法等高风险领域，需要人类专业知识以确保道德和公平的结果。

测试与评估

• **以人为本的测试：** 在评估阶段让人类测试人员参与，以模拟现实世界的干预。

• **审计偏差和公平性：** 定期对人工智能系统进行审计，以检测可能需要人工纠正或干预的任何偏见、错误或伦理问题。

部署和监控

• **持续人工监控：** 确保人类主管能够实时监控人工智能系统，尤其是在关键应用中。

• **紧急干预规程：** 针对人工智能系统故障、不准确或伦理困境的情况，建立清晰的人为干预协议，确保人类监督能够迅速实施。

将原理扩展到智能代理系统

自主式AI由于自主性和涌现行为挑战传统监督。组织必须定义清晰的人工干预阈值。设计人工在环和覆盖场景的协议。监督机制必须包括实时警报、回滚选项和升级工作流。治理团队应定期审查代理的性能和决策。人类问责制应始终可见。

关键工具、技术与进一步阅读

工具和技术：

• 人在回路平台 (HITL) • 亚马逊 SageMaker Ground Truth • 模型升级协议模板 • 审计追踪平台（例如 Fiddler AI） • 普华永道可信人工智能框架 • 普华永道可信人工智能风险与控制矩阵 (RCM)

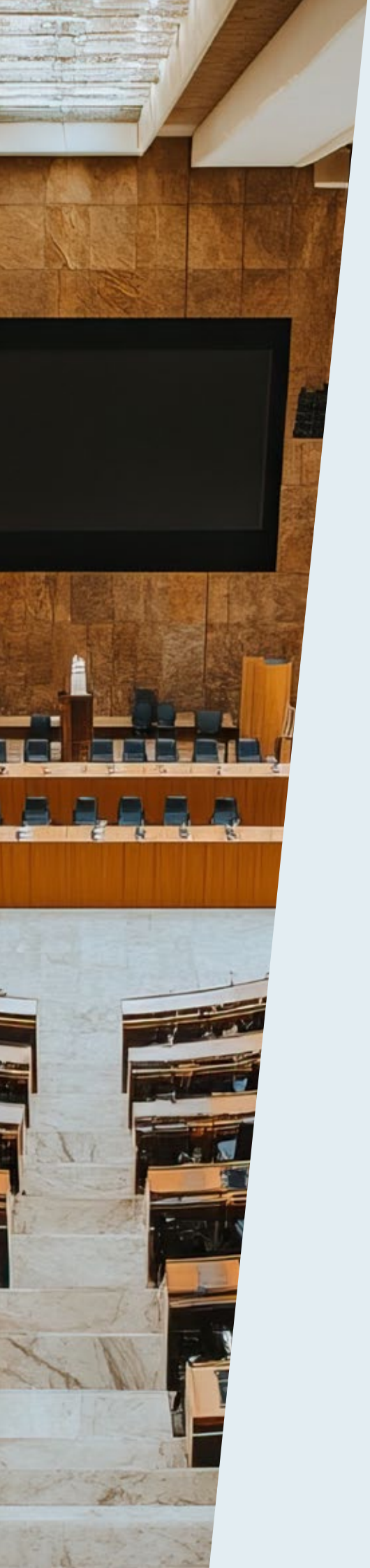
拓展阅读：

• 欧盟人工智能法案：人类监督 • 经合组织原则：以人为本的人工智能 • ISO/IEC 42001 人工智能管理体系草案标准

原则七

治理与问责

阿联酋采取负责任和主动的立场，强调人工智能治理和问责制的重要性，以确保该技术得到合乎道德和透明的使用。



用现实世界的术语理解原理

人工智能治理和问责制是指建立明确监督机制，确保负责任的决策，并保持人工智能生命周期的透明度。这包括明确角色和职责，制定明确政策，以及实施监控、审计和报告的机制。缺乏强有力的治理，人工智能系统面临变得不透明或被误用的风险，可能引发歧视、不公平行为或损害公众信任等意外后果。

现实世界中的例子：

- **金融机构：** 如果缺乏治理结构来确保贷款决策过程中的公平、透明和问责制，那么负责贷款决策的人工智能系统可能会导致结果存在偏见。

- **医疗保健:** 缺乏问责制，人工智能驱动的诊断可能在没有适当监管的情况下应用，导致有害决策，损害患者信任和法规遵从。

将此原则嵌入人工智能治理

要有效地将治理和问责制融入人工智能系统，阿联酋的组织必须制定涵盖人工智能全生命周期的综合框架，明确角色、责任和决策流程。借鉴欧盟人工智能法案等全球标准——这些标准强调监管机构和监督机制的重要性——组织可以建立类似的治理结构，以确保问责制并促进人工智能技术的道德使用。通过采用这些原则以及最佳实践，组织可以促进透明度、建立信任并遵守监管要求，将自己定位为阿联酋人工智能发展领域负责任和具有远见的领导者。

