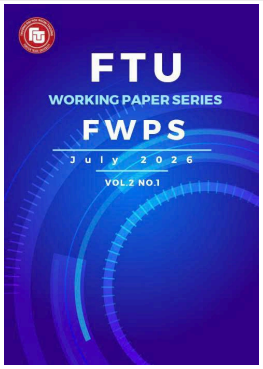

Working Paper 2026.1

- Vol.2 , No. 1

NHẬN DIỆN THÁCH THỨC ĐẠO ĐỨC KHI ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG Y TẾ VÀ HÀM Ý CHO CƠ SỞ Y TẾ

Đỗ Hoàng Linh[1],

Sinh viên K62 CLC Kinh doanh quốc tế – Viện Kinh tế & Kinh doanh quốc tế

Trường Đại học Ngoại thương, Hà Nội, Việt Nam

Phùng Ánh Hồng, Nguyễn Thị Hà, Lương Hồng Hải Ly,

Sinh viên K62 CLC Kinh tế đối ngoại – Viện Kinh tế & Kinh doanh quốc tế

Trường Đại học Ngoại thương, Hà Nội, Việt Nam

Nguyễn Thị Vân Trang

Giảng viên Viện Kinh tế & Kinh doanh quốc tế

Trường Đại học Ngoại thương, Hà Nội, Việt Nam

Tóm tắt

Sự bùng nổ của Trí tuệ nhân tạo (AI) trong quản lý vận hành và xử lý dữ liệu y tế đang tạo ra những bước tiến đột phá, nhưng đồng thời làm nảy sinh những rủi ro đạo đức và pháp lý chưa từng có. Bài báo này khám phá sự dịch chuyển rủi ro từ sai sót con người sang sai sót thuật toán, tập trung phân tích sự mơ hồ trong trách nhiệm giải trình khi xảy ra sự cố. Qua việc đánh giá ba thách thức cốt lõi: (1) quyền riêng tư và thương mại hóa dữ liệu bệnh nhân, (2) thiên kiến thuật toán trong phân bổ nguồn lực y tế, và (3) khủng hoảng niềm tin ảnh hưởng đến tài sản thương hiệu của bệnh viện; nghiên cứu chỉ ra rằng các chuẩn mực đạo đức y sinh truyền thống là chưa đủ. Bài báo đề xuất nên có một khung đạo đức AI toàn diện, được tích hợp trực tiếp vào chiến lược Trách nhiệm xã hội

[1] Tác giả liên hệ, Email: k62.2312550037@ftu.edu.vn

của doanh nghiệp (CSR) và quản trị rủi ro của các cơ sở khám chữa bệnh, nhằm đảm bảo sự cân bằng giữa đổi mới công nghệ và an toàn cho bệnh nhân.

Từ khoá: Trí tuệ nhân tạo, đạo đức y sinh, trách nhiệm giải trình

IDENTIFYING ETHICAL CHALLENGES IN THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE AND IMPLICATIONS FOR MEDICAL INSTITUTIONS

Abstract

The rapid expansion of Artificial Intelligence (AI) in operations management and medical data processing is driving groundbreaking advancements, while simultaneously giving rise to unprecedented ethical and legal risks. This paper explores the shift in risk from "human error" to "algorithmic errors," specifically focusing on the ambiguity of accountability when incidents occur. By evaluating three core challenges: (1) privacy and the commercialization of patient data, (2) algorithmic bias in healthcare resource allocation, and (3) the crisis of trust affecting hospital brand equity; the research indicates that traditional bioethical norms are no longer sufficient. The paper proposes a comprehensive AI Ethics Framework, integrated directly into the Corporate Social Responsibility (CSR) strategies and risk management protocols of medical institutions, to ensure a balance between technological innovation and patient safety.

Keywords: Artificial Intelligence (AI), bioethics, accountability

---

1. Đặt vấn đề

Trí tuệ nhân tạo (AI) đã cho thấy tiềm năng lớn khi được ứng dụng trong lĩnh vực y tế và chăm sóc sức khỏe. AI hỗ trợ bác sĩ trong quá trình ra quyết định, nâng cao chất lượng chăm sóc người sử dụng dịch vụ y tế, đồng thời tối ưu hóa quy trình làm việc và các nhiệm vụ hành chính. Tuy nhiên, việc ứng dụng AI trong y học vào thực tiễn cũng đã làm nảy sinh nhiều vấn đề đạo đức cần được nhận diện và giải quyết (Li và c.s., 2025).

