

如何建立人工智慧的 倫理框架

JULY 2024



企業大數據框架白皮書
Enterprise Big Data Framework White Paper



目錄



- 03、摘要
- 04、人工智慧倫理的介紹
- 05、倫理的原則與價值觀
- 06、什麼是人工智慧倫理？
- 07、治理與問責
- 08、風險評估
- 09、透明度和可解釋性
- 11、隱私和數據保護
- 12、社會和道德影響評估
- 14、持續改進和學習
- 16、TechEthics Inc. 案例研究
- 19、結論

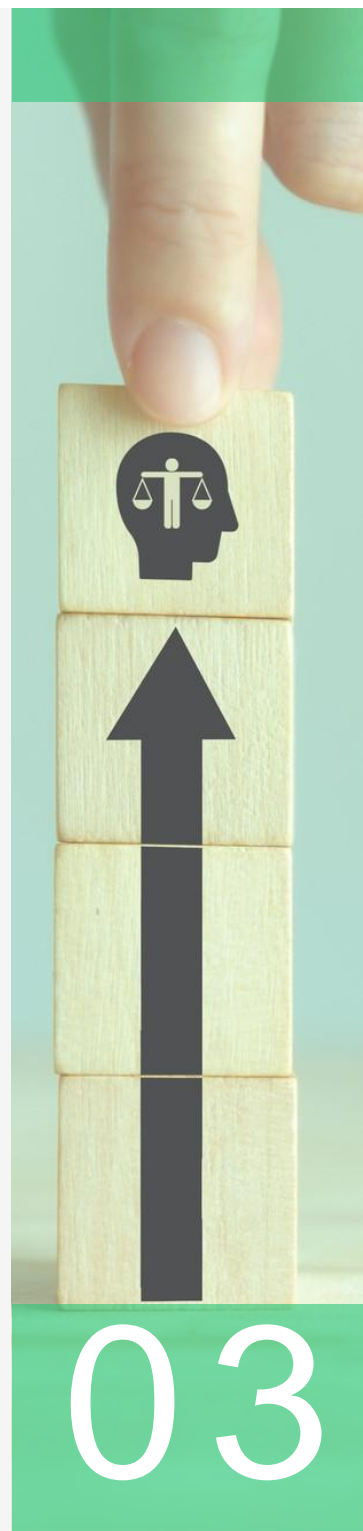


摘要

隨著人工智慧系統越來越多地融入各個領域，建立人工智慧倫理框架，對於確保負責任和道德的部署，是至關重要的行動。本白皮書提供了建立此類框架的詳細指引，重點關注七個關鍵要點。

1. 倫理原則與價值：定義公平、自主、仁慈、正義和透明等核心原則。
2. 治理與問責：實施倫理委員會、明確的政策、明確的角色和監管合規性。
3. 風險評估：識別、分析、緩解並持續監控潛在的風險。
4. 透明度和可解釋性：確保全面的文件化、可解釋的人工智慧技術和有效的溝通。
5. 隱私和數據保護：強調數據最小化、匿名化、強大的安全性和知情同意。
6. 社會和道德影響評估：讓利害關係人參與，進行情境分析，並實施緩解策略。
7. 持續改善與學習：定期檢討、回饋機制、持續培訓和外部協作。

透過遵循這些要點，組織可以使人工智慧計畫與社會價值觀保持一致，促進信任和問責制，並提高人工智慧創新的可持續性和可信度。該框架有助於降低風險並保護個人和社會，避免受到人工智慧系統潛在的不利影響。