最佳实践与方法论

策略与设计

• **定义治理结构：** 建立明确的问责框架，为人工智能决策和监督分配角色和职责。这应包括技术和道德监督委员会。

• **制定道德准则：** 制定明确的伦理准则，以规范人工智能的使用，并确保这些准则在整个人工智能生命周期中得以遵守。

• **设计为人监督（HITL）：** 分配算法所有者，定义自动化级别，通过输入纠正行为，外部记录交互，并为输出检查应用MITL。

数据赋能

• **维护数据文档：** 维护数据来源和决策过程的透明文档记录，以支持问责制，使人们能够轻松地将AI结果追溯到所使用的数据。

• **整合HITL：** 在评估数据质量和标签中嵌入人工监督，以提高可靠性并减少偏差。

模型开发

• **文档AI开发选择：** 维护模型设计过程中做出的决策的清晰记录，包括数据选择、算法选择以及潜在伦理考量。

- **定期审计：** 实施持续审计以确保人工智能系统符合治理政策并满足透明度标准

• **开发 HITL 功能和接口：** 集成人在回路机制，以确保在关键决策点进行监督、验证和控制。

测试与评估

•**问责测试：** 通过确保决策过程在每个步骤中都可以解释和验证，对AI系统进行问责测试。

• **评价：** 定期进行伦理审查，以评估人工智能系统是否正按照与组织价值观和社会期望相符的方式使用。

部署和监控

• **持续监督：** 建立机制，对人工智能系统进行持续监控，以确保它们在其整个生命周期中保持道德、透明和可问责。

• **反馈回路：** 创建反馈回路，允许利益相关者（包括客户、员工和监管者）报告任何道德或治理问题。

将原理扩展到智能代理系统

智能体式AI需要具有分布式但可审计的治理框架。为智能体的训练、部署和适应定义明确的所有权。实施与组织伦理相符的护栏和智能体行为政策。创建“智能体护照”以跟踪进化、授权级别和决策。将智能体嵌入现有的AI治理委员会和审查周期。在智能体决策与负责的人类角色之间建立问责制链接。

关键工具、技术与进一步阅读

工具和技术：

• 人工智能治理平台（例如IBM Watsonx.治理，Credo AI）• 算法风险登记册 • 模型文档模板（例如模型卡）• 道德审查委员会指南 • 普华永道值得信赖的人工智能框架 • 普华永道值得信赖的人工智能风险和控制矩阵（RCM）

拓展阅读：

• WEF模型人工智能治理框架 • ISO/IEC 38507:信息技术治理—人工智能指南 • 阿联酋人工智能伦理与治理倡议

原则八

技术卓越

人工智能应当是创新的灯塔，反映阿联酋在数字、技术和科学卓越方面的愿景。阿联酋通过采用人工智能领域的卓越技术寻求全球领导地位，以推动创新，增强竞争力，并通过针对复杂挑战的创新和有效解决方案提高生活质量，从而为社会带来可持续发展的进步。

用现实世界的术语理解原理

人工智能的技术卓越性在于持续提升人工智能系统的能力以应对紧迫挑战、提高效率并推动经济增长。要实现卓越需要持续投入研发，并由强大的风险和控制机制（例如KPMG的可信赖人工智能框架）提供支持，该框架确保人工智能解决方案既高性能，又符合道德、可靠且安全。欧盟人工智能法案就是一个范例，它强调了严格测试、持续评估和明确性能标准，以在降低风险的同时促进创新。通过将可信赖框架嵌入人工智能的开发和部署中

组织可以在维护公众利益和确保监管一致性的同时，推动持续的技术进步。

现实世界中的例子：

- **连续模型改进：** 一家公司定期用新数据和新技术更新其人工智能模型，以提高准确性和性能，确保它们在变化的环境中保持有效和竞争力。

- **研发投入：** 一个组织成立专门的AI研究实验室，探索生成式AI、边缘计算和联邦学习等新兴技术，以保持创新的前沿。

将此原则嵌入人工智能治理

将技术卓越融入人工智能系统，组织应优先考虑持续创新、研究和开发——同时将这些努力与道德标准、监管框架及其定义的风险偏好保持一致。鼓励实验和快速迭代的环境能够推动突破，但必须以稳健的人工智能风险评估和明确的风险评估方法为基础。这些机制使组织能够在创新的同时，主动识别、管理和减轻潜在风险，通过明确定义的护栏来实现。

