KPMG

人工智能控制

人工智能有助于提升透明性，
可解释性和可信性

2019年6月

kpmg.com

撰文

Martin Sokalski
主管，
咨询及新兴科技风险服务
毕马威美国

Martin Sokalski是毕马威新兴科技风险业务的全球主管。他协助全球企业构思、创新和负责任地应用新技术，以助它们拥抱“可能的艺术”（“art of the possible”），这全赖人工智能等新兴科技的兴起。多年来，他协助不同行业的多个组织评估、设计及实施新的数字化运营及治理模型，以助它们在完成关键的治理、信任及价值要务的同时，实现理想的业务结果。Martin定期在不同会议上发表演讲，并在有关人工智能、数字化转型及新兴科技的领先思维刊物上发表文章。Martin认为，人工智能的大规模应用当前正受到信任与透明度缺乏、可解释性以及无意偏差等方面的阻碍，并致力与行业领袖合作以有效应对这些难题。

Sander Klous 教授、博士
主管合伙人，
数据分析
毕马威荷兰

Sander Klous是毕马威荷兰数据分析主管合伙人和阿姆斯特丹大学商业及社会大数据生态体系教授。他拥有高能物理学博士学位，曾在位于日内瓦的全球最大物理研究组织—欧洲核子研究中心的多个项目上工作超过10年。他的畅销书《我们是大数据》（“*We are Big Data*”）是2015年年度管理书籍奖第二位。他的新书《在智能社会中建立信任》（“*Building Trust in a Smart Society*”）是荷兰最畅销的管理书籍。Sander在大规模分布式计算、实时系统及数据处理技术方面拥有丰富经验。他目前的工作重点是以客户和社会普遍认为有价值的方式大范围应用可靠分析、合乎道德算法和可信赖分析。

Swami Chandrasekaran
执行总监，
毕马威创新与企业方案
毕马威美国

Swami Chandrasekaran是毕马威创新与企业方案小组主管，并领导智能自动化（包括人工智能和新兴技术供应）的建构、技术和创建。他曾在税务及审计、工业自动化、航天安全、联络中心、保险理赔、外勤服务、多媒体增值、社会服务、数字营销及企业并购等领域，克服各种各样的挑战，领导多个复杂的人工智能产品和解决方案的孵化、设计和创建。Swami在业务流程自动化、系统与数据整合方面亦有丰富经验。Swami是一名IBM杰出荣誉工程师。他近期出版了《学习利用Watson人工智能建立应用》（“Learning to Build Apps Using Watson AI”），并已申报20项专利，其中17项已发表，多数涉及人工智能领域。他曾被委任为IBM杰出发明人。

