OLIVER WYMAN

信任机器的大脑

作者

Chris DeBrusk, 合伙人
Ege Gürdeniz, 副董事
Shriram Santhanam, 合伙人
Til Schuermann, 合伙人

概述

“如果你不能浅显易懂地解释它，那么说明你没有很好地理解它”

–爱因斯坦

随着运用复杂机器学习技术应用的不断推广，人类对机器决策中行动的理解和解释能力相关的忧虑也随之上升。这一担忧尤其体现在由于对机器模型运算结果缺乏了解而可能对消费者产生切实不利影响的领域，例如金融服务行业贷款申请人受到不公平的对待，医疗卫生行业患者误诊等。

专家和公司使用人工智能的“解释能力explainability”、“透明度”和“可解释性interpretability”等专业用词表述上述挑战。然而，挑战的根本问题归结于人类对机器运算结果的信任能力：依据一项机器计算结算做出影响他人的重要决定，我们必须充分信任机器结果，为了充分信任机器结果，我们必须：

1. 知道机器结果是准确的
2. 充分了解机器结果运算的方法和依据

大部分机构建立了独立的审核体系并设置具有专业资质的检测人员，确保机器运算得出结果的准确和合理（例如金融机构的模型风险管理职能部门）。具有专业资质的第三方审查并验收批准用于实务应用的机器，了解这一点的确有助于对机器系统建立一定程度的信任。然而，独立审查不一定有助于帮助其他人了解机器：对于机器的用户和消费者等受机器影响的各方而言，经由高度专业化的计算机科学家团队独立审查和批准的高效能机器可能依然是无法理解的谜团。

实际上，并非每台机器都需要达到同等程度的理解和解释（以及信任）。考虑到机器的用途和目的、受到影响的人员和方式，以及监管要求等其他因素，对每台机器可以有不同水平的信任级别要求。因此，信任问题的有效解决方案必须考虑到以上各方面的因素。我们通过本文提出机构建立“人机生态系统”信任的适用体系，该体系还促进以负责的方式开展机器学习应用的推广普及。

人工智能或称AI，是计算机科学的一个分支。 机器学习属于人工智能的一个研究领域，专注于特定类别的算法。1959年亚瑟·塞缪尔（Arthur Samuel）创造了“机器学习”一词，将其定义为“不用具体编程，赋予计算机自行学习能力的研究领域”。

随着欧盟《通用数据保护条例》（简称GDPR）出台、数据隐私和可追溯性整体监督的加强以及公众要求的提高，能够向客户解释机器计算结果已成为一项法律义务（至少在欧盟开展业务的公司面临这项法律义务）。我们建议公司立即着手采取措施建立一套完善的体系，以及时达到和满足不断提高的政府监督和公众要求。

挑战

机器学习算法可以有多种形式，相应的复杂程度各不相同。因此，我们对机器输出结果的理解能力可能因具体的学习算法而受到不同方式的影响。但是，从全面和简化的角度来看，理解机器学习应用程序的输出结果在三个方面的因素而面临挑战性：

- 自编程：学习算法可以自行编程、调整参数，并通过某些数学方式和目标函数映射输入和输出定义自己的规则，而不是遵循人类编程的明确规则和逻辑
- 从输入到输出结果的复杂性：作为学习过程的一部分，机器需要调整大量模型组件、数据层和变量参数，这些自我调整参数之间交互的复杂性使得从输入到输出结果的归因分析和链接都极具挑战性
- 机器的网络效应：自编程机器的输出结果可作为另一个自编程机器的输入，无需任何人为参与，这将形成复杂的、不透明的机器网络，理解的难度很大

