MinterEllison.LLP

企業人工智能系列：價值、控制與責任

探討採用企業人工智能所衍生的法律、商業和治理問題

2026 年 11 月 13 日

企業人工智能究竟需要多高程度的可解釋性？從實務角度看透明度、問責與管治

作者：蘇振國、梁嘉穎

本系列的首篇文章探討了保留對數據的權利及控制權與保留該等數據所產生的價值兩者之間的分別。第二篇文章則探討企業人工智能的價值、控制權及責任將如何分配。本文將焦點轉向可解釋性，探討其在實務上的意義。

機構正將企業人工智能應用於面向客戶的業務、日常營運及決策支援流程。董事會、管理團隊及監管機構日益關注人工智能是否具備可解釋性。然而，採用企業人工智能的機構往往並不擁有、亦無法檢視人工智能系統所採用的技術，更遑論理解其中的「黑箱」問題。所謂「黑箱」問題，是指現代人工智能系統的內部運作並不透明，即使是開發或採用該系統的人，也難以在實際操作中加以檢視。人工智能系統雖能產生結論，但沒有人能像解釋人作出的決定一樣，清楚指出得出該結論的完整推理脈絡。

機構或許無法檢視底層模型，也無法完全理解人工智能如何得出某一結論，但仍可能需要解釋該結論及調查相關事件，以回應監管機構的查詢，並維持持份者及公眾的信任。

企業人工智能往往造成「視野死角」：機構對人工智能系統如何運作的了解可能有限，卻仍須對系統的使用承擔全部責任。

因此，真正的挑戰並非能否解釋人工智能生成內容的每一個環節，而是機構是否充分掌握相關資訊，能夠負責任地管治這項技術，並對其使用持續承擔責任。

> 問題不在於機構是否完全理解人工智能模型如何運作，而在於是否充分掌握相關資訊，能夠對系統的使用持續承擔責任。

不同人士對可解釋性的理解各有不同

就本文而言，可解釋性是指系統本身或採用該系統的機構，即使在底層模型本身未必能夠被完全解讀的情況下，能夠在事後就某項預測或決策提供清晰而有條理的說明。對於人工智能系統的可解釋性，不同人士可能關注不同問題。

- 對董事會及高級管理層而言，重點通常是管治及問責。核心問題很直接：我們能否合理說明為何依賴這個系統？
- 對法律顧問及合規團隊而言，可解釋性往往涉及責任。若出現疑慮，機構能否取得所需資料，以解釋事情的經過、調查問題，並證明已作出適當監督？
- 採購團隊的考慮則有所不同。他們關注供應商會提供哪些資料、承諾及合約權利。
- 監管機構通常從問責角度審視可解釋性：機構能否證明已實施有效的管治、監察及人為監督？

可從人工智能系統供應商取得甚麼，當中又有甚麼限制？

不少機構的做法，是向人工智能系統供應商索取更多有關系統如何作出決策的資料。這種反應合情合理，卻未能解決根本問題。實際上，不論由誰提出要求，又或如何提出，可取得的資料始終有限。

機構通常可以取得系統文件、模型說明卡、管治政策、風險披露、安全認證及效能等資料。這些資料有助機構了解應如何部署及管治人工智能系統。

然而，部分資料很少能夠取得，甚至因「黑箱」問題而根本不存在。客戶一般無法取得完整的訓練數據集、專有源代碼或詳細的內部開發流程。商業保密、知識產權保障及供應鏈的複雜性，令大多數企業人工智能的應用難以做到完全透明。

因此，實際問題並非有哪些資料存在，而是當機構需要交代情況、履行其責任時，有權取得並依據哪些資料。

可解釋性的核心其實是問責

機構或許永遠無法完全理解所採用的人工智能系統的基礎模型。即便如此，機構仍可能需要解釋受該模型影響或由其產生的決策，並調查相關投訴、說明管治安排的理據、回應監管機構的查詢，以及解答客戶和持份者的疑問。

我們的目標並非做到完全可解釋，而是確保機構能夠充分承擔其責任。這表示機構須充分掌握相關資訊，以了解相關風險、監督人工智能系統的部署、監察其效能、調查事件，並在有需要時為管治決策提供支援。因此，當系統產生非預期結果時，機構應能回答以下問題：

