「知情同意」視角下人工智能用於醫療診斷的透明度問題

The Problem of the Non-transparency of Artificial Intelligence Technologies in Medical Diagnosis: A View Based on Informed Consent

徐漢輝

Xu Hanhui

徐漢輝，南開大學醫學院醫學倫理專業講師，中國天津，郵編：300102。
Xu Hanhui, Lecturer in Medical Ethics, School of Medicine, Nankai University, Tianjin, China, 300102.

《中外醫學哲學》XVII:1 (2019 年)：頁 49-58。
International Journal of Chinese & Comparative Philosophy of Medicine XVII:1 (2019), pp. 49-58.

摘要 Abstract

隨著人工智能技術在醫療診斷中得到了越來越廣泛的應用，對於其“不透明性”的擔心也日益加重。這種擔心來自人們對人工智能系統的工作機制尚不清楚。在我們還無法知道其內部工作原理的情況下，根據它的診斷對患者進行治療是否可行？本文嘗試從患者知情同意權的角度去論證，應用於醫療診斷的人工智能系統應該更加透明，以避免對患者造成可能的傷害。

Since the introduction of artificial intelligence technologies in medical diagnosis, ethical issues have emerged. One of these concerns is the “black box,” which can only be seen in terms of inputs and outputs, with no way to understand the AI algorithm. This is problematic because patients, physicians, and even designers do not understand why or how a treatment recommendation is produced by AI technologies. In this paper, I argue that AI technologies should be explained on the grounds that patients have a right to informed consent.

【關鍵字】人工智能 醫療診斷 不透明 知情同意

Keywords: Artificial Intelligence, Medical Diagnosis, Non-transparency, Informed Consent

近年來，人工智能技術（Artificial Intelligence）在醫療診斷中得到了越來越廣泛的應用。2015 年，美國的研究團隊在紐約的西奈山醫院（Mount Sinai Hospital）研發了一套人工智能系統，將其命名為“深度患者（Deep Patient）”。該系統通過對已有的 70 萬多份電子病例進行深度學習，最終能夠準確地預測出新病例中“隱藏”的疾病。不僅如此，“深度患者”系統對於諸如精神分裂症等精神疾病有著異乎尋常的精準預測。而對醫生來說，精神類疾病一直以來都是最難診斷的。在乳腺癌的預測和診斷方

面，人工智能也有著不俗的表現。人工智能系統在大量學習乳腺癌患者的乳房 X 光片之後，能夠對新的乳房 X 光片做出預測，預報就診者是否有得乳腺癌的風險。

人工智能技術在精準預測疾病方面的出色表現並沒有完全打消人們的擔憂。這些擔憂中常被提及的便是人工智能技術用於臨床診斷的“透明度”問題。人工智能的核心是“機器學習（Machine Learning）/深度學習（Deep Learning）”算法，該算法基於人工神經網路（Artificial Neural Network）系統，能夠對輸入的資料進行“自主式”的學習，從而輸出精準的結果，如上文中提到的對疾病的預測。相比之前，基於深度學習演算法的人工智能技術類比人腦神經網路結構，由神經元、層和網路三部分組成。

然而，正像人類尚未弄清人腦的工作原理一樣，應用於醫療診斷的人工智能系統對於患者、醫生甚至程式設計者來說都是不折不扣的“黑箱（Black Box）”。所謂“黑箱”是指，人們對人工智能系統內部的工作機制並不了解。如上文中提到的深度患者系統，研究人員只是知道通過大量的深度學習，該系統可以準確地對疾病進行預測。但是，為什麼能夠達到如此精準的預測以及這一系統是如何得出診斷結論的？這些問題始終困擾著研究人員。由此產生了對於診斷結論“透明度”的擔心，即如果我們無法知道人工智能系統的工作機制，那麼根據它的診斷對患者進行治療是否可行？基於此，一些學者認為，考慮到“黑箱”的不透明性，我們應該完全地或者至少部分地弄清人工智能系統內部的工作原理以增強其可靠性，而不是盲目地信任人工智能系統給出的結論。與之爭鋒相對的一種觀點則認為，“黑箱”不是問題，儘管我們無法了解人工智能系統的工作機制，但其預測結果的精準性已得到了充分驗證。極高的預測精確度足以保障其可靠性。因此，對“黑箱”透明度的擔心並無必要。本文嘗試從患者知情同意的角度論證，患者的知情權要求應用於醫療診斷的人工智能系統更加透明。特別是當人工智能系統給出的結論與醫生的診斷