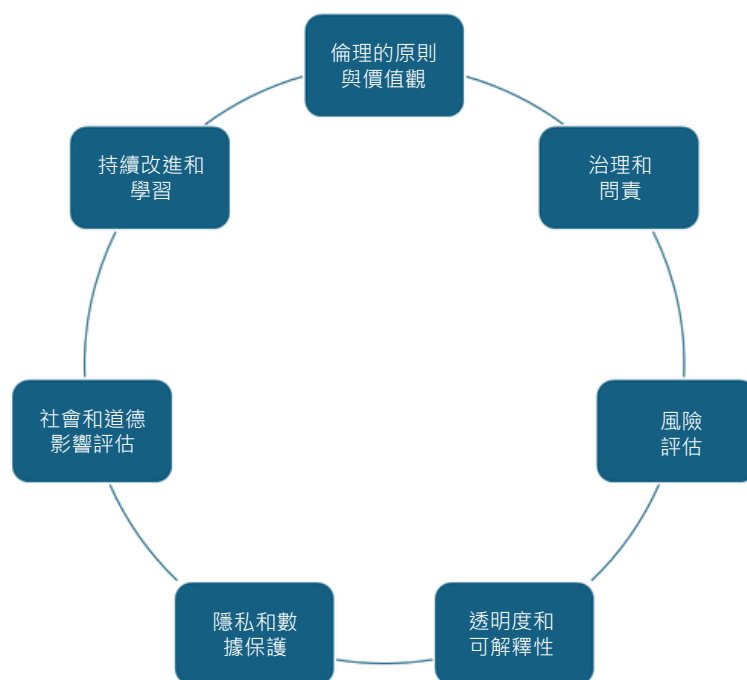


人工智慧倫理的介紹

04

隨著人工智慧 (AI) 系統越來越融入各個領域，倫理的監督變得越來越必要。雖然人工智慧有潛力徹底改變產業、提高生產力和解決複雜問題，但它也可能導致意想不到的後果，例如侵犯隱私、偏見的決策與信任受到侵蝕等。

創建人工智慧倫理框架對於應對這些挑戰，並確保負責任的人工智慧開發和部署是至關重要的工作。本白皮書提供了建立此類框架的全面指引，重點關注關鍵組成部分：倫理的原則與價值觀、治理和問責、風險評估、透明度和可解釋性、隱私和數據保護、社會和道德影響評估以及持續改進和學習。



實施這些要點，可以確保符合倫理的人工智慧開發和部署，促進信任和問責。該框架不僅可以保護個人和社會，還可以提高人工智慧創新的可信度和永續性，使組織能夠負責任的利用人工智慧的變革潛力。

倫理原則與價值

Ethical Principles and Values

任何人工智慧倫理框架的基礎都在於明確定義的道德原則和價值。這些原則充當指導人工智慧系統的開發、部署和使用的倫理指南針。

核心道德原則

公平和非歧視

人工智慧系統的設計應公平、無偏見的對待所有個人和群體。這涉及實施仔細檢查任何潛在偏差的演算法，並確保用於訓練的數據集具有代表性且沒有偏見。

尊重人類自主權

人工智慧系統應該增強而不是削弱人類的自主性。這包括尊重個人做出自己決定的權利，以及維持對人工智慧系統的人類監督。例如，醫療保健領域的人工智慧應用應該為醫生提供幫助，而不是取代他們的批判性判斷。

善行與非惡意

人工智慧應該以造福社會為目標，在積極避免傷害的同時做出積極貢獻。這項雙重任務需要嚴格的測試和評估，以確保人工智慧系統不會無意中造成傷害，無論是意外後果還是故障。

公正與問責

確保公正，包含確保人工智慧系統公平分配利益和負擔。問責意味著有機制讓個人和組織對其開發和部署的人工智慧系統負責。這包括人工智慧系統所採取的決策和行動的可追溯性。

透明和誠實

人工智慧系統的運作，應該是透明的，負責這些系統的實體，應該公開溝通它們的工作方式。這包括揭露人工智慧參與決策的過程，並向受影響的個人提供人工智慧決策之易於理解的解釋。





什麼是人工智慧倫理？



人工智慧倫理確保人工智慧技術符合公平、透明、問責和隱私保護等道德原則，同時也考慮其社會影響。它消除偏見，提高決策透明度，建立問責機制，保護隱私，維護人類自主權，並評估社會影響。

透過將道德考量納入人工智慧實踐，組織可以促進負責任的人工智慧創新，並解決人工智慧領域複雜的倫理挑戰。這種方法可以培養利害關係人之間的信任，促進人工智慧技術的採用，有助於發展更具道德意識的人工智慧生態系統，並最終為社會帶來更永續和有益的人工智慧應用。