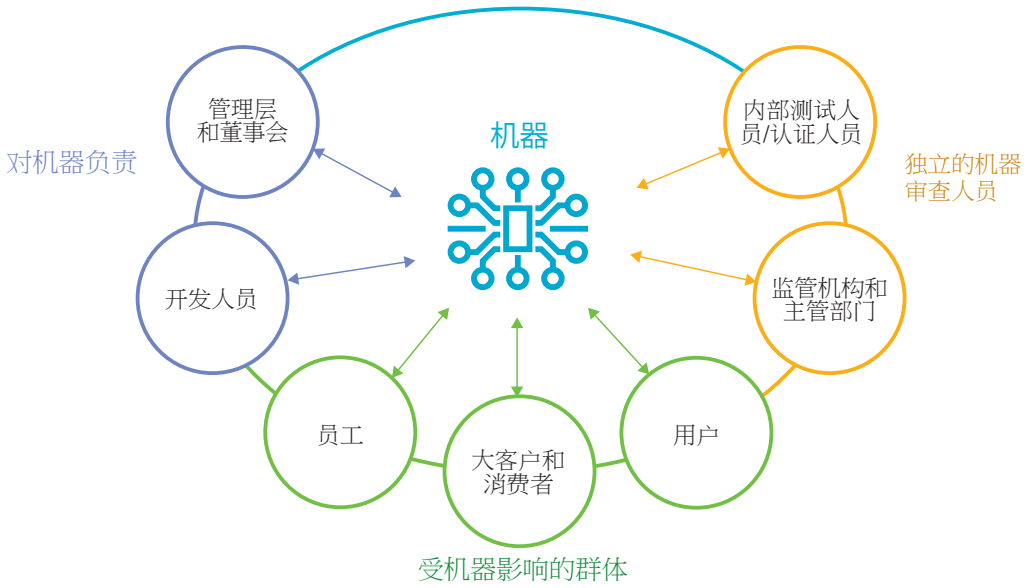
模型可解释性是与模型复杂性程度紧密相关的延续，例如简单回归比多层神经网络更容易理解和解释。然而，虽然通常更容易解释简单的模型，但简单模型往往具有较低的性能（例如较低的预测准确性）。因此，随着公司要求以更高的精确性解决日益复杂的问题，往往需要采用更复杂的方法，如深度神经网络，此类网络包含数十个隐藏层，乃至数千甚至数百万个具有非线性交互的参数，而人类尚无法直观、快速地理解这些参数。随着复杂性的增加，信任机器也越来越困难。

据报道，Google的Deep Mind神经网络拥有约100亿个神经元。另据报道，最大的人工神经网络由Digital Reasoning开发，拥有约1600亿个神经元或参数。相比之下，人类大脑只有约800-1000亿个神经元。

谁需要理解机器输出结果？

“人机生态系统”涵盖与机器发生不同级别、不同类型交互的各种人群，各群体表现出机器输出结果理解需求和能力的水平差异，由此问题变成了“每个群体需要理解多少？”

图1：人机生态系统



要求每个群体完全掌握机器背后的所有原理和计算方式并解释各个方面是既不合理也不现实的想法，另一方面，我们也不能接受每个群体的输出结果理解都非常有限和肤浅。目前业界的传统模型应用数量体现了机器信任问题的上升趋势，很快任何一家公司都可能配置承担起关键决策职能的数百台甚至数千台机器，将进一步加剧问题的复杂性。

鉴于需求和理解能力的差异以及机器类型和数量的多样性，公司无法采用一刀切的方法建立生态系统信任。在遵循一些客观标准的机器分类基础上，奥纬咨询提出“层次化机器信任体系”。

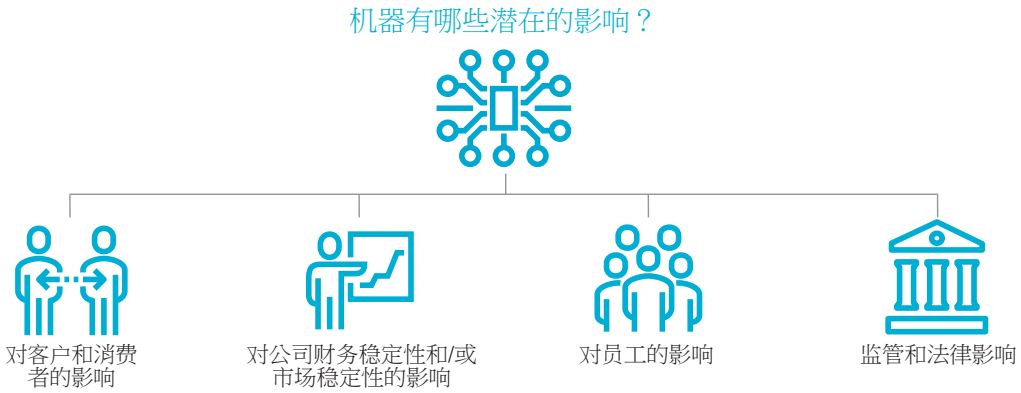
Google翻译在使用人工智能前使用了一套人工定义的规则和逻辑，以词段划分方式翻译文本。现今，Google翻译使用神经网络机器翻译，通过学习语言的语义翻译整个语句，而不是遵循设定的一系列规则。复杂的神经网络机器翻译方法成功将翻译错误减少60%。