最佳实践与方法论

策略与设计

• **培育持续创新：** 鼓励组织内部进行研发，探索人工智能技术的新前沿。

• **定义性能标准:** 建立清晰的技术性能指标和基准，确保人工智能系统达到全球卓越标准。

数据赋能

• **确保高质量数据集：** 使用精心策划的数据集来支持健壮和适应性强的模型开发。

模型开发

• **利用先进技术：** 采用最新的AI算法和方法，以确保AI系统始终处于创新的前沿。

• **自我改进设计:** 构建能够持续学习和适应的模型，以保持技术相关性和性能。

测试与评估

• **绩效评估：** 实施严格的测试协议，以确保人工智能模型能够提供一致的高质量结果。

• **风险和获益评估：** 定期评估人工智能部署的潜在风险和收益，以确保它们在避免未预期后果的同时提供最大价值。

• **进行全面测试：** 定期评估人工智能系统，以确保其性能和功效达到技术卓越标准。

部署和监控

• **持续改进：** 在部署后继续演进人工智能系统，结合反馈和技术进步以保持高标准性能。

将原理扩展到智能代理系统

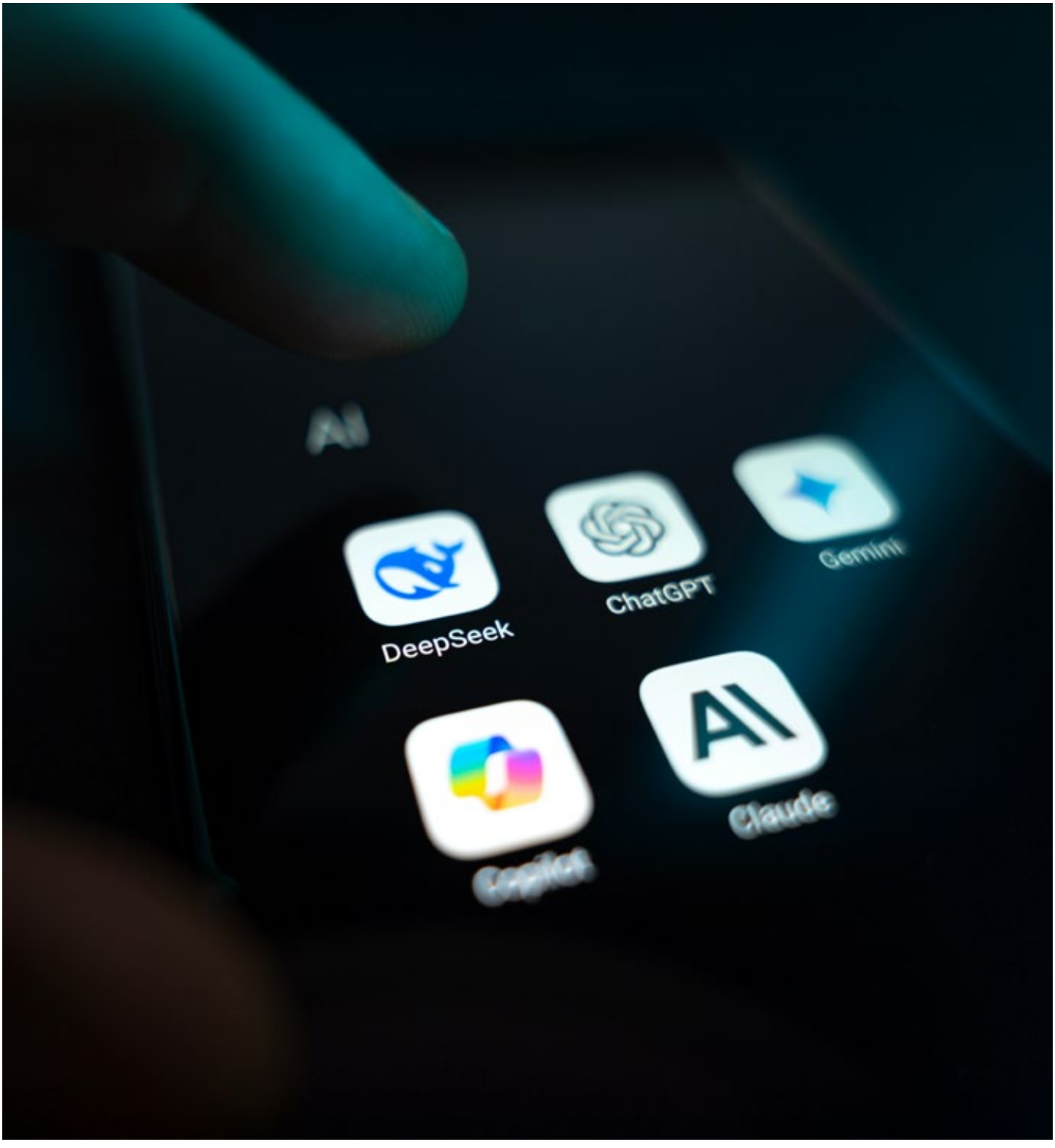
自主型人工智能必须在模型架构、互操作性和弹性方面反映领先标准。持续进行模型调优、场景测试和性能基准测试至关重要。使用先进的模拟环境来模拟复杂智能体行为。采用可组合架构以允许模块化升级。卓越也包括对对抗性操纵和非预期后果的鲁棒性。

关键工具、技术与进一步阅读

工具和技术：

• MLPerf基准测试工具 • 机器学习管道的CI/CD（例如MLflow、TFX） • NVIDIA Triton推理服务器 • 边缘AI联邦学习平台 • KPMG可信人工智能框架 • KPMG可信人工智能风险和控制矩阵

• 斯坦福人工智能指数报告 • 阿联酋国家人工智能战略2031 • 微软负责任人工智能标准



原则9

人类承诺

人工智能中的人类承诺体现了阿联酋的精神，这对于确保该技术的发展服务于公共利益至关重要。它专注于提升人类的福祉和保护基本权利，强调将人类价值观置于技术革新的核心，以确保对社会产生积极而持久的影响。

用现实世界的术语理解原理

人工智能中的人类承诺强调设计和部署优先考虑人类福祉和权利的人工智能系统，并与阿联酋的核心价值观和文化原则相一致。这一原则倡导创造服务于社会利益的人工智能技术，同时保障个人自由和尊严。阿联酋对人类承诺的关注确保人工智能创新惠及所有人，创造促进社会公益和保护基本自由解决方案。

现实世界中的例子：

- **医疗保健中的AI:** 基于人工智能的解决方案，帮助医生做出更好的决策，在保障患者权利和福祉的同时，确保更优质的病人护理。

- **用于残疾人无障碍的AI:** 为残疾人士设计的AI工具，确保教育、就业和日常服务的平等接入。

将此原则嵌入人工智能治理

要将人类承诺融入您的ai开发流程中，请优先考虑用户和社会的福祉。确保ai系统尊重基本权利，包括隐私、平等和自主权。建立强大的治理结构，确保ai模型具有包容性和透明性，重点在于为人类提供实实在在的利益，同时保护个人免受潜在伤害。

