

QUẢN TRỊ AI THEO VÒNG ĐỜI VÀ ĐỊNH HƯỚNG HOÀN THIỆN CƠ CHẾ QUẢN TRỊ THÍCH ỨNG TẠI VIỆT NAM

● HỨA THỊ THÙY ANH

Trường Đại học Kinh tế - Luật - Đại học Quốc gia Thành phố Hồ Chí Minh

Email: httanhcmulaw@gmail.com

TÓM TẮT:

Quản trị trí tuệ nhân tạo (AI) hiện đại đang chuyển dịch từ mô hình quản lý truyền thống sang mô hình quản trị theo vòng đời và quản trị thích ứng, trong đó hoạt động quản trị được thực hiện liên tục xuyên suốt toàn bộ vòng đời của hệ thống AI. Trên cơ sở lý giải cơ chế hình thành rủi ro của hệ thống AI, phân tích giới hạn của mô hình quản trị AI cố định và xu hướng phát triển của quản trị AI quốc tế, bài viết đề xuất một cấu trúc quản trị AI theo vòng đời gồm 3 hợp phần chính: (i) thiết lập các nguyên tắc hướng dẫn vận hành AI định hướng cho quản trị thích ứng; (ii) xây dựng quy trình quản trị theo vòng đời của hệ thống AI; (iii) thiết lập các cơ chế giám sát, đánh giá và hiệu chỉnh khung pháp lý nhằm bảo đảm khả năng thích ứng trước sự thay đổi liên tục của công nghệ AI. Từ đó, bài viết đề xuất một số hàm ý chính sách nhằm xây dựng mô hình quản trị AI thích ứng tại Việt Nam.

Từ khóa: trí tuệ nhân tạo, quản trị AI theo vòng đời, quản trị thích ứng, AIA, sandbox, kiểm toán thuật toán.

1. Đặt vấn đề

Trí tuệ nhân tạo (Artificial Intelligence - AI) đang tạo ra những thách thức mới đối với các mô hình quản lý truyền thống. Khác với các sản phẩm công nghệ thông thường vốn có cấu trúc và hành vi tương đối ổn định sau khi được đưa ra thị trường, các hệ thống AI hiện đại có khả năng tự học, tự tối ưu hóa và liên tục thay đổi hành vi thông qua quá trình tiếp nhận dữ liệu và tương tác với môi trường thực tế. Đồng thời, rủi

ro của AI thường không phát sinh từ một yếu tố kỹ thuật riêng lẻ mà hình thành từ sự tương tác giữa thuật toán, dữ liệu, con người và bối cảnh triển khai thực tế¹. Điều này khiến một hệ thống AI được đánh giá là an toàn trong môi trường thử nghiệm vẫn có thể phát sinh các rủi ro mới sau quá trình vận hành thực tế². Vì vậy, rủi ro của AI không chỉ tồn tại tại thời điểm thiết kế hoặc triển khai ban đầu mà có thể tiếp tục biến đổi xuyên suốt vòng đời của hệ thống.

Trong khi đó, phần lớn các mô hình quản lý truyền thống hiện nay vẫn chủ yếu được xây dựng trên tư duy kiểm soát cố định thay vì tập trung phân tích bối cảnh triển khai, vận hành của hệ thống AI. Cách tiếp cận này có thể phù hợp đối với các sản phẩm công nghệ ổn định nhưng ngày càng bộc lộ nhiều giới hạn đối với AI vốn có tính thích nghi thay đổi liên tục. Một hệ thống AI được xem là an toàn tại thời điểm triển khai vẫn có thể phát sinh các rủi ro mới trong quá trình tự học, cập nhật dữ liệu hoặc thay đổi bối cảnh vận hành thực tế. Điều này làm cho các mô hình quản trị mang tính tĩnh và kiểm tra một lần không còn đủ khả năng kiểm soát hiệu quả các rủi ro của AI.