相矛盾的時候，或者兩套人工智能系統給出不同的結論的時候，醫生需要知道這其中的原因以保障患者的知情權，進而避免對患者造成的不必要的傷害。

一、人工智能用於診斷的不透明性

如上所述，由於人們對人工智能系統的工作機制並不了解，人工智能技術應用於醫療診斷中的“不透明”問題引發了很大的爭議。首先，人工智能系統對於疾病預測的準確度無法達到百分之百。上文中提到的應用於醫療診斷中的人工智能系統在對疾病的預測上，儘管有著高於人類醫生的準確性，但是仍然無法達到完美預測的程度。例如，谷歌（google）人工智能系統在乳腺癌的追蹤和預測上能夠達到 99%的準確率，但仍有百分之一誤判的可能性。不僅如此，一些人工智能系統在對疾病的診斷上出現了違背常識的“錯誤”。例如，李奇·卡魯納（Rich Caruana）教授的研究團隊發現，一個人工智能系統在預測肺炎死亡率方面表現出色，但卻將哮喘患者死於肺炎的可能性排在了正常人群之下。很明顯，這一診斷結論與常識相悖。後經研究發現，該系統之所以得出這一結論是由於哮喘患者會被第一時間送入 ICU 病房救治，使得哮喘患者肺炎死亡率低於正常人群。由於人工智能系統診斷正確率無法達到百分之百，又出現了一些在人類眼中“難以理解”的反常識性“錯誤”。這引發了公眾對於人工智能系統可靠性和安全性的擔憂，而其工作機制的不透明進一步加劇了這種擔憂。要求人工智能系統增強透明性，以便醫生或者程式設計者能夠理解或者“讀懂”這些系統給出的診斷的呼聲日益高漲。

然而，並不是所有人都認為，人工智能系統的不透明問題值得擔心。一些學者聲稱，人工智能系統在預測疾病上表現出了超出人類醫生的高準確性，這種高精確度足以保障其可靠性。那些對人工智能系統可靠性和安全性的擔憂實屬多慮，要知道在很多時候，醫生也是憑著他們無法清楚解釋的直覺和經驗在為病人做

診斷。如果我們能夠接受醫生給出的“難以解釋”的診斷，那麼我們也不應該排斥人工智能系統給出的結論。如韋德·潘傑（Vijay Pande）表示：“人類智能本身就是，而且一直總是，一個黑箱。例如，當一個人類醫生做診斷的時候，患者可能會問這位醫生，她是如何給出這個診斷結論的？這位醫生大概會告訴患者她是如何從已有的資料中得出的結論。然而，這位醫生真的可以清楚地解釋出她是從什麼樣的研究中怎樣得到支撐她結論的資料嗎？當然，她一定會給出自己做診斷的理由，但這其中難免有猜測或者憑藉直覺的成分。”針對這種觀點，筆者嘗試通過一個思想實驗來論證，即使都具有一定程度的不可解釋性，人類醫生給出的診斷和人工智能系統給出的結論卻有可能有著本質的差異。正是這種可能性讓我們對於人工智能系統不透明性的擔憂具有合理性。

假設有四種疾病，分別是疾病甲、疾病乙、疾病丙和疾病丁。就疾病甲而言，幾天的休息就能痊癒。疾病乙則需要服用止痛片外加充足睡眠。疾病丙的治療需要使用抗生素才能治癒。而疾病丁則必須截肢。假設當具有特殊症狀的病人前來就診的時候，人類醫生張聖手通常會首先排除疾病丙和疾病丁，而後在疾病甲和疾病乙中選出判斷。其實，根據病人的症狀，張聖手並不十分清楚為什麼是疾病甲而不是疾病乙。但是根據他多年行醫的經驗和直覺，大多數情況下，他會建議病人回去休息而不是給病人開藥。另一邊，假設人工智能系統“賽華佗”面對相同症狀的病人前來就診時，首先排除的是疾病乙和疾病丙，而後在疾病甲和疾病丁之間做選擇，並在絕大多數情況下最終此類症狀診斷為疾病甲。如上文中提到的，人工智能系統的診斷精確度高於人類醫生，但並非百分之百精準。儘管如此，在這種情況下，大多人就醫的時候可能仍然會選擇人類醫生張聖手而非人工智能“賽華佗”。理由也很明顯：雖然人工智能系統的診斷準確性更高，但一旦出現誤診（將疾病甲誤診為疾病丁），其給患者造成的傷害更大。相比較而言，人類醫生即使誤診，也最多是將疾病甲誤診為疾病乙，

而不會將其誤診為疾病丁。這裡需要指出的是，筆者並不是認定所有應用於醫療診斷的人工智能系統都存在類似風險。但是由於人工智能系統的不透明性，我們並不清楚其內部工作機制，因此，上述思想實驗中的假設就有可能為真。正是這種可能性使得人們對於人工智能系統不透明性的擔憂具有合理性和必要性。即當人工智能系統的誤診有可能對患者造成嚴重傷害時，我們有必要知道其內部工作機制以避免此類傷害。單就這點而言，人工智能系統診斷的高精確性無法成為其不透明的理由。