层次化机器信任体系

如果机器结果产生影响后果，例如影响客户和消费者履行法律义务或影响公司的财务状况，那么在层次化信任体系中应居于较高的位置，同时每个受影响群体的模型输出结果理解要求应适当提高。然而，如果机器结果不存在影响后果，例如机器结果仅用于不重要的、有限的内部工作以提高团队效率，那么这类机器处于信任体系的较低位置，并定义相应的输出结果理解要求。

作为工作的起点，奥纬建议从四个影响维度开展机器评估，在此基础上确定机器输出结果的潜在影响。

图2：机器影响评估



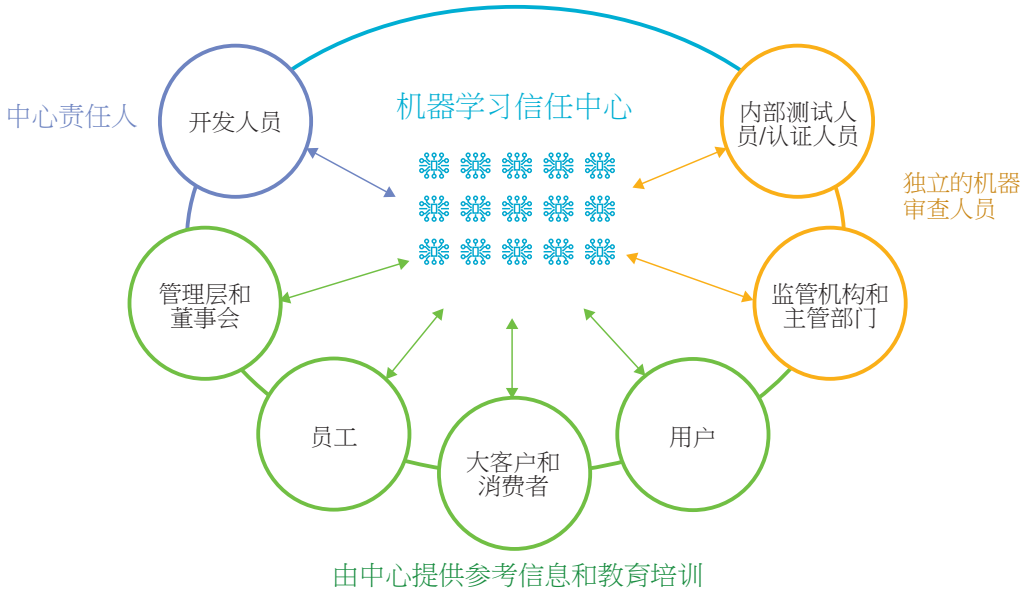
未来每个公司和机构对上述影响分类都将设立自己的临界值和定义标准，然而最终每台机器都将归于四个维度“从高到低的”的评估范围内，为确保受影响的各方群体充分理解机器输出结果还需要开展各项工作和努力。那么，谁将负责开发和主导信任体系？谁负责采用信任体系完成机器评估？谁负责最终向受影响的各方群体解释模型输出结果？

“聚类”是机器学习的常见形式，这一算法可能消耗大量数据集（例如数百万条每日客户交易），发现人类智慧无法识别的隐含特征和发展趋势。

机器学习信任中心

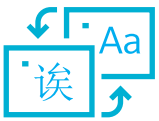
鉴于信任系统所涉及的机器和受影响群体的数量庞大，并且需要确保信任系统在方法论方面保持统一，因此最佳方案是集中应对和解决这一挑战。因此，奥纬建议指派公司现有的机器开发职能部门（数据科学主管、人工智能或数据分析主管）承担起“机器学习信任中心”的工作，一方面负责主导和实施层次化机器信任体系，同时开发帮助受影响群体理解和信任机器所需的工件。