Trước bối cảnh đó, nhiều quốc gia và tổ chức quốc tế đang dần chuyển dịch sang mô hình quản trị AI theo vòng đời và quản trị thích ứng. Theo cách tiếp cận này, hoạt động quản trị AI không chỉ dừng lại ở giai đoạn tiền kiểm mà được duy trì liên tục xuyên suốt toàn bộ vòng đời của hệ thống, từ thiết kế, thử nghiệm, triển khai đến giám sát hậu triển khai và hiệu chỉnh nghĩa vụ quản trị theo sự thay đổi của rủi ro³. Tại Việt Nam, vấn đề quản trị AI theo vòng đời cũng đã bước đầu được quan tâm thông qua quy định tại Luật Trí tuệ nhân tạo 2025 và các văn bản hướng dẫn thi hành. Tuy nhiên, khung pháp lý hiện nay vẫn còn tồn tại dưới dạng các yêu cầu phân tán và chưa hình thành một hệ sinh thái quản trị thích ứng thống nhất đầy đủ. Trên cơ sở đó, bài viết đề xuất một số định hướng nhằm hoàn thiện mô hình quản trị AI thích ứng tại Việt Nam.

2. Lý giải cơ chế phát sinh rủi ro cấu trúc và nhu cầu quản trị AI theo vòng đời

Một trong những đặc điểm quan trọng của AI là rủi ro thường không phát sinh từ một đặc tính kỹ thuật riêng lẻ, mà hình thành từ sự tương tác giữa thuật toán, dữ liệu, con người và môi trường triển khai thực tế⁴. Vì vậy, AI không thể được xem đơn thuần như một phần mềm được thiết kế cố định do AI có khả năng liên tục thích nghi và thay đổi trong quá trình vận hành⁵. Tức, rủi ro của các hệ thống AI không phát sinh dựa

trên những đặc tính nội tại ban đầu của hệ thống AI mà phát sinh khi các đặc tính nội tại của AI tương tác với nhau và được đặt trong một bối cảnh vận hành cụ thể Chẳng hạn, sự kết hợp giữa tính tự động và dữ liệu thiên lệch có thể khiến AI khuếch đại các sai lệch xã hội trên quy mô lớn⁶. Tương tự, khả năng tự học và thích nghi liên tục có thể khiến hệ thống thay đổi hành vi theo thời gian và dần phát sinh các rủi ro ngoài dự liệu ban đầu của nhà phát triển.⁷

Từ đó, có thể thấy rủi ro của AI không phải là trạng thái cố định tại thời điểm thiết kế, mà có khả năng biến đổi liên tục trong quá trình vận hành. Điều này tạo ra sự khác biệt rõ rệt giữa “AI trong phòng thí nghiệm” và “AI trong môi trường thực tế”. Một hệ thống được đánh giá là an toàn tại thời điểm thử nghiệm vẫn có thể trở nên rủi ro sau quá trình tương tác liên tục với dữ liệu và người dùng. Trường hợp chatbot Tay của Microsoft là ví dụ điển hình khi hệ thống, sau quá trình học từ dữ liệu trên mạng xã hội, đã nhanh chóng tạo ra các nội dung mang tính xúc phạm và phân biệt đối xử ngoài dự kiến ban đầu⁸.

Chính cơ chế phát sinh rủi ro mang tính động này đã đặt ra thách thức lớn đối với các mô hình quản lý truyền thống vốn dựa trên hoạt động kiểm tra một lần trước khi sản phẩm được đưa ra thị trường. Nếu các mô hình quản lý truyền thống được xây dựng trên giả định sản phẩm công nghệ sẽ tương đối ổn định sau khi triển khai thì giả định này không còn hoàn toàn phù hợp đối với AI. Do đó, quản trị AI cần chuyển sang mô hình quản trị theo vòng đời, trong đó hoạt động giám sát, đánh giá và kiểm soát rủi ro được duy trì liên tục từ giai đoạn thiết kế, triển khai, vận hành đến hậu triển khai của hệ thống AI.