目录

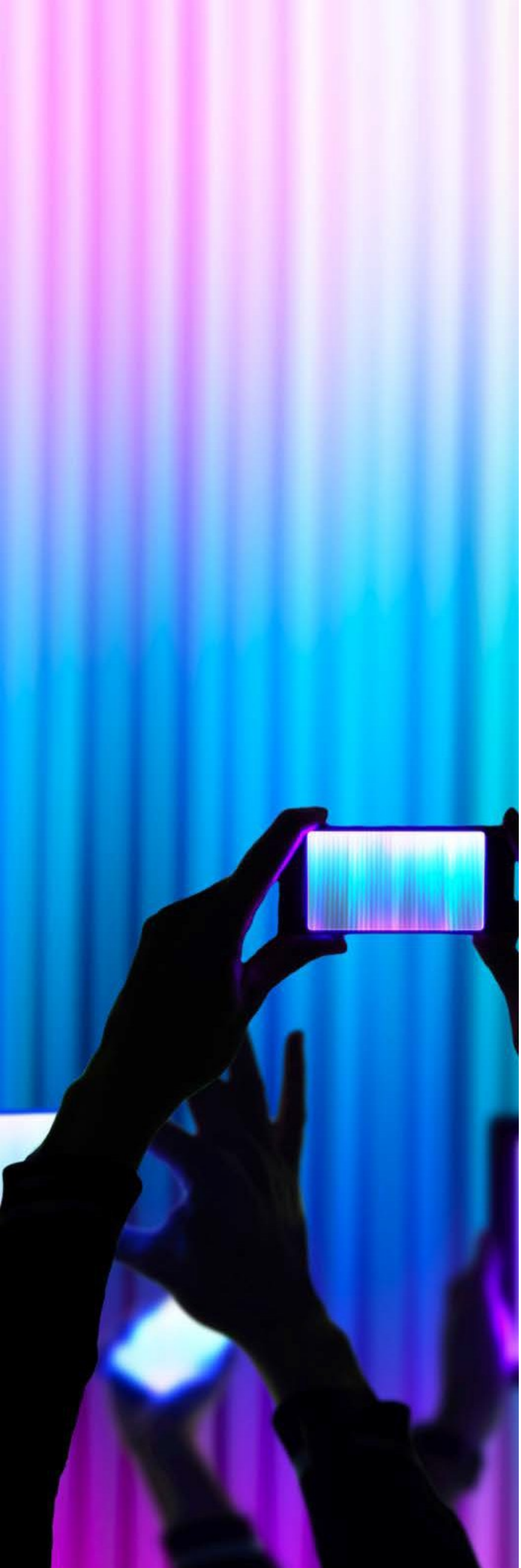
引言

人类历史上所有改变世界的技术均涉及一项共同的元素：控制。

蒸汽与光，以及众多其它发明与技术的兴起是由于我们有能力将自然力量转化为可转换的能量。

若我们缺乏对飞行的充分掌握，航空业将不会存在。

人工智能也有着与上述技术一样改变世界的潜力。



但我们尚未了解人工智能的能力界限。

与其它转换性技术相似，我们唯有了解和控制人工智能的建构和行为，才能完全释放其能力和前景。因此，企业需要确立一套整体的人工智能管理政策，重点是负责任地解放这些技术的能力。

人工智能向我们揭示了隐藏在复杂性后的世界。能够自我学习和持续进化的算法带来的洞见正改变着我们的业务和生活。科学家们可预见一个未来，其中人类面对的某些最深层次的奥秘和难题将会被解决。目前，我们已看到由此而来的裨益—从协助发现亚原子粒子和捕获黑洞的首张照片的算法，到企业层面上，复杂的数据分析在人工智能驱动下，正作出会影响盈亏底线及品牌，以及消费者的健康与安全的关键决策。

人工智能的实际应用

设想一个大型金融机构的消费贷款业务线负责人，牵涉进一起涉及偏见和歧视的事件，并登上新闻头条。在董事会会议上，多名董事会成员要求负责人和首席数字官解释他们拒绝向特定年龄组别或种族的贷款申请人作出贷款这一决定背后的理由。此事件中，起关键作用的是生成结果和提升贷款操作员决策的人工智能算法。这两个负责人面对的问题是：没有人能解释为什么算法会作出这样的决定。

类似情形正在企业及公共部门的不同领域上演：招聘、运输、营销、医疗保健、大学录取、住房和智能城市管理等。开发或应用先进且能不断学习的技术的组织正在挖掘一种远超人类思维能力的洞见和决策力。这是一个巨大的机遇。

但当生成错误或有偏差的结果时，这些算法将是毁灭性的。这是任何希望明确算法应用的企业主管的固有疑虑，这种疑虑在黑匣子效应下进一步放大。因此，虽然人们对人工智能兴致盎然，但在不确定如何作出决策以及决策是否公平且准确的情况下，企业在将决策交托机器上存在犹豫。这即为信任鸿沟。

提升信任

只有当算法可以简明语言向任何人解释清楚并因此被理解时，人工智能的巨大价值才会完全释放。信任鸿沟存在是由于人工智能缺乏透明度；相反，围绕着该技术的，是人们对未知的固有恐惧。提升信任还包括了解人工智能模型的由来，并保护模型（以及构成这些模型的数据）免受各种对抗攻击或未经授权使用。人工智能作出的关键业务决策影响着企业品牌，以及消费者对品牌的信任，可对消费者及市民的福祉或安全产生巨大影响。没有人想说：“因为机器是这样说的。”没有人希望人工智出错。

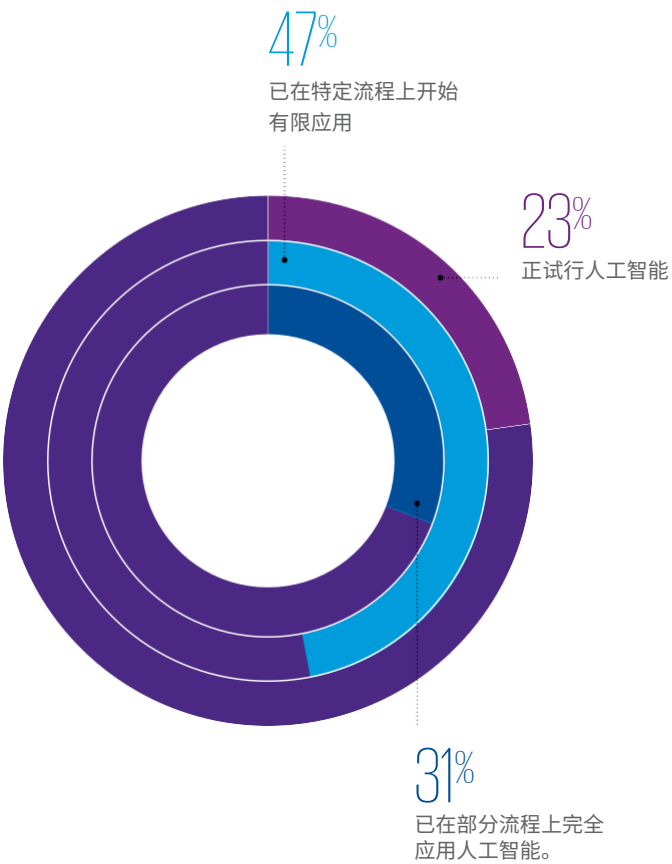
消除信任鸿沟

如今，公平性及可解释的人工智能不仅是企业最高管理层及董事会的重大难题，还是一个必要事项。譬如，毕马威《2019年首席执行官展望》发现[1]，66%的受访高管忽视了计算机驱动数据分析提供的洞见，原因是这些洞见与他们的经验或直觉相反。