最佳实践与方法论

策略与设计

•**以人为本设计：** 开发人工智能系统，在设计和开发阶段始终优先考虑人类需求和福祉。
• **人工智能伦理治理：** 建立清晰的道德准则，将人权和社会福祉作为人工智能项目的核心。
• **公共物品框架：** 将人工智能发展与社会福利相结合，确保人工智能解决方案在实现组织目标和商业目标的同时，也能为公共利益做出积极贡献。

数据赋能

• **尊重隐私：** 确保用于训练人工智能模型的数据在收集、存储和处理时充分尊重个人的隐私和权利。
• **包容数据：** 使用代表不同人口群体的数据，确保人工智能模型满足所有社区的需求，特别是边缘化群体。
• **数据透明度：** 提供清晰的数据来源文档以及确保个人和敏感信息得到合乎道德处理所使用的流程。

模型开发

• **偏差缓解：** 实施策略以减少偏见，并确保人工智能模型不会不公正地损害或有利于某些个人或群体。

• **人工监督：** 在决策过程中保持人类监督机制，确保人工智能系统增强而非取代人类判断。

测试与评估

• **伦理影响评估：** 定期评估人工智能系统的伦理影响，包括潜在危害、偏见和不可预见的后果。

- **人类福祉测试：** 测试AI系统以评估其对人类福祉的影响，确保它们能做出积极贡献社会进步。
- **利益相关者参与：** 持续参与与多元化的利益相关者，包括普通公共的，为了理解社会影响人工智能系统并确保它们与公众兴趣也是如此。

部署和监控

• **持续人权监测：** 追踪人工智能系统的部署，以确保它们继续尊重人权，并且不会无意中伤害社区。
• **成果透明度：** 定期披露人工智能系统的成果和影响，维护公众信任并确保问责。
• **用于改进的反馈回路：** 实施持续反馈系统，以确保人工智能系统以提升人类福祉和社会公义的方式进化。

将原理扩展到智能代理系统

代理机构必须以人类福祉为最终目标进行设计。它们的运作应提升社会各方面的可及性、平等性和幸福感。代理机构必须适应用户偏好、能力和文化价值观。伦理约束应被硬编码以防止剥削或伤害。监控系统应评估人类影响并及早暴露任何伦理问题。

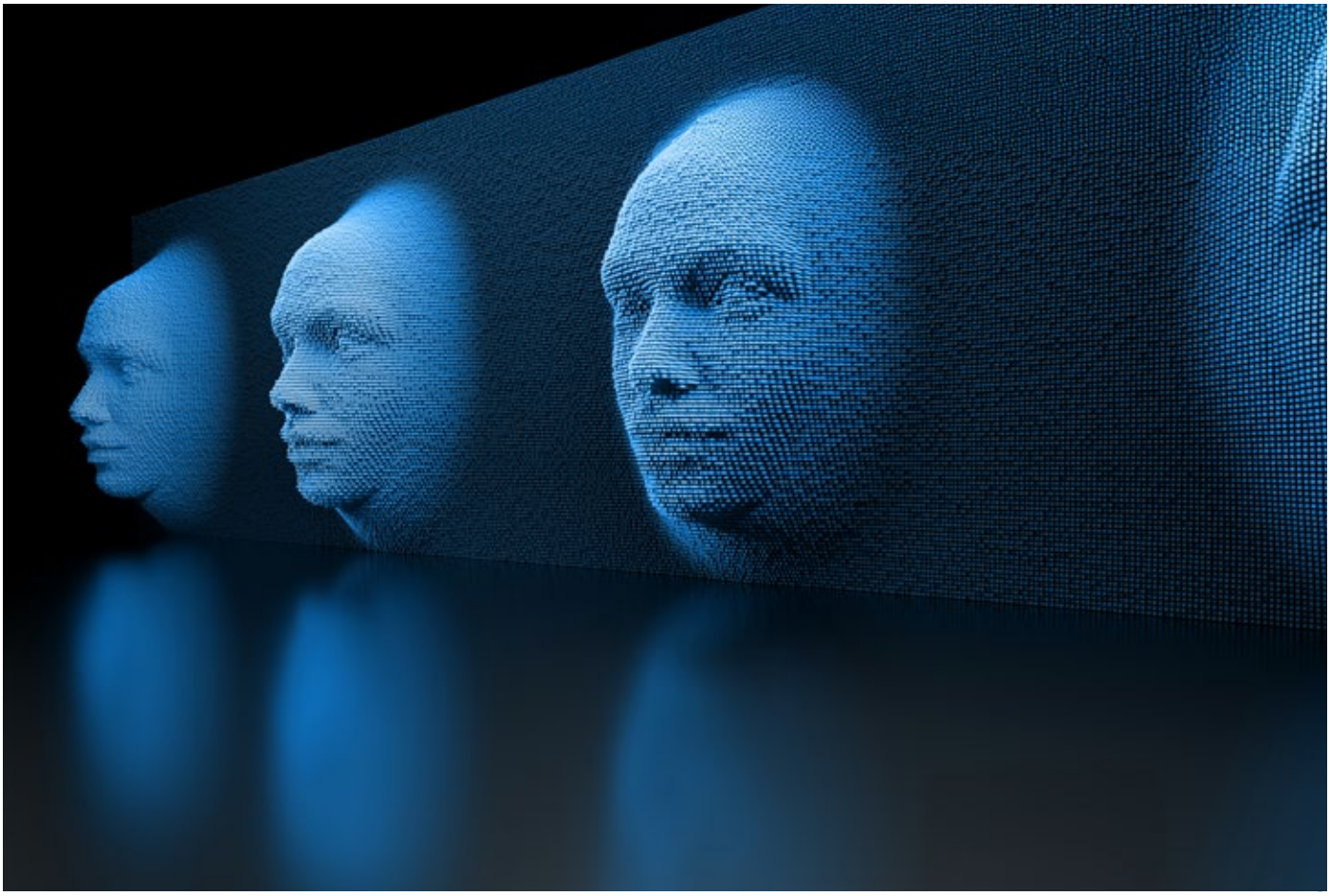
关键工具、技术与进一步阅读

工具和技术：

• 人权影响评估框架• DEI关注型数据标注工具• 人工智能伦理清单• 利益相关者共创平台• KPMG可信人工智能框架• KPMG可信人工智能风险与控制矩阵 (RCM)