治理與問責

Governance and Accountability

有效的治理結構和問責機制，對於實施和維護人工智慧倫理框架，是至關重要的。治理確保道德原則融入人工智慧生命週期的每個階段，從設計和開發到部署和退役等。

關鍵的治理要項

倫理委員會和董事會

建立由不同利害關係人（包括倫理學家、技術專家和各團體的代表等）組成的專門倫理委員會或諮詢委員會。這些機構負責監督人工智慧專案的倫理層面，並就複雜的道德問題提供指導。

政策和程序

制定和執行全面的政策和程序，將倫理考量，納入人工智慧相關活動中。

這包括數據收集、演算法設計、測試協議和部署實踐的指導。

角色和責任

明確定義組織內的角色和職責，以確保倫理實踐的責任。這包括任命人工智慧倫理官員或團隊，負責監控合規性和解決道德問題。

監理合規性

確保人工智慧系統符合相關法律法規和行業標準。這包括遵守 GDPR 等數據保護法規，並遵循 IEEE 或 ISO 等機構制定的人工智慧道德準則。

整合這些治理要素，使組織能夠維護道德標準、提高透明度並建立對人工智慧技術的信任，從而培育負責任和永續的人工智慧生態系統。

案例 - TechEthics Inc.

TechEthics Inc. 是醫療保健人工智慧技術公司，實施以下治理和問責措施：

1. 倫理委員會：倫理諮詢委員會指導人工智慧計畫。
2. 政策和程序：綜合指引，確保道德實踐。
3. 角色和職責：人工智慧倫理官員負責監督實施情況。
4. 監管合規性：遵守 HIPAA 和行業標準，可確保數據保護和道德規範。

該框架促進透明度、問責制和信任，支持醫療保健領域負責任的人工智慧開發。

風險評估

Risk Assessment



進行徹底的風險評估，對於識別和減輕與人工智慧系統相關的潛在倫理和營運風險至關重要。強大的風險評估流程有助於防止不良後果並建立對人工智慧技術的信任。

風險評估流程

風險識別

識別與人工智慧相關的潛在風險，例如偏見、隱私洩露、安全漏洞和社會影響。這涉及與不同的利害關係人合作，以發現廣泛的潛在問題。

風險分析

使用定性和定量方法評估已識別風險的可能性和影響。此步驟包括情境規劃和敏感度分析，以了解不同因素如何影響結果。

風險緩解

制定和實施策略，減輕已識別的風險。這可能涉及演算法審查和公平干預等技術措施，以及道德培訓和包容性設計流程等組織實踐。

持續監控

實施持續監控機制，以發現和應對新出現的風險。這包括針對人工智慧系統中的異常行為設定自動警報，並根據道德標準定期審查人工智慧效能。

透過遵循穩健的風險評估流程，組織可以主動管理風險、維護道德標準並提高人工智慧技術的可靠性和可信度。這種方法促進負責任的人工智慧部署，並有助於在利害關係人和更廣泛的組織之間建立信任。



透明度和可解釋性

Transparency and Explainability



確保人工智慧系統的透明度和可解釋性，對於建立信任和問責制是至關重要的。包括使用者、監管機構和大眾在內的利害關係人，都應該清楚的了解人工智慧系統如何運作和做出決策。透明且可解釋的人工智慧系統，對於獲得大眾信任、促進監管合規，以及創造負責任和有效使用人工智慧技術的環境，是至關重要的。

透明度的策略

文件和報告

維護全面且易於存取的文件，對於透明度至關重要。這包括人工智慧系統設計、開發流程和決策機制的詳細記錄。記錄訓練數據、任何演算法變化，以及決策背後的基本原理，有助於審核和理解人工智慧系統。例如，組織應保留數據來源、預處理方法和模型選擇標準的日誌。可以與利害關係人分享與總結相關的定期更新與報告，以便讓他們了解情況。

可解釋的人工智慧 (Explainable AI, XAI) 技術

實施可解釋的人工智慧技術，對於使人工智慧決策可解釋和理解性是至關重要。局部可解釋性模型 (Local Interpretable Model-agnostic Explanations, LIME) 和 SHAP (SHapley Additive exPlanations) 等技術，透過解釋單一預測來提供複雜模型的見解。這些技術幫助使用者了解影響人工智慧決策的因素，這在醫療保健、金融和刑事司法等高風險領域尤其重要。例如，在醫療診斷人工智慧系統中，XAI 可以幫助臨床醫生理解為什麼建議某種診斷，從而使他們能夠做出更明智的決策。