对多数机构而言，人工智能仍然仅存在于实验室，仅是功能层面上配置，尚未成为企业决策的必要组成部分——虽然此情况正在快速改变。

[1]毕马威《2019年美国首席执行官展望：敏捷还是被淘汰：重新定义适应力》（“Agile or Irrelevant: Redefining Resilience”），2019年6月

根据毕马威《2019年首席执行官展望》，机构正处于不同的人工智能配置阶段。



解决方案是什么？

为了让人工智能服务于共同利益、让机构主管对结果承担责任，机构有必要建立一个架构（并配置相关方法和工具）以促进人工智能的有效实施和规模扩张。

本报告的目标读者是人工智能和机器学习算法领域的机构主管。

就业务和合规工作而言，理解人工智能并对其树立信心已成为必要。

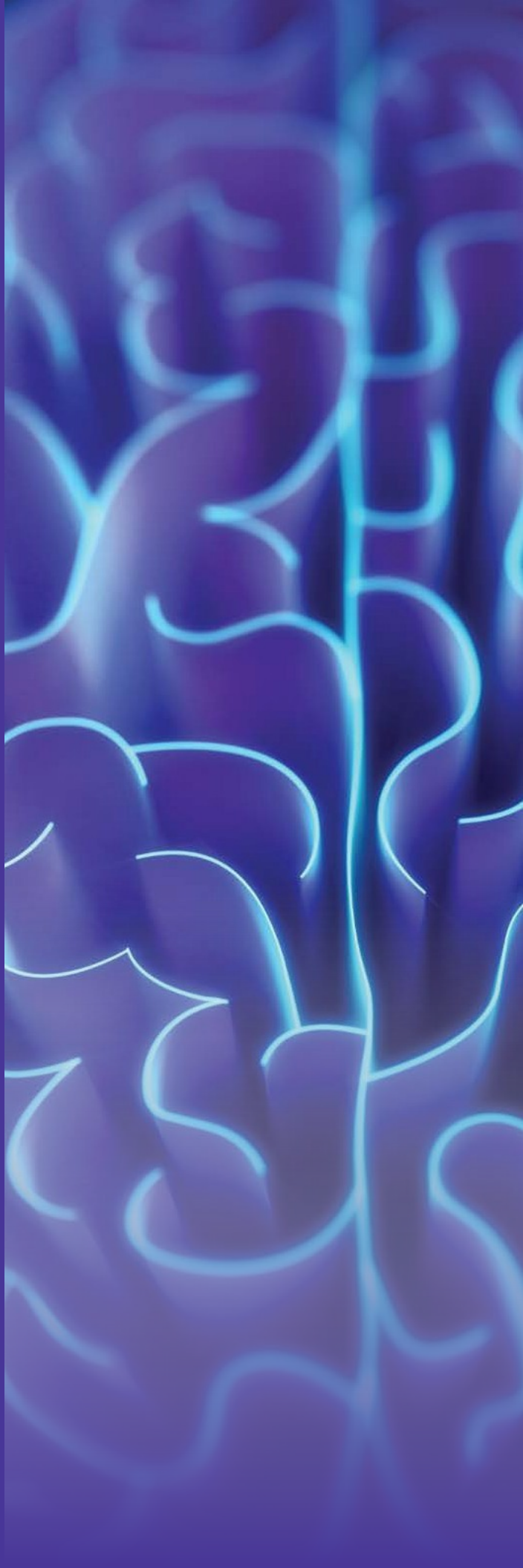
本报告阐述了解决方案的迫切性，并提供了相关方法和工具以协助机构主管对人工智能程序实施治理。

“只要机构能提升信任和透明度，人工智能真正的 “可能的艺术”（art of the possible）便会释放。这可通过纳入基础人工智能程序的必要元素（如完整性、可解释性、公平性和适应能力）来实现。”

Martin Sokalski
主管，
咨询及新兴科技风险服务，
毕马威美国

重要发展

根据大型企业中负责推动人工智能战略的高管的访谈，我们获取了一致的信息：与其它人工智能配置要项相比，多数企业刚开始投资建立人工智能控制架构[2]。



[2]毕马威《2019年企业人工智能应用研究》。

建立对人工智能的信任是机构主管的首要目标。

45%的受访高管表示，他们难以或非常难以信任人工智能系统[3]。

有关数据与人工智能的新规预示着自我监管的结束 以及全新监管模式的出现[4]。

多数机构主管尚不了解人工智能治理方案应是怎样。

70%的受访者表示他们不知道如何对算法实施治理[5]。

企业难以确定谁应对人工智能的程序及结果负责。

在我们的访谈中，我们听到多数企业仍难以确定谁应拥有人工智能配置的权限。部分企业已通过建立人工智能委员会或卓越中心来确立集中权限；其它企业则将相关责任分配至首席科技官或首席信息官等企业高管。