3. Xu hướng quốc tế về quản trị theo vòng đời đối với AI

Xu hướng quốc tế hiện nay đang dần hội tụ về mô hình quản trị AI theo vòng đời. Trên cơ sở đó, nhiều tổ chức quốc tế và quốc gia đã bắt đầu chuyển sang cách tiếp cận quản trị liên tục xuyên suốt toàn bộ vòng đời của hệ thống AI, tiêu biểu là:

Trước hết, OECD đã xây dựng nền tảng khái niệm về “vòng đời hệ thống AI” gồm 6 giai đoạn: (i) lập kế hoạch và thiết kế; (ii) thu thập và xử lý dữ liệu; (iii) xây dựng và sử dụng mô hình; (iv) xác minh và xác nhận; (v) triển khai; và (vi) vận hành và giám sát. Điểm mấu chốt mà OECD nhấn mạnh là các giai đoạn này không nhất thiết phải diễn ra theo tuần tự tuyến tính mà mang tính lặp đi lặp lại theo sự thay đổi, thích nghi của hệ thống AI⁹.

Cách tiếp cận này tiếp tục được thể chế hóa mạnh mẽ hơn trong Đạo luật AI của Liên minh châu Âu (EU AI Act). Đối với các hệ thống AI có rủi ro cao, quản trị không dừng lại ở hoạt động đánh giá sự phù hợp trước triển khai mà được mở rộng sang cơ chế giám sát hậu kiểm. Các nhà phát triển phải duy trì hoạt động lưu vết hệ thống, giám sát hiệu suất và cập nhật đánh giá rủi ro trong suốt quá trình vận hành của hệ thống AI¹⁰.

Xu hướng quản trị theo vòng đời cũng đặc biệt rõ nét trong lĩnh vực y tế - nơi AI có thể tác động trực tiếp đến tính mạng và sức khỏe con người. Cơ quan Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (FDA) cùng Diễn đàn Cơ quan Quản lý Thiết bị Y tế Quốc tế (IMDRF) đã phát triển mô hình “Tổng vòng đời sản phẩm” (Total Product Lifecycle - TPLC)¹¹ đối với các phần mềm AI/ML trong y tế.

Nhìn chung, các xu hướng quốc tế hiện nay cho thấy sự chuyển dịch từ mô hình “kiểm tra một lần trước triển khai” sang mô hình quản trị liên tục và thích ứng trong toàn bộ vòng đời của hệ thống AI. Theo đó, trách nhiệm quản trị không còn chỉ tập trung tại thời điểm đưa sản phẩm ra thị trường, mà được duy trì xuyên suốt quá trình thiết kế, triển khai, vận hành và hậu triển khai nhằm bảo đảm khả năng kiểm soát các rủi ro động của AI trong môi trường thực tế.

4. Sự hình thành các yếu tố quản trị AI theo vòng đời trong pháp luật Việt Nam

Pháp luật Việt Nam đã bắt đầu xuất hiện các yếu tố của mô hình quản trị AI theo vòng đời, bước đầu tiếp cận AI như một hệ thống cần được quản trị xuyên suốt từ giai đoạn thiết kế, phát

triển, triển khai cho đến vận hành và hậu triển khai. Trên cơ sở đó, các yếu tố quản trị AI theo vòng đời trong pháp luật Việt Nam có thể được nhìn nhận qua 4 giai đoạn chính¹²:

Ở giai đoạn nhận diện và phân loại rủi ro, Pháp luật bước đầu tiếp cận AI theo mô hình phân tầng rủi ro, theo đó các hệ thống AI được phân loại thành các mức độ rủi ro khác nhau dựa trên tiêu chí như mức độ tác động đến quyền con người, an toàn, an ninh, lĩnh vực sử dụng và phạm vi ảnh hưởng của hệ thống. Đối với các hệ thống AI có rủi ro cao, yêu cầu quản lý rủi ro được đặt ra ngay từ giai đoạn thiết kế và phát triển nhằm nhận diện sớm các nguy cơ liên quan đến dữ liệu, thuật toán hoặc tác động xã hội của hệ thống.

Ở giai đoạn đánh giá và thử nghiệm trước triển khai, nhiều quy định bắt đầu chú trọng đến hoạt động quản trị dữ liệu, kiểm thử và đánh giá sự phù hợp trước khi hệ thống AI được đưa vào sử dụng. Đồng thời, đối với các hệ thống AI có rủi ro cao, yêu cầu đánh giá sự phù hợp và thông báo kết quả phân loại rủi ro cho cơ quan quản lý nhà nước cũng bắt đầu được đặt ra.