利害關係人溝通

與利害關係人的有效溝通，對於透明度至關重要。組織應主動向所有相關方傳達人工智慧系統的功能、限制和決策流程。這可能涉及建立用戶友好的介面，使用圖表、圖形和其他視覺輔助工具，直觀的顯示決策的製定方式。此外，組織可以定期發布透明度報告，詳細說明人工智慧系統的效能，包括其優點、缺點和任何已識別的偏差。公開分享這些資訊，有助於建立信任，並允許利害關係人提供回饋，從而促進人工智慧開發的協作方法。



監理合規性

遵守監管要求，是透明度和可解釋性的關鍵作為。許多司法管轄區都有具體法規，要求人工智慧系統的透明度，特別是關鍵部門使用的系統。遵守歐盟一般數據保護規範（GDPR）或擬議的人工智慧法案等法規，可確保人工智慧系統的設計和運作透明。組織必須隨時了解監管變化，並將合規性納入其人工智慧開發流程。這不僅可以降低法律風險，也顯示了對倫理的人工智慧實踐之承諾。

使用者教育和培訓

對使用者進行人工智慧系統及其功能的教育訓練，是確保透明度的另一個重要策略。提供培訓課程、研討會和教育材料，可以幫助使用者了解人工智慧系統如何運作，以及如何解釋其產出。這對於依賴人工智慧系統進行決策的專業人士（例如醫療保健提供者、金融分析師和法律從業人員）尤其重要。為使用者提供有關人工智慧的知識，可以增強他們使用這些系統的信心，並增強他們批判性評估人工智慧驅動建議的能力。

協同開發

讓不同的利害關係人群體參與人工智慧開發過程，可以提高透明度和可解釋性。透過讓倫理學家、領域專家和組織代表參與進來，組織可以確保考慮不同的觀點，並在開發週期的早期，就發現潛在的問題。協作開發可以培養共同責任感，有助於創造更符合社會價值和道德標準的人工智慧系統。

持續改進

透明度和可解釋性不是一次性的努力，而是持續的承諾。組織應持續監控其人工智慧系統，並根據新的見解和回饋更新其透明度實踐。定期審核、效能評估以及文件和 XAI 技術的更新可確保 AI 系統隨著時間的推移保持透明且易於理解。這種持續改進週期有助於組織適應新挑戰並維持利害關係人的信任。

透過實施這些策略，組織可以確保其人工智慧系統透明且可解釋，從而促進信任和問責制。透明的人工智慧實踐不僅可以增強利害關係人的信心，還有助於以負責任和道德的方式部署人工智慧技術，最終造福整個社會。

隱私和數據保護

Privacy and Data Protection



隱私和數據保護是人工智慧倫理框架的重要組成部分。保護個人數據並確保遵守數據保護法對於維持信任和維護權利至關重要。

隱私和數據保護措施

數據最小化

將數據收集限制在特定人工智慧應用所需的範圍內。這項原則降低了隱私外洩和數據濫用的風險。例如，使用差異隱私等技術可以幫助聚合數據見解，而無需暴露單一數據點。

匿名化與去識別化

採用科技對個人數據進行匿名化和去識別化，以保護個人身分。這涉及刪除或混淆可識別資訊，以便無法從數據集內輕鬆識別個人。

數據安全

實施強大的數據安全措施，防止未經授權的存取和外洩。這包括加密、存取控制和定期安全審核，以保護靜態和傳輸中的數據。

同意和控制

確保個人對數據收集和使用，提供知情同意，並控制其數據。這涉及到關於正在收集哪些數據、如何使用數據，以及為個人提供選擇，是否退出或管理其數據偏好的選項，進行明確的溝通。

總之，透過實施這些隱私和數據保護措施，組織可以與利害關係人建立和維持信任，維護倫理標準並保護個人權利。這種對隱私和數據保護的承諾不僅增強了人工智慧系統的完整性，而且支持其負責任和道德的部署。