一个包含技术赋能方案的架构有助企业应对人工智能的固有风险和道德问题。

目标是通过四个信任锚，即完整性、可解释性、公平性和适应能力，协助企业用户确立对人工智能程序的控制。

[3]Forrester Research，2018年第二季度全球人工智能在线调查

[4]“人工智能及互联网政策提案预示着自我监管的结束”（“AI, Internet Policy Proposals Signal Shift Away From Self-Regulation”）.《华尔街日报》专业版：人工智能，2019年4月9日

[5]信息来源：毕马威，《为何审计必须应用人工智能》（“Why AI Must Be Included in Audits”），2018年

对了解的需求：信任锚

错误理解人工智能所带来的成本超出财务数字之外，从收入损失、合规失效罚款至声誉、品牌和道德问题。

我们刚刚假设了一名首席数字官和业务线负责人试图向董事会解释单个模型的结果。问责层级从企业高管和银行整个信用卡部门的业务线负责人（对与该业务相关的所有事情负责，包括人工智能模型）一直延伸到客户层，其中一名贷款操作员可能需承担责任。在很多方面，贷款操作员便是品牌的代表。

规模化的关键业务决策对企业取得成功有着决定性影响，譬如：

该部门是否应批准某客户的信用卡申请？

需为每个客户执行的决定包括：年利率、消费限额和其它众多因素。一般而言，机器学习模型正为数百万计的客户执行这些决定。由于规模巨大，业务实际上掌握在少数智能数据科学家手上—以及他们建造及训练的机器—它们利用的是从历史贷款数据中生成的“真实值”（“ground truth”）。

自主算法的今昔

如今的多数算法相对简单和确定性较高：它们从预设状态集和定量规则中生成相同的产出。有关这些算法的有效性及完整性的评估方法一般已确立并被采纳。实际上，根据我们的估算，维持算法准确度及有效性所需的领先实践中超过80%已为人所知。

比如，制造业的专家系统；保险业中运用决定性规则或决策表的精算学；以及金融服务业的机器人流程自动化（RPA）。

用户较易确定这些系统得出的结论是否可接受，且合理、规模可控的监管较容易实施。

这些规则可变得异常复杂，特别是当数据中的属性（又称“特性”或“变量”）数量或记录数量增加时。

机器学习与深度学习——以及其它类型的人工智能——是各不相同的产物。它们被训练成可从数据中学习（一般被称为“真实值”），而不是被明确地编程。这意味着它们能够“理解-学习-揭示”数据中的细微差别和模式，能够处理海量属性集，其运作方式也更为复杂。

譬如，以百万计的贷款申请数据集来训练某个预测模型，该模型运用100个属性；从一百万张核磁共振图像中侦测肿瘤；以及电邮分类。这些模型一经训练和评估，便可从新的或未见过的数据中作出预测。它们会以一定的可信度作出基于概率的结论。

虽然所有这些方面均很好，但用户可能不清楚模型正如何运作：它们正在学习什么（特别是应用神经网络等不透明的深度学习技术时），它们将如何反应或它们是否在持续进化的过程中逐渐养成某种偏见。因此，用户必须了解训练数据中的哪些属性会影响模型的预测结果。

算法风险：对机器的信任

让我们更深入地分析该大型金融机构贷款部门的首席数字官和业务线负责人的潜在问题。

如果一个错误隐藏在算法（或输入到算法或训练算法的数据）内，可能会影响机器所作决定的完整性和公平性。这可能包括对立数据或伪装为“真实值”的数据。企业高管有责任维护企业的品牌声誉，即使是当人工智能的决策越来越让人难以理解，或不符合企业政策、企业价值、指引及公众期望时。贷款部门运用的各种算法均会出现类似问题，因此算法数量越大，问题也越多。因此，信任逐渐减弱，甚至消失无踪。

各种技术，包括那些基于重整化群理论的技术已陆续面世[6]。随着各个人工智能领域——电脑视觉、语音识别和自然语言处理——的模型变得愈加复杂和自主，它们也需承担更高水平的风险和责任。当训练停止一段较长时间后，这些模型便可能出错：运行时间偏差的蔓延、时间推移导致的预测不准确和对抗性攻击等问题将妨碍模型的输出数据。譬如，在某个智能城市里，被入侵的核磁共振扫描或交通灯正被随意操控。

持续学习算法还带来了一系列新的网络安全考量。提前应用者仍在努力应对上述问题带来的业务风险。

其中一项风险是，通过损害模型或篡改训练数据集来攻击算法基础的对抗攻击。这些攻击可能会损害隐私、用户体验、知识产权和其它关键业务因素。设想一下，医疗设备或工业控制系统若遭受对抗攻击，会对人们生命或环境带来何种影响。就零售或金融服务业而言，篡改数据将导致系统提供不合理建议，从而损害消费者体验。此类攻击将可能最终侵蚀掉算法原应创造的竞争优势。

在应用复杂、不断自我学习的算法时，机构除了要了解数据或属性以及它们各自的权重外，还需了解更多信息，以充分意识到人工智能出错或发生误判可能带来的影响；它们需要了解模型开发的背景及原定用途、谁训练模型、数据源头及其任何变动，以及模型以往及当前正作何用途及受到何种保护。此外，它们还需清楚就算法的完整性、可解释性、公平性及适应能力而言，它们应提出的问题以及应关注的关键指标。