Ở giai đoạn giám sát hậu triển khai, Khác với mô hình quản lý truyền thống chỉ tập trung vào giai đoạn tiền kiểm, pháp luật Việt Nam bước đầu hình thành các cơ chế giám sát liên tục trong quá trình vận hành của hệ thống AI. Các yêu cầu về khả năng giám sát và can thiệp của con người, lưu trữ nhật ký hoạt động, báo cáo sự cố và cơ chế hậu kiểm theo mức độ rủi ro cho thấy hoạt động quản lý đang dần chuyển sang mô hình quản trị theo vòng đời.

Ở giai đoạn rà soát và tái hiệu chỉnh hệ thống AI, Do AI có khả năng tự học và thay đổi hành vi theo thời gian nên trong một số trường hợp, hệ thống AI có thể phải thực hiện đánh giá lại sự phù hợp và cập nhật các biện pháp quản lý rủi ro nhằm bảo đảm hệ thống tiếp tục đáp ứng các yêu cầu an toàn trong quá trình vận hành thực tế.

Tuy nhiên, các cơ chế này hiện vẫn tồn tại dưới dạng các yêu cầu phân tán và chưa thống nhất. Nói cách khác, Việt Nam đã xuất hiện các “thành tố” của quản trị theo vòng đời nhưng

chưa hình thành đầy đủ một cơ chế quản trị thích ứng đối với AI.

5. Đề xuất hoàn thiện mô hình quản trị AI theo vòng đời tại Việt Nam

Trên cơ sở kinh nghiệm quốc tế, mô hình quản trị AI theo vòng đời có thể được xây dựng dựa trên 3 trụ cột chính gồm: (i) hệ thống nguyên tắc quản trị thích ứng; (ii) cơ chế quản trị xuyên suốt vòng đời của hệ thống AI; và (iii) cơ chế điều phối, giám sát và hiệu chỉnh chính sách ở cấp độ hệ sinh thái. Trong đó:

Đối với yêu cầu (i), pháp luật cần thiết hoàn thiện hệ thống nguyên tắc định hướng cho quản trị AI thích ứng

Như đã nêu trên, việc quản trị AI phải được xem xét trong bối cảnh triển khai, vận hành cụ thể. Do đó, pháp luật cần từng bước hình thành các nguyên tắc đóng vai trò định hướng cách thức hệ thống quản trị sự bất định, tính thiếu minh bạch và khả năng tiến hóa liên tục của AI. Để làm rõ, các nguyên tắc cơ bản tại Điều 4 Luật Trí tuệ nhân tạo 2025 xác định mục tiêu quản trị cần bảo vệ những giá trị nào, trong khi các nguyên tắc quản trị thích ứng được đề cập tại mục này xác định tiêu chí vận hành như thế nào để bảo vệ hiệu quả các giá trị được đặt ra tại Điều 4 trong suốt vòng đời của hệ thống AI. Một hướng tiếp cận đáng tham khảo là khung nguyên tắc AFIRRM¹³ gồm: thích ứng (Adaptivity), linh hoạt (Flexibility), bao trùm (Inclusiveness), phản tư (Reflexivity), đáp ứng (Responsiveness) và giám sát liên tục (Monitoring).

Đối với yêu cầu (ii), pháp luật cần hoàn thiện cơ chế quản trị xuyên suốt vòng đời của hệ thống AI thống nhất

Pháp luật nên làm rõ vấn đề quản trị AI dựa trên rủi ro trong bối cảnh cụ thể cung như vai trò của các công cụ khi áp dụng vào các giai đoạn quản trị AI theo vòng đời và hoàn thiện các khung hướng dẫn đánh giá rủi ro để nhận diện rủi ro và phân bổ trách nhiệm quản lý rủi ro.