社會和道德影響評估

Social and Ethical Impact Assessment

評估人工智慧系統的社會和道德影響，對於識別潛在的意外後果，並確保與社會價值觀保持一致，至關重要。此過程可協助組織預測並解決其人工智慧計畫的更廣泛影響，促進負責任和道德的人工智慧部署。

影響評估流程

利害關係人參與

讓包括受影響的組織在內的不同利害關係人參與評估過程，以獲取廣泛的觀點。這種參與可以採取大眾諮詢、焦點小組和諮詢小組的形式。透過整合不同利害關係人的觀點，組織可以全面了解潛在影響，並透過透明和包容性的流程培養信任。

情境分析

評估潛在場景，藉以了解人工智慧部署的社會和道德影響。這涉及探索最佳情況和最壞情況，以及考慮對不同群體的長期影響。情境分析有助於識別風險和機遇，採取積極主動的措施來減輕不利影響，並最大限度地提高正面成果。

倫理的審核

定期進行審核，以評估人工智慧系統的倫理的表現。這些審核應審查道德準則的遵守情況，評估緩解策略的有效性，並確定需要改進的領域。倫理的審核提供了一種系統方法，以確保人工智慧系統遵守倫理標準，並幫助組織維持責任。



緩解策略

制定和實施策略，用以解決已確定的社會和倫理影響。這可能涉及修改演算法以減少偏見、加強用戶對人工智慧的教育，或投資措施以抵消負面影響。有效的緩解策略是為了解決特定的道德問題而客製化的，對於維護人工智慧系統的完整性至關重要。

長期監控與回饋

建立人工智慧系統的長期監控機制，以持續評估其社會和倫理影響。這包括建立回饋循環，使用者和其他利害關係人可以報告問題或疑慮，並將這些回饋整合到正在進行的系統改進中。持續監控可確保人工智慧系統負責任的發展，並隨著時間的推移與社會價值觀保持一致。

跨學科合作

促進倫理學家、社會科學家、技術專家和政策制定者之間的合作，以解決複雜的倫理問題。跨學科方法是透過引入不同的專業知識和觀點來豐富影響評估過程，從而實現更全面、更穩健的倫理評估。

將這些要件整合到社會和倫理影響評估流程中，使組織能夠更好的預測和解決，相關人工智慧計畫之各項廣泛的影響。這種全面的方法，確保人工智慧系統的發展和部署方式符合倫理規範、對社會負責，並符合其所服務的社會的價值觀和期望。



“透過跨學科合作，
引入不同的專業知識和觀點，
豐富且影響評估的過程。”

持續改進和學習

Continuous Improvement and Learning

對持續改進和學習的承諾，可以確保人工智慧倫理框架在快速發展的技術環境中，保持相關性和有效性。這涉及根據新的見解和進步定期更新的實踐，從而培育一種適應性強、有彈性的人工智慧倫理的部署方法。

持續改進的策略

回饋機制

為利害關係人建立管道，提供有關人工智慧系統及其道德表現的回饋。這可能包括調查、回饋表以及與倫理官員的直接溝通。有效的回饋機制使組織能夠及時識別和解決倫理問題，從而促進信任和透明度。

定期審查和更新

定期審查和更新人工智慧倫理框架，以納入新的見解、技術進步和監管變化。這確保了該框架保持最新且有效。定期審查有助於組織與最佳實踐和新興標準保持一致，從而降低道德失誤的風險。

培訓和教育

為員工提供有關人工智慧倫理和負責任的人工智慧實踐的持續培訓和教育。這包括研討會和電子學習課程，讓員工了解最佳實踐和新出現的問題。持續教育使員工能夠做出符合倫理的決策，並促進責任和問責文化。

協作與知識分享

與外部合作夥伴、產業團體和學術界進行協作和知識共享。這培育了持續學習的文化，並幫助組織保持在人工智慧倫理開發的最前沿。協作努力可以促進創新解決方案和共享標準的開發，從而增強整體產業實踐。

