

風險：模型、偏差、羊群效應與欺詐

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

金融服務中人工智能的風險可分為監管機構和文獻所識別的若干類別。模型風險是其中最古老的一類。每個模型都是對現實的簡化，在條件發生變化時可能失效。機器學習模型尤其容易受此影響，因為它們從歷史數據中學習規律，而這些規律在未來未必成立。大型模型的複雜性使得人們難以像銀行自 2000 年代以來驗證統計模型那樣對其進行驗證。¹

偏見和歧視構成第二類風險。以歷史決定訓練的信貸模型，即使不直接使用受保護的特徵，也可能複製這些決定中的模式，包括對某些群體的任何不利。瑞士法對私人之間的歧視沒有一般性禁止，但《瑞士民法典》（ZGB）第 28 條對人格的保護以及一些專門法律提供了一定的保護。瑞士金融市場監管局（FINMA）的指引將平等對待列為其關注領域之一，而歐盟（EU）《人工智能法》則通過其對高風險系統數據治理的要求來處理這一問題。²可解釋性與此密切相關。如果模型無法說明它為何拒絕了一筆貸款申請或標記了一筆交易，客戶便無法對該決定提出異議，機構也無法證明其行為合法。

金融穩定是巴塞爾各機構所關注的問題。如果許多市場參與者使用以相似數據訓練的相似模型，他們的決定可能趨於相關，這可能放大市場波動，並減少維持市場流動性的觀點多樣性。Gary

風險：模型、偏差、羊群效應與欺詐

Gensler（在其出任美國證券交易委員會主席之前）與 Lily Bailey 在 2020 年的一篇論文中描述了這種羊群效應風險，認為深度學習可能通過同質化、網絡互聯以及監管空白而加劇系統性風險。³金融穩定理事會補充指出，模型、雲服務和芯片集中於少數幾家提供者，這為金融體系造成了單點故障。⁴

網絡風險和欺詐隨著這項技術一同增長。攻擊者利用語言模型以任何語言（包括瑞士德語）撰寫令人信服的網絡釣魚資訊，並找出軟件中的漏洞。語音克隆和影片深度偽造使犯罪分子能夠冒充高管、客戶或親屬。在 2024 年一起被廣泛報道的案件中，工程公司 Arup 在香港的一名員工在一次影片會議後轉出了約 2500 萬美元而會議中的其他參與者，包括表面上的首席財務官，都是由人工智能生成的。⁵國家網絡安全中心（該中心於 2024 年改組為聯邦網絡安全辦公室）報告稱，瑞士此類企圖的數量正在增加。⁶對銀行及其客戶而言，其後果是：屏幕上的聲音或面孔已不再能證明身份，因此付款指令需要通過另一獨立渠道加以核實。

1 註釋

1. 關於模型風險管理，參見美國聯邦儲備委員會和貨幣監理署的監管指引 SR 11-7，Guidance on Model Risk Management，2011 年 4 月，該指引在國際上對實踐產生了影響。有待核實。
2. 《瑞士民法典》（ZGB，SR 210）第 28 條；FINMA 第 08/2024 號指引；歐盟第 2024/1689 號條例第 10 條，關於數據和數據治理。有待核實。
3. Gary Gensler 與 Lily Bailey， “Deep Learning and Financial Stability”，工作論文，MIT，2020 年 11 月。有待核實。
4. 金融穩定理事會，The Financial Stability Implications of Artificial Intelligence，2024 年 11 月 14 日，關於第三方依賴和市場相關性的章節。有待核實。

5. 香港警方的報告以及 2024 年 2 月和 5 月國際媒體關於針對 Arup 香港辦事處之深度偽造欺詐的報道。數字有待核實。
6. 聯邦網絡安全辦公室（BACS），2024 年以來關於欺詐和深度偽造的半年度報告及每週回顧。現狀有待核實。

僅供一般參考。本文不構成法律或稅務意見，不能取代針對個別情況的專業意見，亦不建立律師與客戶關係。本文反映截至 2026 年 9 月的法律狀況。

網上版本：hk.helveticaregister.com/basel-risks-models-bias-herding-and-fraud