二、患者的知情同意權

上個世紀 50 年代以來，隨著患者權利受到不斷的重視，在治療前需獲得患者的知情同意已成為醫學界的廣泛共識。簡單說來，所謂知情同意（Informed Consent）是指，醫生在治療前向患者提供相關資訊和說明，徵得病人同意再實施治療。這些資訊包括患者的身心狀況、預期的治療效果、以及治療所涉及到的風險等。在這一部分中，筆者嘗試從知情同意的目的及內容的角度來論證，患者的知情同意權要求應用於醫療診斷的人工智能系統更加透明，以降低和避免對患者造成的傷害。

首先，有效的知情同意需具備三種要素：充分知情（Fully Informed）、（患者具有）行為能力（Capacity）以及自願同意（Voluntary）。所謂充分知情是指，“醫生需將所有與治療有關的資訊告知患者。這些資訊包括（治療的）收益和風險、是否具有其他可替代療法、以及治療無法進行將會有什麼樣的結果等。”而患者應具備行為能力是指，“患者需具備同意的能力，這意味著患者能夠理解醫生所提供的資訊，並基於這些資訊做出決定。”自願是指，“患者同意或者不同意的決定是由患者自己做出的，並且沒有來自醫務人員、朋友以及家屬的壓力。”接下來一個問題是，知情同意的目的是什麼？或者說，我們為什麼認為患者知情同意的權利應該得到保障？針對這一問題，有兩種不同的回

應。第一種回應認為，知情同意的目的是尊重患者自主權。即讓患者在充分知情的情況下自己做決定是否接受治療以及接受什麼樣的治療，只有這樣，患者的自主權才能得到尊重。也只有這樣，患者的利益才能得到最大的保護。這裡隱含了一點，即只有患者自己知道究竟什麼才是對自己最有利的。所以，在充分知情情況下，患者做出的決定也最符合其利益。儘管這一觀點得到了大多數學者的支持，但是對知情同意的這種理解存在一個問題，即如果患者的決定很明顯對自己不利甚至會造成嚴重的傷害，是否仍然遵從患者的自主決定？與之相比，另一種回應認為，知情同意的目的是保護患者免受不必要的傷害。對知情同意的這種理解可以追溯二戰之後醫學倫理的幾個綱領性檔，如《赫爾辛基宣言》。可以說，知情同意最初的設定就是為了保護患者免遭醫生的怠忽職守（Malpractice）而造成的不必要的傷害。在這裡，筆者並不打算詳細討論這兩種理解，而是採用第二種觀點，即知情同意的目的是為了避免患者遭受不必要的傷害。

基於這一目的，為了避免患者遭受不必要的傷害，醫生有義務將關於治療的各種資訊提供給患者，以便患者做出有利於自己的決定。具體說來，這些資訊應該包括治療的利弊、可能的不適及風險、預計的副作用，以及患者提出的其他相關問題，特別是這些問題的回答會影響到患者的最終決定。可以看出，患者的知情同意權要求醫生盡可能地披露有關病情和治療的資訊。回到上文中應用於醫療診斷的人工智能系統，由於其不透明性，醫生無法預料診斷是否可靠以及根據其診斷而進行的治療是否會帶來不當的風險。在上文的思想實驗中，筆者假設了一種可能，即在某種情況下，人工智能系統的“誤診”會給患者造成極大的不必要的傷害。而避免此類傷害恰恰是患者知情同意權的目的和要求。因此，基於患者的知情同意權，應用於診斷的人工智能系統應該更加透明，以避免患者遭受不必要的傷害。

三、批評與回應

就筆者所主張的觀點，即應用於診斷的人工智能系統應該更加透明，以避免患者遭受不必要的傷害。一種可能的反駁是，人工智能系統的內部工作機制過於複雜，監控其內部的每一個步驟以理解其工作全過程幾乎是無法實現的。上文中提到的，基於深度學習演算法的人工智能類比人腦神經網路系統，其複雜性也類似於人腦。正像我們尚未完全了解人腦的工作機制一樣，我們對人工智能系統的內部原理也處在探索階段。換句話說，並非人類不想打開人工智能系統的“黑箱”，而是我們現階段尚無能力完全掌握其內部工作機制。既然如此，考慮到人工智能系統在診斷上的精確度，信賴人工智能而不是強求透明性似乎更加可取。

