

代理型AI风险与 金融服务业韧性

关于人工智能治理，10个问题应该被问及：

金融机构一直都在从事风险定价的业务。每一项信贷决策、交易、保险单以及客户互动都包含着机构需要评估和管理并承担相应责任的风险。正是因为这种问责制被高度重视，金融监管的整个架构，从资本要求、模型验证到行为准则和审计标准，才得以存在。

自主型人工智能并未消除这种问责制。改变的只是问责制的执行机制。金融风险治理建立在一个假设之上，该假设在其大部分历史时期内都成立：错误以人类速度发生，一次一个决策。一名贷款员批准了信用不良的贷款。一名交易员做出了错误的判断。一名客户关系经理错误地销售了产品。这些事件代价高昂，会发生，但它们发生在旨在逐个捕获它们的控制框架之内。自动化决策多年来一直在侵蚀这一假设，而自主型人工智能则消除了其残余。

自主AI改变了控制框架需要“捕捉”的内容。 主要挑战并非规模；信用模型多年来一直在加速决策，交易算法自20世纪90年代以来以毫秒级速度调整头寸，欺诈检测系统在过去十年中近乎实时地处理了数百万笔交易。行业已习惯于规模——它在模型风险框架下管理这些系统，并且总体上做得相当称职。AI代理的不同之处不在于它做出了多少决策，而在于有多少决策事先并未指定。它的输出空间并非固定。它选择自己的答案路径，调用工具，并且可以在批准时无人预料的方式行事。这是一种范围而非处理能力的改变，这也是现有控制框架感到挣扎的原因。我们无法用人类速度的治理来管理机器速度的风险。

模型只能在你们预料的方面出错。而智能体可以在无人列举的方面出错，在任何人看到模式之前，这种情况会重复成千上万次。然后，这些错误就被规模化生产了。

横向的，而非各自为政的风险

1987年，一系列各自理性且管理得当的相关投资组合保险计划同时触发，导致道琼斯工业平均指数单日下跌22.6%。十年后，长期资本管理公司（LTCM）进行约25倍杠杆的相对价值趋同交易。当1998年8月俄罗斯违约引发避险情绪时，利差变动远超其模型所认定的合理范围，该基金未能持有头寸足够长的时间来证明其正确性。为避免混乱的清算，纽约联邦储备银行协调14家机构注资36亿美元对其进行资本重组。未动用公共资金。这两起事件都带来了同样的教训：风险已经超出了为控制它而建立的管理机制。

由此而重建于整个行业中的治理架构，如今已成为其脆弱之处。它依然围绕个体风险类型进行组织，例如信用风险、欺诈风险、市场风险和运营风险，每种类型都有其自身的流程、监督机制和治理结构。人工智能并不尊重现有的架构。它具有横向穿透性和快速性。单个人工智能系统可以同时影响信用决策、欺诈检测、定价、客户建议以及运营流程。当其偏离方向、失效或被攻破时，其影响将以机器速度波及所有风险类别。

结果是在该机构应负责的范围与其当前控制力所能实际监管的范围之间存在治理差距。这种差距体现在四个不同的方面。

人工智能带来的风险环境四大变化

自主权风险

经典模型风险框架假设一个具有固定输入、离散输出的模型，并围绕这两者构建验证流程。而智能体打破了这一假设。它可以规划多步任务、调用工具、将任务委托给其他智能体，并在两次运行相同提示时选择不同路径得到不同答案。没有单一决策需要验证，只有一系列相互关联的决策链，其执行速度远快于人肉审查者能够跟进。审计这一决策链，不仅要了解智能体做了什么，还要理解其推理过程，这是大多数机构目前尚未具备的能力。为单一模型和单一输出构建的风险框架不能简单地扩展以覆盖这种情况，它必须被重新构建。

2. 对抗风险

恶意行为者正直接针对人工智能系统，而不仅仅是客户。逆向工程模型、设计对特定系统看似合法的输入，或在人工智能处理的内容中嵌入隐藏指令等对抗性技术，已不再是理论概念。2025年，一个广泛部署的企业级人工智能助手曝出了一个关键漏洞：一封精心构造的电子邮件即可导致系统在无需用户交互的情况下静默地窃取内部数据。攻击面涵盖了所有摄入外部内容（如电子邮件、文档和申报文件）的人工智能系统，而这类系统现在构成了大多数大型金融机构技术架构的重要组成部分。

3. 相关风险

当机构部署基于相似数据训练、以相似目标进行优化的AI系统时，整个市场都存在以机器速度进行关联行为的条件。1987年的股灾是最清晰的先例：每个程序在孤立状态下都是理性的，但在整体汇总时却同时触发。如今，在AI驱动的信贷、交易和风险管理系统中，存在类似动态的条件，覆盖更广泛的资产类别，并且执行速度比当时更快。许多机构都在运行代理程序，用于扫描信号、重新平衡敞口和执行交易。当触发条件出现时，这些系统不会犹豫。当人们恐慌时，他们会打电话给某人，并且可能在行动前等待。而算法不会。