人工智能应用者的首要挑战是数据质量。在毕马威全球人工智能调查中，某政府机构的首席技术官特别提到，如果他们不能信任数据，便不能应用人工智能[7]。

[6]《不同的重整化群与深度学习的精确配对》（“An Exact Mapping Between the Variational Renormalization Group and Deep Learning”），Pankaj Mehta, David J. Schwab。2014年

[7]Forrester Research,《2018年第二季度全球人工智能在线调研》（“Q2 2018 Global AI Online Survey”）

信任之锚

当您细分为建立算法和模型—以及您的品牌—的可信度所需的所有行动和技能时，毕马威认为将出现以下四个维度。

算法完整性

我们可以对房屋的“框架”进行检验，来比喻算法的结构性缺陷及完整性评估。机构主管需要了解的是：训练数据的源头和由来、模型训练控制、建造、模型评估指标、完整的维护过程，并确认未发生改变算法初始目标或意图的变更。此外，关键在于持续监控模型绩效指标，包括概念趋势变化检测。

可解释性

了解模型作出某个预测的原因——并能解释相关原因——对树立系统可信度十分重要，尤其是当用户须根据这些概率性结果作出行动时。

这是人工智能中的主观能力。能够解释模型为何及如何生成某个输出、洞见或决策的能力取决于已确立的成功定义以及算法的整体治理，具体方法包括从清晰、充分及合理的真实值收集到结果的持续评估。目前已有多个方案存在，包括本地可诠释模型怀疑性解释（“LIME”，一种针对本地或分离的决策方面的解释技巧8）和美国国防部先进研究项目局（“DARPA”）的可解释人工智能项目（旨在创建一套机器学习技巧，以提升模型的可解释性，并附带解释界面）。

公平性：道德与问责

如果人工智能和算法是不公平的，它们便不会被信任。要确保公平，它们的设计及建构应尽可能摆脱偏见，且它们应在演变过程中保持公平。

用于训练算法的属性需具相关性、适用于目标，并须被允许使用。但在某些情况下，个人信息对模型十分重要。譬如在医疗保健业，性别或种族可能是某些研究或治疗的重要部分。机构应建立谨慎的监管和治理，以确保模型的训练不会用到代用参数。譬如，邮政编码可被用作民族或收入的代用参数，并在无意中生成带偏见的结果及下游风险——其中一个是监管违规风险。机构必须使用适当的技巧去理解数据中固有的偏见，并运用再平衡、重新分配权重或去偏误化等方法来减少这些偏见。

持续监控及治理工具相当重要，有助于确保持续以使用及反馈数据训练的模型不会导致偏误在运行时间潜入。

适应能力

我们在此探讨所配置或应用的模型或算法的稳健性及适应能力。所应用的模型一般为应用程序接口或嵌入应用内 ， 需具便携性并能在不同的复杂生态系统间运行。

适应能力强的人工智能应能涵盖安全应用的所有方面，并通过谨慎设计的架构和异常侦测（运用生成式对抗网络等人工智能概念，将各个算法对立起来以生成更好更精确的结果）来全面应对风险。目标是确保各个组成部分得到充分的保护和监控。为何要这样？当算法不能对错误或异常数据进行纠正或补偿时，外部境况可能引致错误。保护可用于模型持续训练的使用及反馈数据也同样重要，常规措施包括持续监控模型端点和控制外界对模型的访问。

一个有待解决的核心问题：谁应对人工智能的结果负责？

问责性是一个重要的治理事项，必须纳入所有人工智能项目，下至各个模型。在毕马威2019年企业人工智能应用研究中，我们发现不同机构在分配权限和职责方面存在显著差异。某些机构建立了集中化职权，如人工智能委员会；其它则向首席技术官或首席信息官授权。但少有机构实施了稳健的问责措施，这是一个会削弱内部信任和外部利益相关者信任的领导力差距。此缺失环节的一大原因是，多数机构缺乏必要的工具和专业知识，以充分了解自身算法并提升其透明度。

[8]Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin,《我为什么要相信你？分类器预测解释》（“Why Should I Trust You? Explaining the Predictions of Any Classifier.”）2016年

人工智能治理和道德规范

通过应对复杂算法可能会出错的风险，治理和道德成为有效应用人工智能的途径。

看看航空业治理中的规则和条例，和各个航空公司主导操作的内部最佳实践（上至高管层下至座舱工作人员）。看看所有参与者寄予的信任—工作人员、乘客以及通过飞机运输高价值资产的企业。这就是各个行业须通过人工智能实现的目标。

机构对人工智能的有效治理、负责任应用和规模扩张的需求临界点已到来。多数机构已在制订内部政策和治理职能以监管任何与人工智能有关的事项，从而提升企业内部及外部利益相关者团体（包括消费者）的信任和透明度。

在英国及欧盟，随着《通用数据保护条例》的发展，监管建立的趋势已愈加强烈。就此而言，当前正是大好时机，因人工智能的种子已落地生根并开始茁壮成长。目前虽然仍未扩张至预期规模，但这些技术必定会在企业内部和不同行业板块间扩张，并承担越来越多的自主权和责任。现在是时候建立一个围绕信任为基础的治理和道德框架了。人工智能控制有助确保人工智能能力的负责任扩张。