在這裡，筆者的回應是，讓人工智能系統更加透明並非要求對其內部的每一步都加以監控，也不是要求對於人工智能系統內部機制的完全了解。而是，當人工智能系統給出的結論不同於人類醫生的診斷，或者當不同的人工智能系統給出了不同的結論時，人類醫生能夠察覺並且知道為什麼會出現這種不同。人工智能系統相當於一位“超級專家”，我們並不要求人類醫生對於這些超級專家的所有結論都能完全理解。當超級專家的結論和人類醫生的診斷相一致的時候，人類醫生可以根據自己的專業知識和經驗告知患者這一結論的合理性。而當超級專家和人類醫生的診斷不一致的時候，人類醫生也應該知道為什麼會造成這種不一致，以便告知患者相關情況。也許，人工智能系統的複雜性尚不足以使醫生或者程式設計者知道造成這種不一致的原因。那麼，至少在現階段人工智能系統應該透明到，程式設計者或者醫生知道，對於人工智能系統給出的不同於人類醫生的診斷，有哪些關鍵因素參與其中。這些關鍵因素與疾病診斷的相關性越大，其診斷結論越可靠；反之，則越不可靠。已有的研究表明，人工智能系統的設計可以反映出哪些因素在輸出的結果中起到了關鍵性的作用。並且，確實存在不相關的因素最終影響到輸出結論的情況。

比如，一個預測犯罪可能性的人工智能系統將住址的郵遞區號納入“考量”，最終導致居住在特定地區的人在系統中呈現了高的犯罪可能性。郵遞區號就犯罪可能性而言，明顯屬於不相關因素。對於應用於醫療診斷的人工智能系統，其程式設計中也應該體現出哪些因素對於輸出的結論起到了關鍵性的作用。這樣，當出現了上文所說的不一致的結果的時候，如果人工智能系統的診斷有不相關的因素參與其中，那麼，醫生有義務告知患者這一情況，並提醒患者謹慎對待這一診斷結論。

四、結論

綜上所述，儘管人工智能系統在預測疾病方面表現出了較高的準確性，但是仍無法達到百之百的精準。由於其內部工作機制的不透明，使得我們有理由擔心，人工智能系統的誤診將有可能給患者造成出乎意料的嚴重的傷害。這種傷害甚至要比相同情況下人類醫生的誤診更加嚴重。如上所述，知情同意的目的在於避免患者遭受不必要的傷害。因此，患者的知情同意權要求應用於診斷的人工智能系統更加透明。這裡的“透明”，並不是要求監控人工智能系統內部的每一個步驟以理解其工作全過程，而是通過對人工智能系統的設定，使得當人工智能系統的結論與醫生的診斷不一致的時候，或者當不同的人工智能系統給出不一致的診斷時，醫生能夠知道造成這種不一致的原因，以便向患者解釋相關情況。如果說從技術層面考慮，實現這種透明性仍然存在一定的難度的話；那麼，至少在現階段人工智能系統應該做到，程式設計者或者醫生知道，對於人工智能系統給出的不同於人類醫生的診斷，有哪些關鍵因素參與其中。這些關鍵因素與疾病診斷的相關性越大，其診斷結論越可靠；反之，則越不可靠。

參考文獻 References

Alakwaa, F.M., Chaudhary, K. and Garmire, L.X. “Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data,” *Journal of Proteome Research*, 17.1 (2017): 337-347.

Bathaee, Y. “The Artificial Intelligence Black Box and the Failure of Intent and Causation,” *Harv. JL & Tech.*, 31 (2017): 889-938.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. “Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission,” in proceedings of the *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2015, pp. 1721-1730.

Castelvecchi, D. “Can we open the black box of AI?” *Nature News*, 538.7623 (2016): 20.

Eyal, Nir “Informed Consent,” *Stanford Encyclopedia of Philosophy* 10. https://www.nhs.uk/conditions/consent-to-treatment/. Accessed on July 15th, 2019.

London, A.J. “Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability,” *Hastings Center Report*, 49.1 (2019): 15-21.

Miotto, R., Li, L. and Dudley, J.T. “Deep Learning to Predict Patient Future Diseases from the Electronic Health Records,” in *European Conference on Information Retrieval*, March 2016, pp. 768-774.

Pande, Vijay. “Artificial Intelligence’s ‘Black ox’ is Nothing to Fear,” *The New York Times*. https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html. Accessed on July 15th, 2019.

Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H. and Hipp, J.D. “Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights into the Black Box for Pathologists,” Archives of *Pathology & Laboratory Medicine*, 143.7 (2019): 859-868.

Winfield, A.F. and Jirotka, M. “The Case for an Ethical Black Box,” in *Annual Conference Towards Autonomous Robotic Systems*, 2017, pp. 262-273.