- 發生了甚麼事？
- 為何會發生？
- 當時有哪些監控措施及監督機制？

可解釋性因此不單是技術難題，更是管治問題。關鍵不在於模型的每個部分是否均可解讀，而在於機構是否掌握足夠資料，並具備適當監督機制，以對人工智能系統的使用及其所產生的結果負責。董事會甚少需要了解人工智能模型的架構，卻必須了解自身是否設有適當的管治、監督及上報機制。要求完全可解釋或許並不切實際，但承擔採用人工智能所帶來的責任，並非遙不可及。

可解釋性如何成為管治及合約議題

從管治角度而言，機構日益需要建立監督架構、上報程序、監察安排及問責框架，以監督人工智能的運用。

香港政府轄下的數字政策辦公室最初制訂《人工智能道德框架》，原旨在為政府部門提供協助，現亦普遍供在香港規劃、設計及使用人工智能系統的機構參考。該框架以十二項人工智能道德原則為基礎，包括透明及可解釋、人為監督、公平、問責及合法。在管治架構方面，該框架建議設立三道防線：由項目團隊負責日常的人工智能應用開發和風險評估；由項目督導委員會及項目保證小組負責獨立審核及審批；再由資訊科技委員會或首席資訊官負責持續監察人工智能應用的策略及風險，並在適當情況下由外部顧問提供支援。

香港其他監管機構的取向亦相近。香港個人資料私隱專員公署發佈的《人工智能（AI）：個人資料保障模範框架》着重强調整個人工智能生命周期的責任分配、風險評估、人為監督、透明度及管治。香港金融管理局（「金管局」）就面向消費者的應用生成式人工智能發出指引，要求認可機構採用「人在環中」的方式，即在決策過程中保留人類控制權，並就人工智能生成的結論維持清晰的管治及問責架構。金管局最近發佈《負責任應用人工智能的監管指引》，並推出「ART」框架，支持負責任地使用人工智能，其中一項重點是貫穿人工智能整個生命周期的問責框架。上述文件雖不具備法律效力，但它們有可能影響監管機構對負責任人工智能管治的合理標準，以及機構應具備哪些管治措施。

在中國內地，具約束力的法規將透明度及問責轉化為具體義務。《生成式人工智能服務管理暫行辦法》（「暫行辦法」）適用於在中國內地向公眾提供的生成式人工智能服務，但不涵蓋並無向公眾提供此類服務的企業研發及應用活動。在此適用範圍內，《暫行辦法》要求採取有效措施，提高服務透明度及生成內容的準確性和可靠性，並披露服務的適用人群及投訴處理機制。重要的是，服務提供者必須配合監督檢查，按要求說明訓練數據來源、算法機制等事項，並提供必要的技術及數據支持。雖然機構無須記錄所使用的每一項資料，但其仍應備存應對監督檢查所需的資料，並作出相應合規安排，同時遵守適用的紀錄保存義務，以及有關保留用戶輸入內容及使用紀錄的限制。

歐盟《人工智能法案》提供了更詳細的人工智能系統分類。該法案將人工智能系統分為四個風險等級：（1）「不可接受風險」系統被全面禁止；（2）「高風險」系統須遵守法案最嚴格的義務，包括建立風險管理系統、數據管治監控措施、技術文件、紀錄保存（運作日誌）及人為監督安排，以及符合準確性、穩健性和網絡安全要求；（3）「有限風險」系統須履行的義務較少，例如告知用戶其正在與人工智能互動；（4）「最低風險」系統則不受法案的特定義務規限。

美國的監管情況則有所不同，亦較為分散。在聯邦層面，美國國家標準與技術研究院（National Institute of Standards and Technology）的《AI Risk Management Framework》雖屬自願採用而非強制性規範，但已成為廣泛使用的參考依據。該框架將人工智能管治分為四項功能：（1）管治（Govern）：在機構內建立風險管理框架；（2）識別風險背景（Map）：確立相關背景以界定人工智能系統的風險；（3）衡量（Measure）：運用不同工具、技術及方法，分析、評估、比較和監察人工智能風險及相關影響；（4）管理（Manage）：定期就已識別及衡量的風險分配資源，並制訂事故應對、復原及溝通計劃。美國各州亦逐漸出現類似方向。例如，《Colorado Automated Decision-Making Technology Act》將於 2027 年 1 月 1 日生效，訂明披露預期用途、已知限制及訓練數據，以及就人工智能輔助決策進行人為覆核及問責等要求。