治理与道德规范

人工智能信任基础的评估和保障可基于全新的、旨在保持机构对人工智能及机器学习算法的控制的领先实务和方法。一个有效的治理战略可通过建立可持续测量人工智能的机制和工具，为信任和透明度打下基础。机构主管将能够在掌握充分信息的情况下作出决策，并在机构中建立更稳固的负责任文化，自觉反映机构的道德规范。

治理

人们在深究人工智能的工作方式时会发现很多问题，其中多数与人有关。某些用例为何及如何被选为人工智能的候选？团队为何选择这些特性（而排除另一些特性）？我们应如何计量和展示成功或解释失败？算法为何这样做？谁应对结果负责？“因为这是算法说的算”此说法将不能说服机构主管和社会大众，因这些系统已变得越来越强大和无处不在。

机构需要：洞悉大局，从一开始设定正确的基调。若您未为人工智能建立治理或运营模型，您将难以实现理想的业务结果或信赖自身人工智能的完整性、可解释性、公平性或适应能力。就信任和透明度而言机构也应建立人工智能的治理体系。这意味着机构将通过全新视角，以贯穿整个过程的员工、业务流程和技术出发，审视企业架构和治理：从模型的初始阶段至战略制定、交付、监控、训练及能力建设以及持续测量。

道德规范

无论是对于企业和大众面对的问题与困境，还是就控制人工智能所需的步骤及安全保障而言，道德都是人工智能领域的一大话题。道德与信任密切相关。两者均为人工智能向造福社会的方向发展提供动力。解决方案及相关法规已在实施。《通用数据保护条例》便是一个显著例子；其它法规也陆续出台，如欧盟委员会近期发布的可信任人工智能道德指引[9]。

道德的另一个方面是个人自主。我们乐意将哪些决定交给机器？哪些决定仍应由人来作出？这是科学界和政府讨论的焦点。值得一提的是，欧洲理事会部长委员会近期宣布，人工智能和机器学习技术“不能用于不当地影响或操纵个人的思想及行为”[10]。

机构需要：在人工智能项目的初始阶段建立道德保障体系，这要求机构了解及持续监控人工智能生命周期，即从战略制定到执行直到以后的演进阶段。

[9]《可信任的人工智能道德指引》("Ethics Guidelines for Trustworthy AI”）。欧盟委员会人工智能高级别专家组，2019年4月。
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines- trustworthy-ai

[10]欧洲理事会部长委员会，《部长委员会就算法流程对人的操纵能力所作的声明》（“Declaration by the Committee of Ministers on the Manipulative Capabilities of Algorithmic Processes”）。2019年2月
https://search.coe.int/cm/pages/result details.aspx?objectid=090000168092dd4b

受控的人工智能
阿姆斯特丹市

毕马威正与阿姆斯特丹市开展合作，以评估一项数字市政服务。通过此服务，市民能就街上垃圾等问题提交在线请求。机器学习算法即可识别问题类型、优先等级并确定应作出响应的市政服务。

阿姆斯特丹市官员将毕马威的人工智能应用于其控制架构，以有效、持续地评估不断进化的人工智能应用，防止它们无意中应用了可能导致错误或偏误决策的学习模式。

这些城市取得的成功越来越取决于它们对待数据的智能及道德程度。

目标结果

提升公众对安全及维护得当的城市的信心；协助城市保护市民的数字权利。

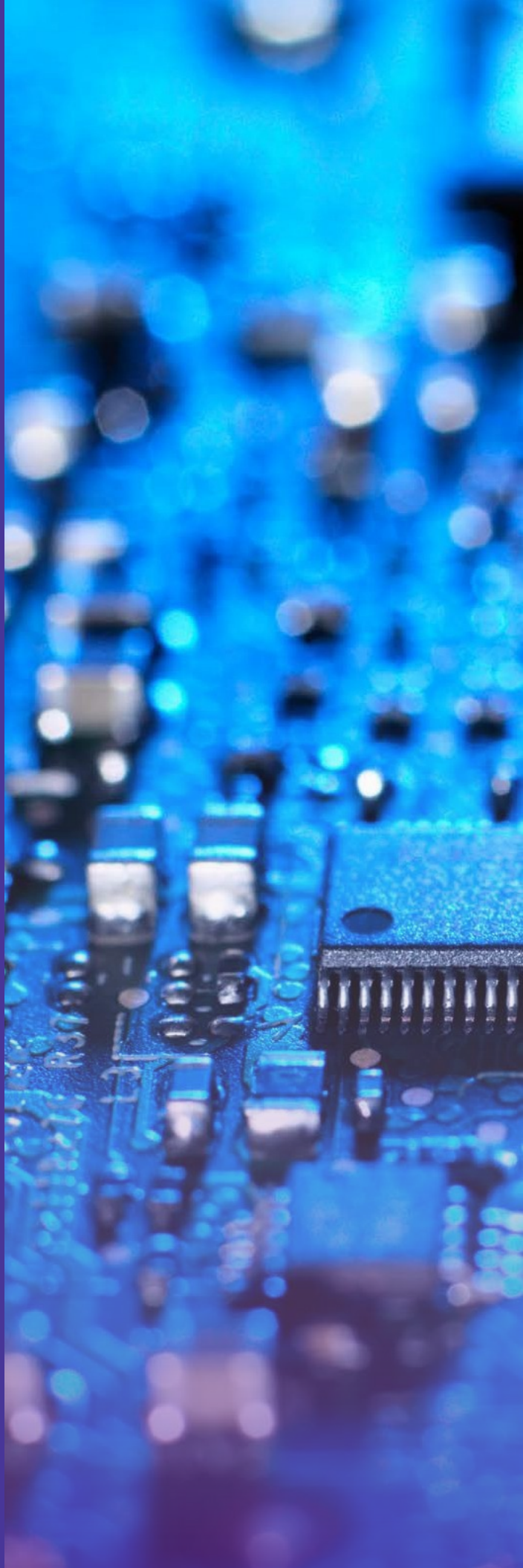
“阿姆斯特丹市致力于保护市民的数字权利，我们有责任保持我们用以支持市政服务及项目的机器学习算法的包容性及透明度。因此，我们力图与我们的合作伙伴毕马威开发一个模型，以助我们制定适用的筛选方法以验证和审批这些算法。”

