MinterEllison LLP

企业人工智能系列：价值、控制与责任

探讨采用企业人工智能所衍生的法律、商业和治理问题

2026 年 11 月 13 日

企业人工智能究竟需要多高程度的可解释性？从实践角度看透明度、问责与治理

作者：苏振国、梁嘉颖

本系列的首篇文章探讨了保留对数据的权利及控制权与保留这些数据所产生的价值两者之间的区别。第二篇文章则探讨企业人工智能的价值、控制权及责任将如何分配。本文将焦点转向可解释性，探讨其在实践中的意义。

机构正将企业人工智能应用于面向客户的业务、日常运营及决策支持流程。董事会、管理团队及监管机构日益关注人工智能是否具备可解释性。然而，采用企业人工智能的机构往往并不拥有、也无法审视人工智能系统所采用的技术，更不用说理解其中的“黑箱”问题。所谓“黑箱”问题，是指现代人工智能系统的内部运行机制并不透明，即使是开发或采用该系统的人，也难以在实际操作中加以审视。人工智能系统虽能得出结论，但没有人能像解释人类作出的决定一样，清楚说明得出该结论的完整推理过程。

机构或许无法审视底层模型，也无法完全理解人工智能如何得出某一结论，但仍可能需要解释该结论并调查相关事件，以回应监管机构的问询，并维持利益相关方及公众的信任。

企业人工智能往往造成“视野盲区”：机构对人工智能系统如何运行的了解可能有限，却仍须对系统的使用承担全部责任。

因此，真正的挑战并非能否解释人工智能生成内容的每一个环节，而是机构是否充分掌握相关信息，能够负责任地治理这项技术，并对其使用持续承担责任。

> “问题不在于机构是否完全理解人工智能模型如何运行，而在于是否充分掌握相关信息，能够对系统的使用持续承担责任。

不同群体对可解释性的理解各有不同

就本文而言，可解释性是指系统本身或采用该系统的机构，即使在底层模型本身未必能够被完全解读的情况下，仍能够在事后就某项预测或决策提供清晰、有条理的说明。对于人工智能系统的可解释性，不同群体可能关注不同问题。

- 对董事会及高级管理层而言，重点通常是治理及问责。核心问题很直接：我们能否合理说明为何依赖这个系统？
- 对法律顾问及合规团队而言，可解释性往往涉及责任。如果出现疑虑，机构能否获取所需资料，以解释事情的经过、调查问题，并证明已实施适当监督？
- 采购团队的考虑则有所不同。他们关注供应商会提供哪些资料、承诺及合同权利。
- 监管机构通常从问责角度审视可解释性：机构能否证明已实施有效的治理、监测及人工监督？

可以从人工智能系统供应商处获取什么，又有哪些限制？

不少机构的做法，是向人工智能系统供应商索取更多有关系统如何作出决策的资料。这种反应合情合理，却未能解决根本问题。实际上，无论由谁提出要求，又或如何提出，可获取的资料始终有限。

机构通常可以获取系统文档、模型卡、治理政策、风险披露、安全认证及性能等资料。这些资料有助于机构了解应如何部署及治理人工智能系统。

然而，部分资料很难获取，甚至因“黑箱”现象而根本不存在。客户一般无法获取完整的训练数据集、专有源代码或详细的内部开发流程。商业保密、知识产权保护及供应链的复杂性，使大多数企业人工智能应用难以做到完全透明。

因此，实际问题并非有哪些资料存在，而是当机构需要说明情况、履行责任时，有权获取并依据哪些资料。

可解释性的核心其实是问责

机构或许永远无法完全理解所采用的人工智能系统的基础模型。即便如此，机构仍可能需要解释受该模型影响或由其产生的决策，并调查相关投诉、说明治理安排的依据、回应监管机构的问询，以及解答客户和利益相关方的疑问。

我们的目标并非实现完全可解释，而是确保机构能够充分承担责任。这意味着机构须充分掌握相关信息，以了解相关风险、监督人工智能系统的部署、监测其性能、调查事件，并在需要时为治理决策提供支持。因此，当系统产生非预期结果时，机构应能回答以下问题：