4. 检测延迟

当一个AI模型带有缺陷假设时，它不会立即失效。它能在规模上持续稳定运行数周或数月，而问题则会逐渐累积。信用损失、有害行为、模型漂移和定价错误等问题可能要到后期才会显现，届时风险敞口早已累积扩大。AI风险在执行上可能很快，但在发现上却很慢。这种结合使其与现有周期性控制旨在捕捉的风险在结构上有所不同。

在当今大多数机构的AI治理问题上，人人都有份，但又无人拥有。

发展最快的机构正在为AI风险指定一名负责的负责人——不是委员会，也不是由首席风险官、首席技术官和首席数字官分担的共同责任——而是一个单一的、有明确姓名的个人，其绩效评估取决于这一点。高级问责制度已经在许多市场存在，包括英国的“高级管理人员和认证制度”和澳大利亚的“财务问责制度”。然而，在大多数机构中，目前还不存在将这些制度与单个AI系统进行映射的做法。通常，一名高管可以说明他们负责哪些业务服务。但能说出哪些AI系统代表他们行事、这些系统被允许做什么以及谁批准了这种权限的人却少得多。

响应不能超过人为疏忽

治理缺位时，本能的制度性应对是加强监管：增加审查者、增加委员会和增加检查点。然而，在人工智能赋能的环境下，这种应对方式将无法扩展。

无风险团队无法以AI生成信用决策的体量和速度进行逐一审核。没有任何合规职能能够监控每一次AI辅助的客户互动。任何欺诈团队都无法实时调查每一个异常模式。这不符合逻辑。

更有效的应对方式是将人工智能视为新风险环境的来源，同时也是管理它的主要工具。那些持续监控模型漂移、异常行为、行为风险信号和对抗性模式的AI系统，相当于从定期到实时风险管理转变所要求的持续控制。

玻璃翼计划（Project Glasswing），一个于2026年4月启动的联盟，由亚马逊网络服务（Amazon Web Services）、Anthropic、苹果（Apple）、博通（Broadcom）、思科（Cisco）、CrowdStrike、谷歌（Google）、摩根大通（JPMorgan Chase）、Linux基金会（Linux Foundation）、微软（Microsoft）、英伟达（NVIDIA）和帕洛阿尔托网络（Palo Alto Networks）作为创始合作伙伴共同发起，具体展示了大规模AI防御的形态。该项目利用前沿AI模型扫描了超过1000个开源项目，并识别出超过6000个高或严重等级的漏洞，其中一些漏洞已存在于广泛部署的软件中近三十年。数个合作伙伴报告称，与先前的安全审查流程相比，漏洞检测率提高了十倍以上。能够被用于攻击金融基础设施的同类能力，正被有意地用于防御它。

对人类判断的影响是，它并没有变得不那么重要，而是在更小的决策集中变得更为重要。

这些问题是治理设计层面的问题。从根本上说，它们也是领导力层面的问题。

2026年人工智能治理的两大优先事项

承认人工智能风险与实际治理人工智能风险之间的差距很大。这也正逐渐成为监管期望以及良好实践。新加坡金融管理局和香港金融管理局均已发布直接针对人工智能模型治理的指引，澳大利亚的CPS 230要求机构将关键系统的运营弹性（包括由人工智能驱动的系统）视为董事会层面的问责责任。对于在区域内运营的机构而言，以下两项优先事项正逐渐成为底线，而非差异化因素。

1. 人工智能实际影响风险决策的地图

与其拥有全局概览，不如专注于人工智能正在作出或实质性影响具有财务、行为或运营后果的决策之处，例如信贷审批、定价、理赔、建议、欺诈检测、催收和客户沟通。此类映射练习的输出几乎总会揭示比高层领导预期的更广泛的足迹，包括人工智能在那些原本未被评估为人工智能部署的供应商工具中运行。第三方覆盖并非可选项，因为你的风险边界已扩展到人工智能代表你作出影响的任何地方。在实践中，这更像是一个在决策层面持有的登记册，而非用例列表：对于每个重大的决策，是什么在影响它，该系统被允许做什么，决策由谁拥有而非系统拥有，以及其推理记录保存在哪里。

2. 制定失败计划，而非假设其不会发生

金融机构已为信用损失计提准备金、持有资本以防范市场风险，并进行流动性压力测试。它们这样做是因为假设损失不会发生并非可接受的治理立场。同样的逻辑也应适用于人工智能。如果一个机构无法说明其关键性人工智能系统在发生故障时的表现，或无法说明如何使受影响的客户得到妥善处理，那么无论其部署前的验证多么稳健，都存在治理缺陷。