图3：机器学习信任中心



机器学习信任中心（简称“中心”）主要承担以下机器解释责任：

- 测试：开展大量的定量测试，评估模型输入因素的重要性及其对输出结果的影响 – 例如：
  - 创建和评估体现每个变量特征影响模型预测方式的部分因变量关系
  - 运用各种测试技术评估输入扰动对模型输出结果的影响（如LIME1）
  - 审查变量特征重要性评分，评分代表变量特征在确定输出结果方面的价值和效用
  - 检查变量系数



长期以来机器翻译一直依赖英语作为不同目标翻译语言之间的桥梁。另一方面，谷歌神经网络机器翻译算法通过自学进行翻译，摒弃英语作为桥梁。该神经网络通过开发自有的“机器语言”完成自学，“机器语言”实现了多种语言之间的高效翻译。

- 数据审查：完整分析数据来源的系统方法，追溯用于模型培训的数据，旨在识别和矫正潜在的偏差区域
- 文档记录：创建用户使用友好的文档记录，文档记录全面汇总定量测试结果和完成的全部定性评估（例如机器输出结果的对比性解释），并且提供非技术、直观的模型输出结果驱动因素解释
- 步骤程序：定义和实施机器开发规范标准和步骤程序，确保机器开发达到透明公开和统一，并且机器输出结果可由独立第三方进行复制
- 监控和报告：根据机器解释和信任层级，以适当频率持续或定期监控模型输入变量和输出结果，并将结果报告给管理层等相关方
- 教育培训：设计和执行针对性的教育培训计划、研讨会和沟通宣传（内部或外部）。例如，实施非常重要的机器时可以开展相应的培训研讨会，为相关方提供新机器培训
- 消费者支持：为消费者/员工提供机器相关问题的服务支持。例如，消费者询问机器人拒绝其借款申请的原因，或者销售人员询问机器向客户建议推售某类产品的原因

由机器学习信任中心决定机器信任体系每个层次所适用的上述组成方面。然而，为确保所采用方法的合适，最终由内部审计、风险管理职能部门和主管等第三方独立审核和质询中心的各项决策和体系。

结论

各种行业出版物、学术论文和主流媒体正在全力宣传和描绘大规模成功应用机器学习带来的收益潜力，每天不断出现的新实务案例、应用和实验进一步助推了机器学习为公司和消费者创造价值潜力的乐观激动情绪。

然而，由于机器输出结果造成非故意不利后果的风险太大，因此缺乏对人机生态系统的信任可能阻碍机器学习的大规模应用。另一方面，公司可能也缺乏面对潜在的监管、法律、道德或财务后果的主观意愿。为避免实务应用过程中的障碍，公司亟待明确适合自身特点的“机器学习信任”模式，并立即着手实施相应的指导原则和要求。

Deep Blue（深蓝）是一款国际象棋电脑，遵循人类编写的硬编码规则（相对易于观察和理解），曾击败Gary Kasparov。AlphaGo通过观察成千上万的游戏和数百万次动作进行自学，使用深度神经网络（非常难以观察和理解），击败了中国围棋大师柯杰。

奥纬咨询是全球领先的管理咨询公司，融合专深的行业知识与战略、运营、风险管理和组织转型课题洞见。

欲了解更多信息，欢迎与奥纬咨询市场部联系。敬请发送电邮至：insights.digital@oliverwyman.com，或致电全球各地区办事处：

美洲	欧洲、中东及非洲	亚太
+1 212 541 8100	+44 20 7333 8333	+65 6510 9700

作者

Chris DeBrusk, 合伙人
Chris.DeBrusk@oliverwyman.com

Ege Gürdeniz, 负责人
Ege.Gurdeniz@oliverwyman.com

Shriram Santhanam, 合伙人
Shriram.Santhanam@oliverwyman.com

Til Schuermann, 合伙人
Til.Schuermann@oliverwyman.com

OLIVER WYMAN