TELEHEALTH AND MEDICINE TODAY

討論/評論/觀點

可解釋的人工智能和評估框架在醫療保健安全有效整合大型語言模型中的作用

Sandeep Reddy, MBBS, MSc, PhD[1] ; Alexandre Lebrun, MS, MSEE[2] ; Adam Chee, PhD[3] ; and Dimitrios Kalogeropoulos, PhD[4,5]

[1]澳洲維多利亞州迪肯醫學院院長；[2]法國巴黎 Nabla 共同創辦人兼執行長；[(3) Saw]Swee Hock 副教授。

[(3)]新加坡國立大學 Saw Swee Hock 公共衛生學院副教授，新加坡；[4]全球健康與數位創新基金會創辦人兼執行長，英國倫敦；[5]UCL 全球健康商學院，英國倫敦。

通訊作者：Sandeep Reddy；電子郵件： sandeep.reddy@deakin.edu.au DOI:

https://doi.org/10.30953/thmt.v9.485

Keywords：AI、人工智慧、Converge2Xcelerate、醫療保健、大型語言模型、LLM

提交：2024 年 2 月 8 日；接受：2024 年 4 月 2 日；發表：2024 年 4 月 30 日

整合人工智慧 (AI)，特別是大型語言模型 (LLM：一種在大量文字上訓練的專門 AI)。以理解現有內容並產生原始內容）與醫療保健的整合持續加速，因此需要周詳的評估與監督，以確保安全、合乎道德且有效的部署。在由本刊出版商 Partners in Digital Health 舉辦的 2023 Converge2Xcelerate 活動中，人工智慧專家針對將 LLM 等創新技術深思熟慮地轉化為醫療照護的機遇與挑戰，進行了專題討論，並分享了重要觀點[(1)]。

涵蓋的主要主題：

- 可解釋的人工智能在促進透明度和信任方面的潛力；
- 將人工智慧與多變的全球醫療照護協議相結合的挑戰；
- 針對醫療照護環境，透過轉譯與治理架構進行評估的重要性；
- 對於會話介面過度擴大使用 LLMs 持懷疑態度；
- 在考慮風險等級時，需要明智地驗證 LLM。

此外，討論還強調了可解釋性（即機器學習模型及其輸出可被解釋且「有道理」的概念）、評估，以及與醫療照護專業人員的慎重商議，這些都是在醫療領域更廣泛採用 AI 的同時，實現效益並積極應對風險的關鍵。其他討論主題集中在醫療領域人工智慧的關鍵評估實務、可解釋人工智慧的必要性與限制，以及醫療領導者在部署前必須進行的討論。

[2]儘管人工智能有望增強人類的能力，並改善醫療服務的可及性、品質和效率，但人工智能的整合除了對患者和醫療專業人員的責任和影響外，還帶來了複雜的技術、倫理和法規方面的考慮，包括透明度[3,4]。

可解釋的人工智能：建立信任與透明度

討論的焦點之一是可解釋的 AI 技術在建立對 AI 系統的信心方面所扮演的角色。與會者一致認為，雖然 LLM 可以提供有用的輸出，但其本質上是一種「黑箱」功能，掩蓋了結論背後的推理。在高風險的醫療照護環境中，這種不透明性變得令人擔憂。

與會者強調，可解釋的人工智慧是解決這些透明度問題的積極研究領域。

[5]然而，利用大型神經網路實現全面的可解釋性仍然充滿挑戰。臨床環境的複雜性、多變性和動態性可能會進一步混淆針對更受限制的設置而調整的解釋方法[6]。

評估：確保有效且合乎道德的轉譯

這些觀察突顯了在人工智慧系統的設計、部署與運作過程中，針對醫療照護應用進行嚴格評估的必要性。與會專家提倡廣泛、持續的評估架構，例如 TEHAI (Translational Evaluation of Healthcare AI)[7]，涵蓋與醫療照護 IT 系統的整合、臨床採用因素、最新的效能監控，以及治理元件[8]，例如責任與道德。專題討論小組成員也注意到與單獨針對安全性和傷害的狹隘監管評估的區別。多層面的評估可以說明優點、缺點和適當的使用個案，以負責任的方式引導人工智慧的採用。

自動化照護流程，而非對話

在小組討論中提出的另一個考量是，人們越來越傾向於將人工智慧系統人格化，尤其是會話介面。與會者提醒，過度人格化會導致對於完整性與效能的不切實際期望。是的，LLM 可以幫助關鍵照護流程自動化，但目前完全自動化的診斷或治療仍不可靠。

此外，與會專家建議將自動化的重點放在增強臨床醫師的文件記錄、編碼與照護協調支援。人工智慧的解釋有助於為其能力和限制建立適當的心智模型，而非使其人性化[9]。

謹慎進步

總括而言，與會者強調謹慎、共同合作[8]與評估的方法，是在臨床實務中積極應用人工智慧進展的必要條件。與其取而代之，不如強調從支援性、低風險的角色著手，並透過實際部署來評估其好處與缺點[7]。

持續的進步有賴於跨領域領導者進行開放且誠實的評估，以便在實際醫療服務中為 LLM 等新興創新技術找到最具建構性的利基。可解釋的評估架構可促進透明度，在負責任地推進人工智慧整合方面扮演不可或缺的角色。

經費

手稿編寫過程中未使用任何外部經費。

財務及非財務關係與活動

Sandeep Reddy 是 Healea Services 和 Centre for Advancement of Translational AI in Medicine 的主席和董事。

貢獻者

Sandeep Reddy 負責手稿的起草和定稿。所有作者對手稿內容皆有貢獻。

資料可用性聲明 (DAS)、資料分享、可重複性及資料庫

本文編寫過程中未使用任何原始資料。

應用人工智能產生的文字或相關技術

本文使用 AI 技術製作小組討論的文字記錄，並協助撰寫文稿的文法。

鳴謝

作者感謝 Partners in Digital Health 舉辦此次活動和小組討論。

參考文獻

1. Telehealth and Medicine Today.ConV2X 座談會 [網際網路]。[於 2024 年 2 月 1 日引用]。網址：https://telehealth.conv2xsym-posium.com
2. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR.Artificial Intelligence (AI) and global health: How can AI con-tribute to health in resource-poor settings?BMJ Glob.Health.2018;3:e000798. https://doi.org/10.1136/bmjgh-2018-000798
3. Char DS, Shah NH, Magnus D. Implementing machine learn-ing in health care-addressing ethical challenges.N Engl J Med.2018;378:981–3. https://doi.org/10.1056/NEJMp1714229
4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. 利用人工智能提供臨床影響力的關鍵挑戰。BMC Med.https://doi.org/10.1186/ S12916-019-1426-2
5. Adadi A, Berrada M. Peeking inside the black-box: a sur-vey on explainable artificial intelligence (XAI).IEEE Access.2018;6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052
6. Rudin C. Stop explaining black box machine learning mod-els for high stakes decisions and use interpretable models instead.Nat Mach Intell.2019;1:206-15. https://doi.org/10.1038/ s42256-019-0048-x
7. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings.BMJ Health Care Inform.2021;28(1):e100444. https://doi.org/10.1136/bmjhci-2021-100444

8. Kalogeropoulos D, Barach P. Telehealth's role enabling sustain-able innovation and circular economies in health.Telehealth Med Today.2023;8(1). https://doi.org/10.30953/thmt.v8.409
9. Miller T. 人工智能中的解釋：來自社會科學的見解。Artif Intell.2019;267:1-38. https://doi. org/10.1016/j.artint.2018.07.007