Các quy định pháp lý ở cấp toàn cầu và khu vực, bao gồm các hướng dẫn do Tổ chức Y tế Thế giới và Liên minh châu Âu ban hành, cung cấp định hướng để thích ứng với sự thay đổi nhanh chóng của thế giới và giảm bớt những lo ngại ngày càng gia tăng về tác động đạo đức của AI trong việc cung cấp dịch vụ chăm sóc sức khỏe. Tại Việt Nam, với sự ra đời của Luật Trí tuệ nhân tạo 2025 (có hiệu lực từ tháng 3/2026) đã đặt ra những nền tảng pháp lý đầu tiên cho việc kiểm soát và quản lý khi áp dụng AI trong y tế.

Tuy nhiên, các khuôn khổ đạo đức hỗ trợ xây dựng quy định pháp lý về AI y tế thường được phát triển mà thiếu sự đối thoại giữa các nhà đạo đức học, nhà phát triển, bác sĩ và cộng đồng. Những mối quan ngại đạo đức do các nhà hoạch định chính sách và học giả nêu ra có thể không trùng khớp với những vấn đề mà bệnh nhân, bác sĩ và nhà phát triển thực sự gặp phải (Tang, Li và Fantus, 2023). Mối quan hệ pháp lý chưa rõ ràng giữa AI và người sử dụng hiện vẫn chưa thể được giải quyết dứt điểm (Arnold, 2021). Sự phát triển của AI và việc ứng dụng nó trong chăm sóc bệnh nhân sẽ đòi hỏi một quá trình đối thoại liên tục và lặp đi lặp lại nhằm bảo vệ các giá trị nhân văn trong các mô hình chăm sóc y tế tương lai. Khi xác định được các bên liên quan nhìn nhận những rủi ro đạo đức của AI như thế nào có thể thúc đẩy đối thoại và xây dựng các hướng dẫn thực hành, tiêu chuẩn tổ chức và quy định pháp lý phù hợp với thực tiễn. Do vậy, bài báo tập trung nhận diện thách thức đạo đức khi sử dụng trí tuệ nhân tạo trong y tế, từ đó đề xuất các hàm ý quản trị cho các cơ sở y tế, nhằm đảm bảo sự hài hòa giữa tiến bộ công nghệ và các giá trị nhân văn cốt lõi.

2. Phương pháp nghiên cứu và câu hỏi nghiên cứu

2.1. Phương pháp nghiên cứu

Nghiên cứu này tổng hợp, phân tích tài liệu và phân tích trường hợp điển hình. Các từ khóa tìm kiếm liên quan đến "AI trong y tế", "Quyền riêng tư", "thiên kiến", "trách nhiệm", …. Các tài liệu bao gồm bài báo khoa học, hướng dẫn quốc tế (WHO, EU, Anh, Mỹ,...), và văn bản pháp luật và báo cáo chính sách được thu thập và phân loại theo chủ đề.

2.2. Câu hỏi nghiên cứu

Trên cơ sở mục tiêu tổng quát và tính cấp thiết của vấn đề, nghiên cứu đặt ra các câu hỏi chính sau:

- Có những thách thức đạo đức nào khi triển khai AI trong chăm sóc sức khỏe?
- Những bên liên quan (bác sĩ, bệnh nhân, nhà phát triển công nghệ) nhận thức thế nào về các thách thức này và ưu tiên giải quyết những vấn đề nào?
- Các khung pháp lý, quy định và nguyên tắc đạo đức hiện nay có đáp ứng đầy đủ hay cần bổ sung, cải tiến ra sao xử lý rủi ro của AI trong y tế?

3. Cơ sở lý thuyết

3.1. Tổng quan về Trí tuệ nhân tạo và Trí tuệ nhân tạo trong y tế

Trí tuệ nhân tạo (AI)

AI được cấu thành từ hai từ: "nhân tạo" (artificial), chỉ những thứ do con người tạo ra, và "trí tuệ" (intelligence), chỉ khả năng tự suy nghĩ. Do đó, AI được định nghĩa là một dạng năng lực tư duy do con người tạo ra (Limna, Siripipatthanakul và Phayaphrom, 2021).

Trí tuệ nhân tạo trong y tế

AI được ứng dụng trong y học nhằm nâng cao chất lượng chăm sóc bệnh nhân thông qua việc phát hiện và chẩn đoán sớm hơn, cải thiện quy trình làm việc. AI trong chăm sóc sức khỏe đề cập đến việc sử dụng các thuật toán học máy và các công nghệ nhận thức khác nhằm hỗ trợ chẩn đoán, điều trị và quản lý toàn diện việc chăm sóc bệnh nhân (Zeb và c.s., 2024).