- 发生了什么事？
- 为什么会发生？
- 当时有哪些监控措施及监督机制？

因此，可解释性不仅是技术难题，更是治理问题。关键不在于模型的每个部分是否均可解读，而在于机构是否掌握足够资料，并具备适当的监督机制，以对人工智能系统的使用及其产生的结果负责。董事会很少需要了解人工智能模型的架构，却必须了解自身是否设有适当的治理、监督及上报机制。要求完全可解释或许并不切实际，但承担采用人工智能所带来的责任，并非遥不可及。

可解释性如何成为治理及合同议题

从治理角度而言，机构日益需要建立监督架构、上报程序、监测安排及问责框架，以监督人工智能的使用。

香港政府下设的数字政策办公室最初制定《人工智能道德框架》，原本旨在为政府部门提供帮助，现也普遍供在香港规划、设计及使用人工智能系统的机构参考。该框架以十二项人工智能道德原则为基础，包括透明及可解释、人为监督、公平、问责及合法。在治理架构方面，该框架建议设立三道防线：由项目团队负责日常的人工智能应用开发和风险评估；由项目督导委员会及项目保证小组负责独立审核及审批；再由资讯技术委员会或首席资讯官负责持续监测人工智能应用的策略及风险，并在适当情况下由外部顾问提供支援。

香港其他监管机构的导向也相近。香港个人资料私隐专员公署发布的《人工智能（AI）：个人资料保障模范框架》着重强调人工智能全生命周期的责任分配、风险评估、人工监督、透明度及治理。香港金融管理局（“金管局”）就面向消费者的应用生成式人工智能发布指引，要求认可机构采用“人在回路中”的方式，即在决策过程中保留人类控制权，并就人工智能生成的结论维持清晰的治理及问责架构。金管局最近发布《负责任应用人工智能的监管指引》，并推出“ART”框架，支持负责任地使用人工智能，其中一项重点是贯穿人工智能全生命周期的问责框架。上述文件虽不具备法律效力，但可能影响监管机构对负责任人工智能治理的合理标准，以及机构应具备哪些治理措施。

在中国内地，具有约束力的法规将透明度及问责转化为具体义务。《生成式人工智能服务管理暂行办法》（以下简称“暂行办法”）适用于在中国内地向公众提供的生成式人工智能服务，但不涵盖未向公众提供此类服务的企业研发及应用活动。在此适用范围内，《暂行办法》要求采取有效措施，提高服务透明度及生成内容的准确性和可靠性，并披露服务的适用人群及投诉处理机制。重要的是，服务提供者必须配合监督检查，按要求说明训练数据来源、算法机制等事项，并提供必要的技术及数据支持。虽然机构无须记录所使用的每一项资料，但仍应留存应对监督检查所需的资料，并作出相应的合规安排，同时遵守适用的记录保存义务，以及有关保留用户输入内容及使用记录的限制。

欧盟《人工智能法案》提供了更详细的人工智能系统分类。该法案将人工智能系统分为四个风险等级：（1）“不可接受风险”系统被全面禁止；（2）“高风险”系统须遵守法案最严格的义务，包括建立风险管理系统、数据治理监控措施、技术文档、记录保存（运行日志）及人工监督安排，以及符合准确性、稳健性和网络安全要求；（3）“有限风险”系统须履行的义务较少，例如告知用户其正在与人工智能互动；（4）“最低风险”系统则不受法案特定义务的约束。

美国的监管情况则有所不同，也较为分散。在联邦层面，美国国家标准与技术研究院（National Institute of Standards and Technology）的《AI Risk Management Framework》虽属于自愿采用而非强制性规范，但已成为广泛使用的参考依据。该框架将人工智能治理分为四项功能：（1）治理（Govern）：在机构内建立风险管理框架；（2）识别风险背景（Map）：明确相关背景，以界定人工智能系统的风险；（3）衡量（Measure）：运用不同工具、技术及方法，分析、评估、比较和监测人工智能风险及相关影响；（4）管理（Manage）：定期针对已识别及衡量的风险分配资源，并制定事故应对、恢复及沟通计划。美国各州也逐渐出现类似趋势。例如，《Colorado Automated Decision-Making Technology Act》将于 2027 年 1 月 1 日生效，规定了披露预期用途、已知限制及训练数据，以及对人工智能辅助决策进行人工复核及问责等要求。