拓展阅读：

• IEEE道德对齐设计 • 联合国B-Tech人权与人工智能项目 • 伯克曼克莱因中心——人工智能与公民自由



第十条原则

与人工智能和平共存

阿拉伯联合酋长国强调与人工智能和平共处，以确保这项技术被用来促进社区福祉和进步，同时不损害人类安全或基本权利。

用现实世界的术语理解原理

与人工智能和平共处要求确保人工智能系统补充人类生活和社会，在保持对技术的控制的同时提高生产力和福祉。这一原则着重于人工智能的伦理整合，确保人工智能系统不对人权、隐私或安全构成威胁。

现实世界中的例子：

- **工作场所自动化：** 通过再培训和创造就业岗位来保障员工福利，从而支持工人的人工智能系统。

- **监控中的AI** 监控中的人工智能正以符合人权和安全的方式使用。

将此原则嵌入人工智能治理

为确保和平共存，企业必须开发高效且符合社会价值观的AI技术——尊重隐私、人权，并促进社会凝聚力。伦理应用应包括变革管理、持续培训和关键绩效指标（KPIs）以监测影响。像KPMG的可信AI（Trusted AI）这样的框架可以通过将伦理考量、负责任的数据使用和适应性嵌入AI系统来强化这些优先事项——帮助组织建立信任并应对不断变化的法规，例如欧盟AI法案。

最佳实践与方法论

策略与设计

• **合规人工智能设计：** 从一开始就优先考虑人工智能系统对人体的影响，通过设计促进公平、公正和尊重隐私的人工智能解决方案。

• **利益相关者包容性：** 让伦理学家、法律专家和受影响的社区等多元化利益相关者参与人工智能设计过程，以确保与更广泛的社会目标保持一致。

数据赋能

- **偏差缓解：** 确保人工智能系统得到训练在多样、有代表性的数据集上防止可能造成伤害的偏见决策易受攻击的社区。

• **隐私保护：** 使用匿名化和加密技术来保护用户数据，确保人工智能系统不会侵犯个人隐私或泄露敏感信息。

模型开发

• **将可解释性构建到模型架构中：** 优先考虑那些天生易于解释的模型，以便利益相关者更容易理解、质疑和信任人工智能决策。

测试与评估

• **影响评估：** 在部署人工智能系统之前，进行社会和伦理影响评估，评估对人权、安全和社會和諧的潛在風險。

• **公平性审计：** 定期审查人工智能系统，确保其不会无意中歧视或侵犯基本人权。

部署和监控

• **持续人工监督：** 实施对人工智能系统的持续人工监督，特别是在执法、医疗保健和金融服务等高风险领域，以确保人工智能与人类价值观和社会规范和谐共处。

• **道德人工智能报告：** 定期发布关于人工智能伦理使用的报告，详细说明任何问题或关注点，以及为解决这些问题所采取的措施。

将原理扩展到智能代理系统

智能体应促进社会和谐，而非制造混乱。确保智能体尊重用户边界，并以情境恰当的方式进行操作。避免鼓励依赖或操纵的设计。支持多方利益相关者在公共空间评估智能体使用情况。将设计中的伦理原则融入其中，以使智能体与和平共存相协调。监控智能体对人类行为和社会规范的影响。