道德標竿

將組織的人工智慧倫理框架和績效，與行業基準和最佳實踐進行比較。這涉及分析其他領先組織如何解決道德問題，並確定需要改進的領域。基準測試幫助組織不僅滿足而且超越行業標準，可以推動持續的倫理進步。



技術觀察和廣泛關注

隨時了解新技術、方法以及可能出現的潛在倫理挑戰。廣泛的關注是涉及展望未來，為可能影響倫理實踐的人工智慧未來發展，做好準備。隨時了解新興趨勢，可以確保組織能夠主動調整其倫理框架以應對新形態的挑戰。

影響指標和報告

開發和使用指標來衡量人工智慧倫理倡議的影響。定期向利害關係人報告這些指標，可以確保透明度和問責制。影響力指標提供有關道德實踐有效性的具體數據，突顯成功和需要改進的領域。

將這些策略納入其持續改進流程，使組織能夠確保自身的人工智慧倫理框架維持穩健與強壯的適應性，並符合社會期望。這種對倫理卓越的持續承諾，不僅可以降低風險，還可以提高人工智慧技術的可信度和可持續性，有助於其負責任與有益的使用。

持續改進和學習

TechEthics Inc. 案例研究

TechEthics Inc. 是一家領先的醫療保健人工智慧技術公司，致力於確保自身人工智慧系統遵循最高的倫理標準。為了實現這一目標，該公司在自身的人工智慧道倫理架內實施了全面的持續改進和學習的策略。本案例研究，探討了 TechEthics Inc. 如何應用這些策略，在快速發展的技術環境中，保持其倫理實踐的相關性和有效性。

背景

TechEthics Inc. 專注於開發人工智慧驅動的診斷工具，幫助醫療保健提供者做出準確、及時的醫療決策。該公司的人工智慧系統，分析大量醫療數據，提供可改善患者治療結果的見解。鑑於醫療保健數據的敏感性以及醫療決策的關鍵影響，倫理考量對於 TechEthics Inc. 至關重要。

TechEthics Inc. 的持續改善策略

回饋機制

TechEthics Inc. 已經為利害關係人（包括醫療保健提供者、患者和內部員工）建立了多種管道，以提供有關人工智慧系統的回饋。這些管道包括：線上調查、整合到人工智慧系統介面中的回饋表，以及與倫理諮詢委員會的定期會議等。例如，在部署新的診斷工具後，該公司收集醫生和患者對其準確性、可用性和感知道德問題的回饋。然後分析該回饋，以確定需要改進的領域。

定期檢討和更新

該公司每季對自身的人工智慧倫理框架進行審查，納入最新研究、技術進步和監管要求變化的新見解。在這些審查過程中，多學科團隊會評估框架的有效性，並確定必要的更新。例如，為了回應新的 GDPR 準則，TechEthics Inc. 更新了自身的數據保護政策，並對員工進行了合規性要求的再培訓。

培訓和教育

TechEthics Inc. 為其員工提供持續的培訓計畫，重點是人工智慧倫理和負責任的人工智慧實踐。這些計畫，包括有關偏見檢測和緩解的研討會、有關人工智慧最新道德挑戰的研討會，以及有關監管合規性的電子學習課程。最近，該公司推出了關於可解釋人工智慧 (XAI) 技術的新培訓課程，以幫助員工理解和實施，使人工智慧決策更加透明的方法。

協作與知識共享

公司積極與大學、研究機構和產業團體合作，始終走在人工智慧發展的前沿。TechEthics Inc. 參與聯合研究計畫、出席會議並為業界標準做出貢獻。例如，該公司最近與一所領先大學合作，研究人工智慧在個人化醫療中的倫理影響，分享有助於改善其自身人工智慧系統的見解。

倫理的標竿

TechEthics Inc. 定期將自身的人工智慧倫理實踐與其他領先的醫療保健技術公司的實踐，進行比較。透過將自身的方法與行業最佳實踐進行比較，該公司確定了優勢和需要改進的領域。例如，基準測試顯示 TechEthics Inc. 可以增強其數據收集的同意流程，從而顧及到更完整的知情同意權的實踐。