各地监管发展的细节虽有不同，方向却一致。机构应具备足够的监督安排及文档记录，以说明人工智能系统的使用情况，而非解释底层模型如何运行。但在实际操作中能否符合上述要求，也取决于人工智能系统开发者与采用该系统的机构之间的合同安排。

与其追求无法实现的技术透明度，机构应考虑自身是否享有适当权利，以获取负责任地治理及监督人工智能系统所需的资料。以下五个问题最为重要：

- 人工智能系统供应商必须披露哪些资料？
- 哪些资料可以被独立核实？
- 如果出现问题，能否重现相关输入内容、输出结果及决策过程？
- 人工智能系统供应商是否会协助调查、审计及回应监管问询？
- 重要的输出结果能否被复核、质疑或推翻？

事件处理及调查支持可能尤为重要。当人工智能辅助决策引起疑虑时，机构通常需要解释事情的经过、评估影响，并证明已实施适当监督。即使相关技术由第三方提供，机构仍可能须承担这些义务。因此，机构应考虑是

否需要在合同中要求供应商及时配合，提供相关文档、记录及技术协助，并支持调查、审计及回应监管问询。

这五个问题决定了机构在实际操作中能否持续履行监督责任。不少机构所称的“可解释性”，其实是获取资料的权利、监督机制及治理监控措施的结合，使机构充分掌握相关信息，以了解及调查人工智能系统的使用情况，并就此作出说明。

结论

机构经常问，某个人工智能系统是否具备可解释性。然而，这未必是最有用的问题。企业人工智能往往涉及采用系统的机构无法全面审视或理解的技术，但机构不会仅因掌握的信息有限，就无须承担责任。

更实际的问题是机构是否具备足够的透明度、监督安排及资料，能够负责任地治理人工智能。在不少企业人工智能应用中，机构能否承担监督责任，并不取决于能否全面掌握模型本身的运行机制，而在于能否获取所需资料，以解释、调查及治理这项技术的使用。机构未必需要完全了解人工智能模型如何运行，但必须充分掌握相关信息，才能对其使用持续承担责任。

我们可以提供哪些帮助

随着机构持续采用企业人工智能，各司法管辖区对治理的要求和期望也在迅速变化。我们可以协助客户审视其人工智能治理及问责框架、开展企业人工智能采购及供应商评估、审查合同及谈判涉及透明度的条款、向董事会及管理层介绍人工智能治理义务，以及制定人工智能风险管理及监督框架。

机构所需的可解释程度取决于它的目标、风险状况及监管环境。目前的难题并非如何做到完全透明，而是如何确保即使未能完全解释系统的运行机制，机构仍能承担监督责任。

关于作者

苏振国是铭德有限法律责任合伙律师事务所的合伙人。他为企业提供治理、监管和风险管理相关的法律意见，包括因采用新技术而产生的问责、监督和技术风险。

梁嘉颖是铭德有限法律责任合伙律师事务所的见习律师。她尤其关注科技监管与数据治理，以及采用新技术所涉及的风险及相应的风险管理策略。



苏振国
合伙人，香港
T: 2841 6879
E: eddy.so@minterellison.com



梁嘉颖
见习律师，香港
T: 2841 6951
E: betty.liang@minterellison.com

铭德及有关办事处：

阿德莱德 奥克兰 布里斯班 堪培拉 达尔文 黄金海岸 香港 伦敦 墨尔本 珀斯 悉尼 乌兰巴托 惠灵顿

本通讯由铭德有限法律责任合伙律师事务所编写，用于略举相关时期的某些事项，仅供参考，并不旨在提供全面内容。本通讯并非用于提供法律意见，其内容也不构成法律意见，任何人亦不应出于任何一般目的或就特定交易及/或情况依赖本通讯的内容。您应在将本通讯的信息应用在特定情况之前寻求专业意见。如果您对本通讯有任何疑问或意见，请联络您在铭德有限法律责任合伙律师事务所的日常联系人。铭德有限法律责任合伙律师事务所对于基于依赖本通讯的信息而造成的任何损失不承担任何责任。

© 2026 铭德有限法律责任合伙律师事务所