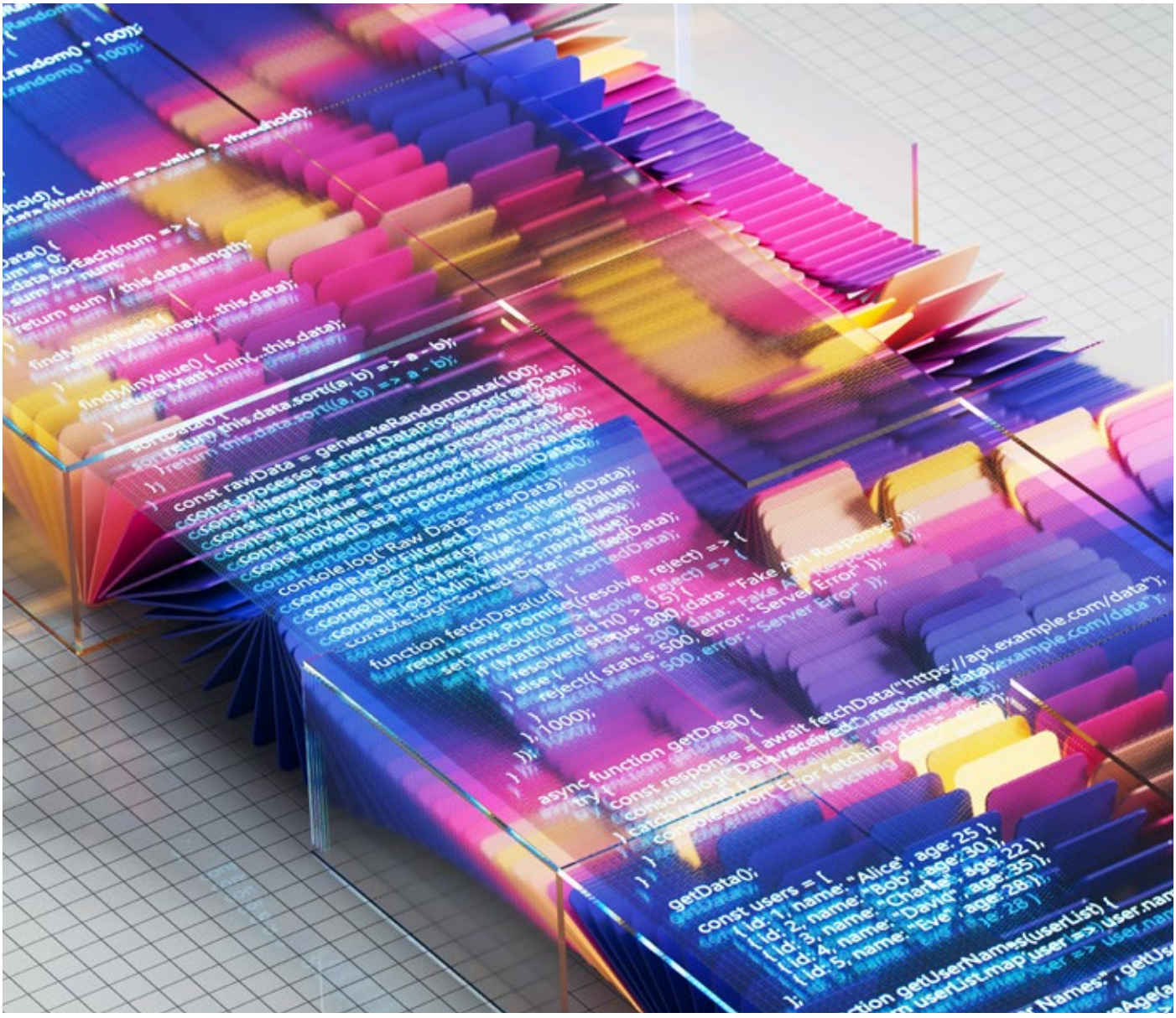
关键工具、技术与进一步阅读

工具和技术：

• 伦理影响评估模板 • 偏见审计工具 • 社区反馈循环（通过应用程序/网络界面） • 为社会福祉设计的AI工具包 • 普华永道可信人工智能框架 • 普华永道可信人工智能风险和控制矩阵（RCM）

拓展阅读：

• 联合国教科文组织人工智能伦理建议书 • 生命未来研究所——人工智能政策资源 • 阿联酋关于社会人工智能的部长简报



原则11

促进人工智能意识，共创包容未来

创建一个包容性的未来至关重要，该未来确保每个人都能从人工智能的进步中受益，保证社会各界都能公平地获得这项技术和其优势。

用现实世界的术语理解原理

提升人工智能意识可以确保每个人——商界领袖、学生、政策制定者和普通公众——都能参与和了解人工智能技术。如果没有广泛的知识和意识，人工智能就有可能成为一种专属工具，将其益处限制在小范围内。通过培养不同人群的人工智能知识和参与度，阿联酋确保人工智能技术为社会发展做出贡献，并包容所有人。

现实世界中的例子：

- **学校里的AI教育：** 将基础人工智能概念引入学校课程，以帮助来自各种背景的学生从小就理解和参与人工智能。

- **社区人工智能工作坊：** 举办免费公开课，提升人们对人工智能如何影响日常生活认知，目标群体为技术或数字素养有限的人群。

将此原则嵌入人工智能治理

要整合这一原则，需专注于构建人工智能素养项目，确保人工智能解决方案对所有社会群体具有包容性和可及性。鼓励促进人工智能参与意识和认知的举措，并确保您的AI系统设计用于服务所有用户，特别是那些可能在技术或知识方面有限制访问的用户。拥抱透明度，并为每个人提供学习和受益于人工智能发展的机会。

最佳实践与方法论

策略与设计

• **人工智能素养计划：** 为企业、学校和公众设计人工智能培训计划。

- **包容性拓展：** 服务未得到满足的需求社区，使人工智能更容易为更广泛的人群所使用观众。
- **以用户为中心的设计:** 设计人工智能系统那些考虑了不同的用户需求并提供基于专业知识的不同级别的交互。
- **政策制定：** 确保人工智能政策促进公平地获取和培养包容性人工智能生态系统。

模型开发

• **可访问性功能：** 集成使人工智能解决方案对残疾人和非专家更易用的功能。

测试与评估

• **包容性测试:** 用多样化的用户群体测试人工智能系统，以识别挑战并确保公平性。
• **持续参与** 使用用户反馈循环来完善人工智能系统，确保它们对所有社区部分保持相关性。

部署和监控

• **公众参与：** 维护面向公众的渠道（研讨会、网络研讨会），以持续与社区互动并提高AI意识。

• **影响追踪：** 追踪人工智能解决方案在不同人口群体中的采用情况，以确保没有人被落下。

• **公平性审计：** 开展审计，以评估人工智能系统是否在不同人口群体中促进公平的结果。

• **持续改进：** 定期根据社会反馈更新人工智能解决方案，以保持其包容性和可及性。

将原理扩展到智能代理系统

用知识赋予用户理解和管理自主式AI系统的能力。构建可访问、多语言和包容性的代理界面。在代理工具中提供培训、入职和帮助功能。推广普及数字素养计划，揭开代理功能之谜。代理本身可以通过可解释的交互来设计，以促进意识和包容性。