各地監管發展的細節雖有不同，方向卻一致。機構應具備足夠的監督安排及文件紀錄，以交代人工智能系統的使用情況，而非解釋底層模型如何運作。但在實際操作中能否符合上述要求，亦取決於人工智能系統開發者與採用該系統的機構之間的合約安排。

與其追求無法實現的技術透明度，機構應考慮其是否享有適當權利，以取得負責任地管治及監督人工智能系統所需的資料。以下五個問題最為重要：

- 人工智能系統供應商必須披露哪些資料？
- 哪些資料可被獨立核實？
- 若出現問題，能否重現相關輸入內容、輸出結果及決策過程？
- 人工智能系統供應商會否協助調查、審計及回應監管查詢？
- 重要的輸出結果能否被覆核、質疑或推翻？

事件處理及調查支援可能尤其重要。當人工智能輔助決策引起疑慮時，機構通常需要解釋事情的經過、評估影響，並證明已作出適當監督。即使相關技術由第三方提供，機構仍可能須承擔這些義務。因此，機構應考慮合

約是否需要要求供應商及時配合，提供相關文件、紀錄及技術協助，並支援調查、審計及回應監管查詢。

這五個問題決定了機構在實際操作中能否持續履行其監督責任。不少機構所稱的「可解釋性」，其實是取得資料的權利、監督機制及管治監控措施的結合，讓機構充分掌握相關資訊，以了解及調查人工智能系統的使用情況，並就此作出交代。

結論

機構經常問，某個人工智能系統是否具備可解釋性。然而，這未必是最有用的問題。企業人工智能往往涉及採用系統的機構無法全面檢視或理解的技術，但機構不會僅因掌握的資訊有限，便無須承擔責任。

更實際的問題是機構是否具備足夠的透明度、監督安排及資料，能夠負責任地管治人工智能。在不少企業人工智能的應用中，機構能否承擔其監督責任，並不取決於能否全面掌握模型本身的運作，而在於其能否取得所需資料，以解釋、調查及管治這項技術的使用。機構未必需要完全了解人工智能模型如何運作，但必須充分掌握相關資訊，才能對其使用持續承擔責任。

我們可以如何提供協助

隨着機構持續採用企業人工智能，各司法管轄區對管治的要求和期望亦迅速演變。我們可以協助客戶檢視其人工智能的管治及問責框架、進行企業人工智能採購及供應商評估、審閱合約及磋商涉及透明度的條款、就人工智能管治義務向董事會及管理層提供簡介，以及制定人工智能風險管理及監督框架。

機構所需的可解釋程度取決於其目標、風險狀況及監管環境。目前的難題並非如何做到完全透明，而是如何確保即使未能完全解釋系統的運作，機構仍能承擔其監督責任。

關於作者

蘇振國是銘德有限法律責任合夥律師事務所的合夥人。他為企業提供管治、監管和風險管理相關的法律意見，包括因採用新技術而產生的問責、監督和技術風險。

梁嘉穎是銘德有限法律責任合夥律師事務所的見習律師。她尤其關注科技監管與數據管治，以及採用新技術所涉及的風險及相應的風險管理策略。



蘇振國



梁嘉穎

合夥人，香港
T: 2841 6879
E: eddy.so@minterellison.com

見習律師，香港
T: 2841 6951
E: betty.liang@minterellison.com

銘德及有關辦事處：

阿德萊德 奧克蘭 布里斯班 坎培拉 達爾文 黃金海岸 香港 倫敦 墨爾本 珀斯 悉尼 烏蘭巴托 威靈頓

本通訊由銘德有限法律責任合夥律師事務所編寫，用於略舉相關時期的某些事項，僅供參考，並不旨在提供全面內容。本通訊並非用於提供法律意見，其內容也不構成法律意見，任何人亦不應出於任何一般目的或就特定交易及/或情況依賴本通訊的內容。您應在將本通訊的資訊應用在特定情況之前尋求專業意見。如果您對本通訊有任何疑問或意見，請聯絡您在銘德有限法律責任合夥律師事務所的日常連絡人。銘德有限法律責任合夥律師事務所對於基於依賴本通訊的資訊而造成的任何損失不承擔任何責任。

© 2026 銘德有限法律責任合夥律師事務所