Ger Baron
首席科技官，
阿姆斯特丹市

人工智能控制

人工智能治理关键：一个有助提升透明度的架构

实施适用准则以助企业透明、有效地治理算法，并建立对业务决策质量的信任。



多数机构主管不知道如何消除信任鸿沟，因他们不了解人工智能治理，不能洞悉其全局运作。

机构主管需要的是一个基层视角，以揭示关键绩效指标和关键风险指标。训练数据中的偏差是一大问题，整体的模型风险也是另一个问题。此外，还有合规与安全，以及其它众多因素。在本报告的引言中，我们回到令人不安的董事会场景：系统为何会生成错误的结果？在某些情况下，可能是由于系统代码或数据中的底层错误。机构只有实施严密措施才能预防或侦测到这些错误，从而消除疑虑和信任鸿沟。

人工智能控制

一个有效的框架体系有助于机构树立对人工智能技术的信心。此类方案应深入企业及个体模型层面，协助确保主要信任要素已在整个过程中整合和受控。方案应包含相关方法、控制及工具以保护整个生命周期中（从战略制定到演变）的信任基础，从而持续评估并保持对复杂、不断演进的算法的控制。此外，方案还应为机构——不同管理及监督职能的利益相关者——提供明确指引，以使它们能清晰自觉地管理人工智能的点对点生命周期。成功范例及关键考量如下：

战略

人工智能治理始于项目之始。为企业人工智能或特定模型制定合适的战略始于一个明确的愿景、抱负和目标结果。我们特此提及道德规范和问责概念。

设计

协助确保算法的原定用途已被明确定义，且模型的设计是为了实现该原定用途，具体方法是通过工程特点、数据差异和真实值。设计需符合原则（价值及道德）、安全、质量准则及指引和合规要求。

模型与训练

一旦达到设计标准，模型建造及训练便可启动。在此阶段，为保持模型完整性、公平性、可解释性及适应能力，机构需考虑偏差侦测、超参数、特征源头以及其它变量。模型特征及数据需遵循机构贯彻的原则、政策、业务要求和规定。

评价

这是人工智能生命周期中的关键一步；机构应能验证人工智能模型及它们生成的结果符合有关完整性、公平性、可解释性及适应能力的要求。机构需知道应提出哪些问题，应关注哪些关键绩效及风险指标，并具备必要的执行能力。人工智能评价及监控职能的效力及完整性将直接提升机构（或外部监管者）对企业人工智能的信心。

配置与发展

人工智能与机器学习模型并非静态模型。即使已投产，这些模型还可通过与新数据集或其它模型的互动而不断进化。因此，此阶段应考虑的主要因素包括运行时间监控与控制汇报、合规、关键绩效及风险指标以及模型准确度、完整性、公平性、可解释性及适应能力指标。除此之外，对这些指标作出响应的能力（包括动态模型校准）也是企业必须具备的能力之一。

受控的人工智能：算法治理架构

一个包含各种方法和工具的稳健架构的透明度有助提升人们对人工智能的信任，并能创造一个鼓励创新和适应力的环境。

从事人工智能建造及配置的机构正在挖掘一种远超人类思维能力的洞见和决策力。对企业和社会大众而言，这是巨大的机遇。但当算法生成错误或带偏差的结果时，则可能是毁灭性的。因此，在不了解机器如何作出决策及决策是否公平和准确时，机构主管不愿意将决策交托于机器。

只有当算法结果可以简明、直白的语言解释并被理解时，人工智能的力量和潜力才会完全释放。不重视人工智能治理及算法控制的企业将很可能损害自身的整体人工智能战略，使其业务计划（甚至品牌）蒙受风险。

“借鉴领先实践有助于降低人工智能在可解释性及偏差方面的固有风险。”

Martin Sokalski
主管，
咨询及新兴科技风险服务
毕马威美国[11]