关键工具、技术与进一步阅读

工具和技术：

• 人工智能素养平台（例如 Elements of AI、AI4ALL）• 多语言自然语言处理模型（例如 mBERT、BLOOMZ）• 无障碍API（例如WCAG合规工具）• 适合低技术用户的负责任设计模板 • KPMG可信人工智能框架 • KPMG可信人工智能风险和控制矩阵（RCM）

• 阿联酋人工智能夏令营报告 • 世界经济论坛未来就业报告——人工智能技能提升 • 经济合作与发展组织数字包容工具包

原则12

履行条约和适用法律

阿联酋强调在人工智能的开发和使用中遵守国际条约和地方法律的重要性，确保人工智能技术在跨境部署时负责任且合法。

用现实世界的术语理解原理

该原则强调企业和组织确保人工智能技术合法使用的重要性，遵守当地法律和全球条约。这确保了人工智能应用不仅技术可靠，而且符合保护隐私、安全和基本权利的监管框架。人工智能的合法使用至关重要，公司必须避免以违反法律或道德标准的方式部署人工智能技术。

现实世界中的例子：

- **跨境数据传输：** 确保在跨境传输个人数据时，人工智能系统符合国际数据保护条约，例如欧盟的通用数据保护条例（GDPR）。

- **出口管制条例:** 在国防或关键基础设施等敏感行业中使用符合国际出口管制法的ai，以防止被滥用。

将此原则嵌入人工智能治理

为确保符合当地法律法规和国际条约，企业应将法律合规性嵌入人工智能的开发和部署流程中。这包括进行定期的法律审查，并确保人工智能技术符合国际条约、公约以及其他相关监管框架的要求，例如欧盟人工智能法案。例如，该法案也强调了遵守当地法律法规和国际条约的重要性，特别是在确保人工智能系统尊重基本权利和促进公共安全方面。这有助于建立人工智能系统与全球标准协调的清晰框架。通过坚持这一原则并遵循最佳实践，组织将更有能力预防法律问题，并符合全球不断发展的人工智能法规。

最佳实践与方法论

策略与设计

• **人工智能设计中的法律审查：** 在开发人工智能系统之前，应咨询法律专家以确保系统符合当地和国际条约以及适用法律。

• **全球合规战略:** 制定全球战略，确保符合地区法律（如数据保护法）和国际条约，以避免在新市场扩张时陷入法律困境。

数据赋能

• **跨境数据合规：** 在跨境处理个人数据时，确保遵守国际协议和数据保护法律（如GDPR），确保数据主体在各司法管辖区的权利得到尊重。

• **数据最小化:** 应用数据最小化等原则，避免收集可能违反隐私法律法规或条约的过多数据。

模型开发

• **人工智能模型的法律法规审计:** 对AI模型实施定期的法律审计，以确保符合相关法律法规，包括知识产权、隐私保护和出口管制。

• **人权影响评估：** 确保人工智能模型不侵犯基本人权，并遵守保护此类权利的国际人权条约和国家法律。

测试与评估

• **法律风险评估：** 定期评估人工智能系统的法律风险，例如侵犯隐私法、数据保护法规或个人权利，确保符合国际条约和地方立法。

• **法律合规指标：** 建立可衡量的指标来评估人工智能系统的法律合规性，以及它们对当地法律的遵守情况。

部署和监控

• **持续法律监管：** 监控部署后的人工智能系统，以确保其符合不断变化的当地法律和国际条约。这包括更新人工智能治理实践，以反映相关法律或法规的任何变化。

• **法律透明度报告：** 发布透明度报告，说明人工智能系统如何遵守地方和国际法律框架，确保问责制。

将原理扩展到智能代理系统

自主型人工智能必须遵守不断发展的法律法规，包括国际条约和地域性法规。实施符合法律合规需求的模块，跟踪监管要求的变化。使代理能够根据当地法律环境调整其行为。维护日志和审计追踪以保障法律合规性。与法律、合规和审计团队协作，确保全过程监管。代理绝不能超越其授权范围或违反数据、安全或人权法规。

关键工具、技术与进一步阅读

工具和技术：

• 合规管理系统 (CMS) • 法律风险扫描工具（例如 Relativity AI、基于 NLP 的法律条款标记器）• 跨境数据传输框架（例如 SCC 工具、OneTrust）• 出口管制合规监控器 • 普华永道可信人工智能框架 • 普华永道可信人工智能风险和控制矩阵 (RCM)

拓展阅读：

• 经合组织人工智能原则 • 阿联酋网络法与数据法规门户 • 欧盟人工智能法案（关于法律责任章节）

行动号召

伦理AI的基础

清晰的、可操作的AI原则是负责任AI的基础。正如KPMG所观察到的，这些原则建立信任、问责制和创新。阿联酋的AI宪章列出了12项关键原则，作为一项全面的治理框架，但只有通过实施它们（将其转化为流程和控制），组织才能使AI与社会价值观保持一致。

框架对齐

宪章的原则与KPMG的值得信任的人工智能框架紧密一致，为实施提供了实用指导。通过将宪章与值得信任的人工智能相结合，企业获得了逐步实施的、按设计进行伦理指导，这使得人工智能 **系统透明，可解释** 和 **合规。**

企业必须超越口号式的伦理声明，走向具体的治理。正如一位专家指出，“挑战在于从口号式的AI伦理原则……过渡到一个成为现实的空间”。在实践中，这意味着构建问责机制和监督机制（审计、测试、审查委员会），将宪章的意图嵌入到AI生命周期的每个阶段。

