

风险：模型、偏差、羊群效应与欺诈

Dr Alexander Ressos, TEP
Attorney-at-Law (Switzerland), Solicitor (Hong Kong),
Solicitor (England & Wales)

金融服务中人工智能的风险可分为监管机构和文献所识别的若干类别。模型风险是最古老的一类。每个模型都是对现实的简化，在条件发生变化时可能失效。机器学习模型尤其容易受此影响，因为它们从历史数据中学习规律，而这些规律在未来未必成立。大型模型的复杂性使得人们难以像银行自 2000 年代以来验证统计模型那样对其进行验证。¹

偏见和歧视构成第二类风险。以历史决定训练的信贷模型，即使不直接使用受保护的特征，也可能复制这些决定中的模式，包括对某些群体的任何不利。瑞士法对私人之间的歧视没有一般性禁止，但《瑞士民法典》（ZGB）第 28 条对人格的保护以及一些专门法律提供了一定的保护。瑞士金融市场监管局（FINMA）的指引将平等对待列为其关注领域之一，而欧盟（EU）《人工智能法》则通过其对高风险系统数据治理的要求来处理这一问题。²可解释性与此密切相关。如果模型无法说明它为何拒绝了一笔贷款申请或标记了一笔交易，客户便无法对该决定提出异议，机构也无法证明其行为合法。

金融稳定是巴塞尔各机构所关注的问题。如果许多市场参与者使用以相似数据训练的相似模型，他们的决定可能趋于相关，这可能放大市场波动，并减少维持市场流动性的观点多样性。Gary

Gensler（在其出任美国证券交易委员会主席之前）与 Lily Bailey 在 2020 年的一篇文章中描述了这种羊群效应风险，认为深度学习可能通过同质化、网络互联以及监管空白而加剧系统性风险。³金融稳定理事会补充指出，模型、云服务和芯片集中于少数几家提供者，这为金融体系造成了单点故障。⁴

网络风险和欺诈随着这项技术一同增长。攻击者利用语言模型以任何语言（包括瑞士德语）撰写令人信服的网络钓鱼信息，并找出软件中的漏洞。语音克隆和视频深度伪造使犯罪分子能够冒充高管、客户或亲属。在 2024 年一起被广泛报道的案件中，工程公司 Arup 在香港的一名员工在一次视频会议后转出了约 2500 万美元而会议中的其他参与者，包括表面上的首席财务官，都是由人工智能生成的。⁵国家网络安全中心（该中心于 2024 年改组为联邦网络安全办公室）报告称，瑞士此类企图的数量正在增加。⁶对银行及其客户而言，其后果是：屏幕上的声音或面孔已不再能证明身份，因此付款指令需要通过另一独立渠道加以核实。

1 注释

1. 关于模型风险管理，参见美国联邦储备委员会和货币监理署的监管指引 SR 11-7, Guidance on Model Risk Management, 2011 年 4 月，该指引在国际上对实践产生了影响。有待核实。
2. 《瑞士民法典》（ZGB, SR 210）第 28 条；FINMA 第 08/2024 号指引；欧盟第 2024/1689 号条例第 10 条，关于数据和数据治理。有待核实。
3. Gary Gensler 与 Lily Bailey, “Deep Learning and Financial Stability”, 工作论文, MIT, 2020 年 11 月。有待核实。
4. 金融稳定理事会, The Financial Stability Implications of Artificial Intelligence, 2024 年 11 月 14 日，关于第三方依赖和市场相关性的章节。有待核实。

5. 香港警方的报告以及 2024 年 2 月和 5 月国际媒体关于针对 Arup 香港办事处之深度伪造欺诈的报道。数字有待核实。
6. 联邦网络安全办公室（BACS），2024 年以来关于欺诈和深度伪造的半年度报告及每周回顾。现状有待核实。

仅供一般参考。本文不构成法律或税务意见，不能替代针对个案的专业意见，亦不建立律师与客户关系。本文反映截至 2026 年 9 月的法律状况。

在线版本: cn.helveticaregister.com/basel-risks-models-bias-herding-and-fraud