Đối với yêu cầu (iii), hình thành cơ chế điều phối, giám sát và hiệu chỉnh chính sách ở cấp độ hệ sinh thái

Theo đó, cơ chế này được thiết lập thông qua yêu cầu quản trị: (i) Thiết lập cơ chế điều phối liên ngành và thiết lập Hạ tầng dữ liệu sự cố quốc gia; (ii) Chuẩn hóa hệ thống quản lý tiêu chuẩn các tổ chức đánh giá kỹ thuật và cơ chế kiểm toán độc lập và (iii) Cơ chế thích ứng và hiệu chuẩn quy tổng thể của khung pháp lý AI.

Về Thiết lập cơ chế điều phối liên ngành và thiết lập Hạ tầng dữ liệu sự cố quốc gia, bên cạnh các nghĩa vụ quản trị áp dụng đối với doanh nghiệp, Việt Nam cần xây dựng cơ chế điều phối liên ngành nhằm bảo đảm khả năng giám sát và hiệu chỉnh chính sách trong bối cảnh công nghệ AI liên tục thay đổi.

Về Chuẩn hóa hệ thống quản lý tiêu chuẩn các tổ chức đánh giá kỹ thuật và cơ chế kiểm toán độc lập, trong quản trị AI thích ứng, pháp luật chỉ thiết lập các quy phạm mang tính bắt buộc quy phạm vĩ mô về mặt nguyên tắc, đặt ra tiêu chuẩn và chế tài đối với các tổ chức đánh giá sự phù hợp và tổ chức kiểm toán độc lập. Nhiệm vụ chuyển hóa các nguyên tắc thành các thông số định lượng và đo lường có thể được ủy quyền cho các tổ chức uy tín ở cấp độ quốc gia và quốc tế (như ISO, IEEE) kiến tạo.

Về Cơ chế thích ứng và hiệu chuẩn quy tổng thể của khung pháp lý AI, pháp luật cần xây dựng cơ chế cập nhật chính sách dựa trên dữ liệu thực tiễn thu thập từ hoạt động hậu kiểm, kiểm toán thuật toán và phản hồi xã hội nhằm bảo đảm khả năng thích ứng.

6. Kết luận

Sự phát triển của trí tuệ nhân tạo đang đặt ra yêu cầu chuyển đổi từ mô hình quản lý tĩnh sang mô hình quản trị AI theo vòng đời và quản trị thích ứng. Đối với Việt Nam, mặc dù đã bước đầu hình thành các yếu tố của mô hình quản trị theo vòng đời, khung pháp lý hiện nay vẫn cần tiếp tục được hoàn thiện theo hướng quản trị rủi ro động, tăng cường cơ chế giám sát liên tục và xây dựng cơ chế điều phối chính sách thích ứng đối với AI. Điều này sẽ góp phần bảo đảm sự cân bằng giữa kiểm soát rủi ro, bảo vệ lợi ích công cộng và thúc đẩy đổi mới sáng tạo trong bối cảnh phát triển nhanh chóng của công nghệ AI■

TÀI LIỆU TRÍCH DẪN:

- ¹Justin B. Bullock, Yu-Che Chen et al., *The Oxford Handbook of AI Governance* (Oxford University Press, 2024), chapter 22.
- ²Nikos Th. Nikolinakos, *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies: The AI Act* (Springer, 2023), pp. 567-568.
- ³Organisation for Economic Co-operation and Development (OECD), *OECD Framework for the Classification of AI Systems* (2022).
- ⁴National Institute of Standards and Technology (NIST) - U.S. Department of Commerce, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023)
- ⁵Nikos Th. Nikolinakos, *tlđđ sTh.*, pp. 273-274.
- ⁶Wim Naudé, Thomas Gries, and Nicola Dimitri, *Artificial Intelligence: Economic Perspectives and Models* (Cambridge University Press, 2024), p.227-229.
- ⁷Nikos Th. Nikolinakos, *tlđđ sTh.*, pp. 273-274.
- ⁸Luciano Floridi, *Ethics, Governance, and Policies in Artificial Intelligence* (Springer, 2021), pp. 130-132.
- ⁹OECD, “Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI”, *OECD Digital Economy Papers No. 349* (OECD Publishing, 2023), p.06-21.
- ¹⁰The AI Act - Regulation (EU) 2024/1689
- ¹¹Shin-yi Peng, Ching-Fu Lin and Thomas Streinz, *Artificial intelligence and international economic law* (Cambridge University Press, 2021), p. 79-82.
- ¹²Luật Trí tuệ nhân tạo (Luật số: 134/2025/QH15) ngày 10 tháng 12 năm 2025; Nghị định số 142/2026/NĐ-CP quy định chi tiết một số điều và biện pháp thi hành Luật Trí tuệ nhân tạo do Chính phủ ban hành ngày 30 tháng 4 năm 2026.
- ¹³Dubber, M. D., Pasquale, F., & Das, S., *The Oxford handbook of ethics of AI* (Oxford University Press, 2020), p.713-717.