[11]《华尔街日报》专业版，《人工智能及互联网政策提案预示着自我监管的远去》（“AI, Internet Policy Proposals Signal Shift Away From Self-Regulation”），2019年4月

人工智能治理

毕马威在控制架构中开发人工智能，通过经验证的人工智能治理体系以及贯穿整个人工智能生命周期（从战略制定到演进）的方法及工具，协助机构提升信心和透明度。此架构是为应对上文指出的固有风险而设计，并就建立人工智能治理、执行人工智能评价和建造持续的人工智能监控及视图提出主要建议和领先实务。

在鼓励创新和适应力的环境中**制定人工智能设计准则和建立相关控制。**

设计和实施一个贯穿整个生命周期（战略制定、建造、训练、评价、配置、运营及监控）的点对点人工智能治理及运营模型。

评估现行治理架构并执行差距分析以识别机遇和需改进的领域。

设计一个通过指引、模板、工具和加速器来**交付人工智能解决方案和创新的治理架构，**以快速、负责任地交付人工智能解决方案。

整合一个风险管理架构以识别和优先处理业务关键算法，并纳入敏捷风险缓释战略以应对设计及运营过程中的网络安全、完整性、公平性及适应力考量。

设计及制定标准以在不妨碍创新及适应力的情况下，**维持对算法的持续控制。**

“首先，我们需确保数据是整洁、充足及合理的。然后，确保算法能提供一致的结果，在开始假设时无需依赖小变动。最后，在不对个别利益相关者产生过多负面影响的情况下，确保整体目标的实现。”

Cathy O’ Neil
顾问及“Weapons of Math Destruction, from the introduction to Building Trust in a Smart Society”作者(Sander Klous, 毕马威荷兰)

人工智能评估

对企业人工智能项目及治理进行诊断以评估现行人工智能治理元素的现状和适用性、当前运作模式和人工智能规模扩张的准备情况。这包含能力及成熟度评估以及实现目标状态的路线图和建议。

评估各个人工智能及机器学习算法：控制测试、设计评估以及根据四个信任基础，即完整性、可解释性、公平性及适应能力进行的算法执行及运作。

持续监控及控制面板

创建信任要素相关指标的**全面视图**，包括关键绩效及风险指标，如针对主要相关的人工智能关键绩效及风险指标的**董事会、高管及项目**层汇报。

实现关键控制及指标的持续监控—贵机构的人工智能/机器学习模型哪些凑效或不凑效。

根据控制及测试，**展示某个时间段内的向上/向下趋势。**

可在问题产生时应对及纠正问题。如当学习模型引入偏差或被禁止的特征正被用于决策时。

评估贵机构的人工智能模型或对贵机构更 广泛的企业人工智能项目进行健康检查。

毕马威受控人工智能的主要优势

- **平台中立性**
- 对偏差及准确度以及其它因素的**持续监控**
- **持续保护及安全性**，包括训练数据，以避免对抗性及其它网络攻击
- 能将数据科学术语及概念与关键业务风险指标映射
- 建立展示人工智能模型运作的**全面视图**
- 建立**点对点架构**以管治模型的建造、配置和进化
- 有助**提升人工智能**在企业内部**的应用和规模**

KPMG

人工智能在控制中是如何运作的?

核心组件包括：

全面人工智能架构

人工智能架构通过透明、高效地治理算法，协助机构提升对自身技术绩效的信任

专业的人工智能知识配对

审视人工智能的整体治理及管理架构，从风险角度将其与企业政策及指引进行映射

原型架构能力

一个有助提升数字化及弹性人工智能控制的环境，以计量算法风险

可见度及风险管理面板

向用户展示与信任要素有关的各种指标。

完全释放人工智能的潜力

如今的机构在很大程度上依赖基于算法的应用以作关键的业务决策。此做法固然会带来机遇，但同时也提出了有关信任的问题。随着我们进入了一个算法治理年代，机构必须思考对算法的治理，以提升它们对结果的信任，并释放人工智能的全部潜力。

这就是毕马威的受控人工智能起作用的地方。毕马威成员所认为，人工智能的治理与人的治理同样重要。毕马威专家在一个技术中立的环境中运作，其建议是基于最适合贵机构需求的方案。我们的成员所致力为您提供一个全面、范围广泛的方案，在您的人工智能旅程中助您实现业务目标，从现在，到未来。

read.kpmg.us/AIincontrol

联系我们

Martin Sokalski
主管，
咨询及新兴科技风险服务
毕马威美国

电话：+1312 665 4937
电邮：msokalski@kpmg.com

Sander Klous 教授、博士
主管合伙人，
数据分析，毕马威荷兰

电话：+ 31206 567186
电邮： klous.sander@kpmg.nl

Swami Chandrasekaran
执行总监
毕马威创新与企业方案
毕马威美国

电话：+1214 840 2435
电邮：swamchan@kpmg.com

read.kpmg.us/AIincontrol

kpmg.com/cn/socialmedia

如需获取毕马威中国各办公室信息，请扫描二维码或登陆我们的网站：
https://home.kpmg.com/cn/en/home/about/offices.html