针对该对话，对于任何单一的物料AI系统，一个起点是一套董事会无需请求技术团队翻译就能回答的问题：

- 该人工智能系统正在做出何种决策或产生实质性影响？如果该系统做出错误决策，客户或资产负债表将面临何种后果？
- 谁对整个AI系统及其做出的决策负有个人责任——是以姓名，而非以职能来界定？
- 它是自主运行，还是需要人工批准，并且那条线是故意画的，还是默认设置的？

- 哪些决定，无论性能如何，我们都已完全排除了自动化？
有文献记载吗？
- 如果它失败了，或者开始表现得与它被批准当天不同，那它算不算失败呢？
安全地，我们多久才会注意到？

事后，我们能解释它为何做出特定决策吗？

- 在人类察觉之前，它可能造成多大的损失？

我们上次尝试像攻击者那样输入敌对数据来破解它时，是什么时候，以及发生了什么？

我们其他的AI系统以及我们竞争对手的系统，有多少是运行在相同的模型、供应商或数据之上的，如果它们在同一天以相同的方式移动，会发生什么？

- 谁能把它关掉，而且有多快？

人工智能治理也是该地区监管机构关注的重点。新加坡金融管理局和香港金融管理局均已发布直接针对人工智能模型治理的指引。澳大利亚的CPS 230要求董事会对其关键运营的运营弹性负责，包括现在由人工智能支持的那些运营，而针对特定个人的问责则根据《财务问责制度》另行处理。日本采取了有意的咨询路线：金融服务局于2025年3月发布了其《人工智能讨论文件》，即《促进金融领域人工智能健康应用初步讨论要点》，同年下半年通过公私合办的人工智能论坛进行磋商，并于2026年3月发布了1.1版，目前仍开放意见征询。

当一个监管机构仍在形成其观点时，该机构承担着定义自身治理、证明其运营并日后捍卫它的工作，而上述问题正是检验其能否做到这一点的公正标准。

一种不同的比赛

金融领域AI应用的下个阶段，将不取决于谁部署了最强大的模型，而是取决于谁能构建治理它们的制度能力，包括理解AI的运作方式、其优化目标、其可能失败之处，以及在机器速度下后果会呈现何种样貌。

将人工智能治理视为合规性工作的机构会构建框架然后继续前进。将人工智能治理视为战略能力的机构则会构建与之匹配的人工智能速度韧性，以应对人工智能速度风险。在一个信任是产品的行业中，这种区别至关重要。

机器速度的风险不会等待人类速度的治理。

关于 Quotient — 奥利弗·怀曼的 AI

由 Oliver Wyman 提供的商誉，AI（商誉）结合了深厚的行业洞察力和实用的AI专业知识，旨在提升绩效、增强韧性、调动资本并驱动增长。我们与组织紧密合作，助其自信扩展规模，并构建持久的竞争优势——将AI的潜力转化为实际成果。

威灵顿咨询公司（Oliver Wyman）， Marsh（纽约证券交易所代码：MRSH）旗下的业务部门，是一家管理咨询公司。该公司凭借深厚的行业洞察力、大胆的创新精神以及一种能够化繁为简、协作共赢的方法，帮助组织应对其最具定义性的转型时刻。Marsh是全球在风险、再保险和资本、人力资源与投资以及管理咨询领域的领导者，为130个国家的客户提供建议。公司年营收达270亿美元，拥有超过9.5万名同事，Marsh致力于通过独特的视角，帮助客户建立信心，从而蓬勃发展。

欲了解更多信息，请访问 oliverwyman.com，或在领英（LinkedIn）和推特（X）上关注我们。

美洲
+1 212 541 8100

欧洲
+44 20 7333 8333

亚太地区
+65 6510 9700

印度、中东和非洲
+971 (0) 4 425 7000

作者

Jayant Raman
Partner, Head of AI Partnerships
jayant.raman@oliverwyman.com

Ashwini Karandikar
Principal
ashwini.karandikar@oliverwyman.com

版权所有。未经 Oliver Wyman 书面许可，本报告的全文或部分内容均不得被复制或重新分发。
且， Oliver Wyman 对此方面第三方行为的任何责任概不承担。

本报告中的信息及观点由 Oliver Wyman 整理。本报告不构成投资建议，亦不应被视为投资建议或替代就相关事宜咨询专业会计师、税务师、律师或财务顾问。Oliver Wyman 已尽合理努力确保所使用的信息和分析可靠、最新且全面，但所有信息均未经保证，亦不构成任何明示或暗示的保证。Oliver Wyman 不承担更新本报告中信息或结论的责任。对于因依据本报告或本报告中引用的任何其他报告或信息来源所采取或未采取的任何行动而导致的任何损失，或对于任何后果性、特殊或类似损害，即使已被告知此类损害的可能性，Oliver Wyman 亦不承担责任。本报告不构成买卖证券的要约或招揽买卖证券。未经 Oliver Wyman 书面同意，本报告不得出售。