技術觀察和廣泛的關注

該公司擁有一支專門的團隊來監控新興技術、方法和潛在的倫理挑戰。該團隊進行廣泛的關注，以預測可能影響倫理實踐的人工智慧未來發展。最近，該團隊發現了基因組數據分析中，使用人工智慧相關的潛在道德問題，促使 TechEthics Inc. 在其倫理框架中主動解決這些問題。

與監管機構的合作

TechEthics Inc. 與 FDA 和歐洲藥品管理局等監管機構保持開放的溝通管道。公司隨時了解不斷變化的法律要求和道德準則，確保合規性和道德健全性。例如，透過與監管機構的定期互動，TechEthics Inc. 預見了新的人工智慧透明度要求，並相應的調整了自身的做法。

影響指標和報告

為了衡量人工智慧道德措施的影響，TechEthics Inc. 制定了一套指標，包括道德問題報告的頻率、使用者對人工智慧決策的滿意度，以及倫理準則的遵守率。該公司發布年度倫理報告，為利害關係人提供有關自身倫理道德實踐表現的見解。這種透明度促進了信任和問責制。

結論

透過整合這些持續改進策略，TechEthics Inc. 確保自身人工智慧倫理框架的穩健、適應性強並符合社會期望。該公司對倫理卓越的持續承諾，可降低風險並提高自身人工智慧技術的可信度和可持續性。透過與利害關係人的積極參與、持續學習和定期更新，TechEthics Inc. 為醫療保健領域的人工智慧倫理的部署，制定了高標準。

結論

Conclusion

19

對於尋求以負責任和道德的方式部署人工智慧系統的組織來說，建立人工智慧倫理框架是至關重要的一步。透過建立明確的倫理原則和價值觀、健全的治理結構、徹底的風險評估流程、透明的實踐、嚴格的隱私和數據保護措施、全面的社會和道德影響評估以及持續改進的承諾，組織可以確保自身人工智慧系統與社會發展保持一致。

這個框架不僅保護個人和社會，也增強了人工智慧創新的可信度和永續性。透過讓不同的利害關係人參與、進行情境分析並實施緩解策略，組織可以主動解決潛在的道德和社會影響。定期審查和更新、持續培訓和教育，以及與外部合作夥伴的合作，進一步加強了框架，確保自身在快速發展的技術環境中，保持相關性和有效性。

優先考慮人工智慧倫理的組織，能夠更好的利用人工智慧的變革潛力，同時降低風險，並維護大眾信任。透過這樣做，他們可以負責任的推動創新，在遵守道德標準的同時，保持競爭優勢。最終，人工智慧倫理框架不僅是監管的必需品，而且是促進永續和值得信賴的人工智慧發展的戰略資產，使組織和更廣泛的社會，可以受益。

“人工智慧倫理
框架至關重要，
它可以保護社會、
提高人工智慧的可
信度，並幫助
組織負責任的利
用人工智慧。”

有關於 EBDFA

關於企業大數據框架聯盟

Enterprise Big Data Framework Alliance,
EBDFA

透過企業大數據框架，我們向成千上萬的人和組織提供有關數據、分析、機器學習和人工智慧的知識。

企業大數據框架聯盟，透過會員資格建立一個從業者的社區，透過獨立於供應商的認證計畫，來推廣知識，並舉辦學習和啟發社區的活動。

免責聲明

EBDFA 設計並建立了「如何建立人工智慧倫理框架」（「工作」），主要作為專業人士的教育資源。EBDFA 並不聲稱使用任何作品將確保將會成功的結果。

此「工作」不應被視為已經包含所有適當的資訊、程序和測試，或排除針對獲得相同結果的其他資訊、程序和測試的合理性。在確定任何特定資訊、程序或測試的適當性時，專業人員應根據特定系統或資訊科技環境所呈現的特定情況，運用自己的專業判斷。



Endenicher Allee 12
53115, DE Bonn
Germany

W: www.bigdataframework.org

E: info@bigdataframework.org

人工智慧國際認證，請參考：
<https://www.learnmydata.com/AIC.html>

Provide Feedback:

communities.kajabi.com/ebdfacommunity

Join our Communities:

LinkedIn:

www.linkedin.com/company/big-data-framework

YouTube:

www.youtube.com/@bigdataframework

Twitter (X):

twitter.com/bdframework