TÀI LIỆU THAM KHẢO:

- Antunes, H. S., Freitas, P. M., Oliveira, A. L., Pereira, C. M., Sequeira, E. V., & Xavier, L. B. (2024). *Multidisciplinary perspectives on artificial intelligence and the law*. Springer.
- Barfield, W. (2020). *The Cambridge handbook of the law of algorithms*. Cambridge University Press.
- Bullock, J. B., Chen, Y.-C., et al. (2024). *The Oxford handbook of AI governance*. Oxford University Press.
- Chesterman, S. (2021). *We, the robots? Regulating artificial intelligence and the limits of the law*. Cambridge University Press.
- Corrales, M., Fenwick, M., & Forgó, N. (2018). *Robot, AI and the future of law*. Springer.
- Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.
- The AI Act - Regulation (EU) 2024/1689
- Floridi, L. (2021). *Ethics, governance, and policies in artificial intelligence*. Springer.
- Kovač, M. (2024). *Generative artificial intelligence: A law and economics approach to optimal regulation and governance*. Springer.
- Mathis, K., & Tor, A. (2023). *Law and economics of the digital transformation*. Springer.
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.
- Nikolinakos, N. T. (2023). *EU policy and legal framework for artificial intelligence, robotics and related technologies: The AI Act*. Springer.

Organisation for Economic Co-operation and Development (2022), OECD framework for the classification of AI systems.

Organisation for Economic Co-operation and Development (2023), “Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI”, OECD Digital Economy Papers No. 349, OECD Publishing, p.06-21.

Singh, T. (2024). Artificial intelligence and ethics: A field guide for stakeholders. CRC Press.

Smuha, N. A. (2025). The Cambridge handbook of the law, ethics and policy of artificial intelligence. Cambridge University Press.

Shin-yi Peng, Ching-Fu Lin and Thomas Streinz (2021), Artificial intelligence and international economic law, Cambridge University Press, p. 79-82.

Voeneky, S., Kellmeyer, P., Mueller, O., & Burgard, W. (Eds.). (2022). The Cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives. Cambridge University Press.

Ngày nhận bài: 05/5/2026

Ngày phản biện đánh giá và sửa chữa: 20/5/2026

Ngày chấp nhận đăng bài: 05/6/2026

ARTIFICIAL INTELLIGENCE LIFECYCLE GOVERNANCE AND DIRECTIONS FOR DEVELOPING AN ADAPTIVE GOVERNANCE FRAMEWORK IN VIETNAM

● **HUA THI THUY ANH**

University of Economics and Law, Vietnam National University,

Ho Chi Minh City, Vietnam

Email: httanhcmulaw@gmail.com

ABSTRACT:

Contemporary artificial intelligence (AI) governance is shifting from traditional regulatory models toward lifecycle-based and adaptive governance approaches, in which governance activities are continuously implemented throughout the entire lifecycle of AI systems. Based on an analysis of the mechanisms underlying AI risk formation, the limitations of static governance models, and global trends in AI governance, this study proposes a lifecycle-based governance structure consisting of three core components: (i) establishing guiding principles for AI operation to support adaptive governance; (ii) developing governance processes across the AI system lifecycle; and (iii) implementing monitoring, evaluation, and regulatory adjustment mechanisms to ensure adaptability to the evolving nature of AI technologies. Based on these findings, the study proposes policy implications for developing an adaptive AI governance model in Vietnam.

Keywords: artificial intelligence, AI lifecycle governance, adaptive governance, AIA, sandbox, algorithmic auditing.