早期行动的益处

主动与宪章（及待批准法规）保持一致是值得的。早期采用者建立 **利益相关方信任** 并可信度 **减少AI风险** (bias, safety, reputational harm) and **激发负责任的创新** ，获得竞争优势。他们也为自己定位 **合规与韧性** ——以较小干扰迎接新兴法律（例如欧盟AI法案）。

可操作的建议

进行原则差距和风险清单

对照阿联酋宪章（以及欧盟人工智能法案草案）评估当前的人工智能项目。目录化系统并根据风险对其进行分类，以识别治理、隐私或公平方面的差距。

建立完善的AI监督

设立一个跨职能的人工智能伦理或治理委员会（包括法律、技术和业务领导者），以在所有人工智能计划中实施人类监督、问责制和合规性。

在整个人工智能生命周期中整合隐私保护、可解释性和偏差缓解。采用最佳实践（例如KPMG的可信人工智能框架），以便模型默认情况下是透明、公平和安全的。

准备合规

根据欧盟风险等级对人工智能系统进行清点和分类，并开始建立所需的文件/测试。采取早期行动（例如指定高风险用例和编写符合性检查清单）将简化对欧盟人工智能法案和类似法律的合规。

教育和参与利益相关者：

对团队进行符合道德的人工智能标准和法规新要求的培训，并公开向客户和合作伙伴传达您的人工智能政策。关于人工智能控制和用例的清晰、持续的对话能够建立信任，并推动包容性、负责任的应用。

通过在结构化治理框架内内化阿联酋宪章的原则，组织可以将伦理转化为行动——降低今天的风险，并为明天的AI机遇和需求奠定基础。

下一步

早期开始协调旅程的组织将更有能力适应变化、证明问责制，并以有效和道德的方式部署人工智能。虽然本文提出的方法提供了一个坚实的基础，但人工智能治理必须根据组织的独特环境、战略重点和风险状况进行定制。从原则到实践需要领导层、监督职能和技术执行之间的持续协调。KPMG 通过其全球认可的可信赖人工智能框架支持这种转型，辅以人工智能风险和控制矩阵（RCM）以及治理运营模型等工具。这些服务共同帮助组织评估人工智能成熟度、设计定制化的治理结构、实施强大的风险控制、培养负责任的创新文化，并最终与本地和全球的人工智能原则和监管要求保持一致。对于准备行动的组织，KPMG 带来跨学科专业知识，以嵌入面向未来的人工智能治理。

KPMG 中东有限责任公司

KPMG中东有限责任合伙公司是KPMG全球组织的一部分，该组织由在143个国家和地区运营并隶属于KPMG国际有限责任公司的独立成员公司组成。我们为沙特阿拉伯、阿拉伯联合酋长国、约旦、黎巴嫩、阿曼和伊拉克的公共和私营部门客户提供审计、税务和咨询服务，并通过独立的法律实体进行签约。我们在该地区拥有悠久的传统，已经建立了超过50年。KPMG中东有限责任合伙公司与其全球成员网络联系紧密，并将本地知识与国际专业知识相结合。

毕马威为企业的多样化需求、政府、公共部门机构、非营利组织以及资本市场提供服务。

我们对质量和卓越服务的承诺是我们一切行动的基础。我们努力为利益相关者提供最高标准的服务，通过我们的行为和举止建立信任，无论是在职业上还是个人上。

我们的价值观指导我们的日常行为，影响着我们的行动、决策以及我们如何与彼此、客户以及所有利益相关者合作。正直：我们做正确的事。卓越：我们永不停止学习和改进。勇气：我们大胆思考和行动。团结：我们互相尊重，并从差异中汲取力量。为更好：我们做重要的事。

我们的宗旨是激发信心，赋权变革。通过激发我们的人民、客户和社会的信心，我们帮助赋能变革，以解决最艰巨的挑战并引领前进的方向。

毕马威的“我们的影响计划”指导我们对客户、人群和社区的承诺，涵盖四个类别：地球、人群、繁荣和治理。这四个优先领域帮助我们界定和管理我们的环境、社会、经济和治理影响，以创造一个更可持续的未来。我们致力于实现有意义的增长。我们汇聚毕马威的精华，帮助客户实现他们的使命，并为联合国可持续发展目标做出贡献，使所有我们的社区都能蓬勃发展。

我们致力于以目标为导向实现增长，帮助客户实现他们的目标，并推动可持续进步，以确保我们所有的社区都能繁荣发展。秉持我们的价值观，并致力于我们的目标，我们的员工是我们最大的力量。一起，我们正在建设一个以价值观为导向的未来组织。为了更好。

免责声明：此处所述的部分或全部服务可能对KPMG审计客户及其附属公司或关联实体不适用。

埃米利奥·佩拉
中东地区副首席执行官，波斯湾地区首席执行官KPMG中东 emiliopera@kpmg.com

罗伯特·普塔辛斯基
数字经济与创新及托管服务主管 KPMG 中东地区
rptaszynski@kpmg.com

马廷·朱兹达尼
伙伴，数据，分析和人工智能KPMG(lowerGulf) mjouzdani1@kpmg.com

乔·德瓦西
总监，战略联盟 KPMG 下加利福尼亚 jdevassy@kpmg.com

WORLD GOVERNMENTS SUMMIT

JOIN THE CONVERSATION

@WorldGovSummit

www.worldgovernmentssummit.org