Về phân loại, ứng dụng của AI trong y học có hai nhánh chính: nhánh ảo và nhánh vật lý (Hamet và Tremblay, 2017). Nhánh ảo được thể hiện qua học sâu, hồ sơ bệnh án điện tử nhằm hỗ trợ cho bác sĩ trong việc ra quyết định điều trị. Trong khi đó, nhánh vật lý được thể hiện qua các đối tượng vật lý, thiết bị y tế và các robot tham gia vào quá trình chăm sóc bệnh nhân hoặc hỗ trợ bác sĩ phẫu thuật.

Trí tuệ nhân tạo trong quản lý hành chính y tế

Bằng cách thực hiện các nhiệm vụ lặp đi lặp lại như nhập dữ liệu bệnh nhân, tự động phân tích kết quả xét nghiệm và hình ảnh, AI giúp giải phóng thời gian để bác sĩ tập trung vào chăm sóc trực tiếp cho bệnh nhân. Dịch vụ y tế có thể sử dụng các thuật toán học máy tối ưu để hỗ trợ lập lịch khám bệnh và ưu tiên bệnh nhân, từ đó giảm thời gian chờ đợi và nâng cao hiệu quả sử dụng dịch vụ. AI còn giúp dự đoán thời gian nằm viện của bệnh nhân ngay từ trước khi nhập viện, góp phần sử dụng hiệu quả hơn nguồn lực bệnh viện (Reddy, Fox và Purohit, 2019).

3.2. Các nguyên tắc đạo đức y sinh

Khung phân tích do Beauchamp và Childress (2009) đề xuất đã thiết lập bốn nguyên tắc cơ bản của Đạo đức y sinh bao gồm: quyền tự quyết, không gây hại, lòng nhân từ và công bằng. Đạo đức AI trong y học được củng cố bởi bốn nguyên tắc đạo đức y sinh này cùng với các giá trị như công bằng, an toàn, minh bạch, bảo vệ quyền riêng tư, trách nhiệm và niềm tin (Tang, Li và Fantus, 2023). Những nguyên tắc và giá trị này định hướng cho các quy định pháp lý, tiêu chuẩn tổ chức và chính sách nhằm tập trung giải quyết các thách thức đạo đức cản trở việc ứng dụng AI trong y tế.

3.3. Khung pháp lý liên quan đến Trí tuệ nhân tạo trong y tế

3.3.1. Trên thế giới

(1) Tại Liên minh châu Âu

Liên minh châu Âu (EU) bắt đầu xây dựng cách tiếp cận đối với trí tuệ nhân tạo thông qua các hướng dẫn không mang tính ràng buộc pháp lý, bao gồm "Hướng dẫn đạo đức về AI đáng tin cậy" và "Khuyến nghị về chính sách và đầu tư", được công bố vào năm 2019.

Tháng 4 năm 2021, EU đã đề xuất Đạo luật Trí tuệ nhân tạo, thiết lập khuôn khổ pháp lý thống nhất cho các sản phẩm và dịch vụ AI, áp dụng phương pháp tiếp cận dựa trên mức độ rủi ro để quản lý các hệ thống AI. Đặc biệt trong lĩnh vực y tế, các hệ thống AI có rủi ro cao nên càng đòi hỏi nhà sản xuất phải đảm bảo tuân thủ.

(2) Tại Hoa Kỳ

Hiện nay, Hoa Kỳ chưa có quy trình quản lý riêng biệt, tuy nhiên Cơ quan Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (FDA) vẫn đánh giá công nghệ dựa trên AI theo khuôn khổ quản lý hiện hành đối với thiết bị y tế.

Vào tháng 4 năm 2019, FDA đưa ra "Khung quản lý đề xuất cho các thay đổi đối với phần mềm thiết bị y tế dựa trên AI/ML". Theo đó, các nhà phát triển phải chịu trách nhiệm về hiệu suất thực tế của hệ thống và cần cập nhật cho FDA về những thay đổi liên quan đến hiệu suất và dữ liệu đầu vào.

Tháng 1 năm 2021, FDA ban hành "Kế hoạch hành động cho phần mềm thiết bị y tế (SaMD) dựa trên AI/ML", trong đó đề ra 5 hướng hành động chính dựa trên cách tiếp cận vòng đời của sản phẩm để giám sát các thiết bị y tế tích hợp AI, áp dụng cách tiếp cận lấy bệnh nhân làm trung tâm, đảm bảo tính minh bạch của thiết bị.

(3) Tại Vương quốc Anh

Viện Y tế và Chăm sóc Sức khỏe Quốc gia phối hợp với Dịch vụ y tế Quốc gia công bố "Khung tiêu chuẩn bằng chứng cho công nghệ y tế số" vào năm 2019. Tài liệu này đưa ra các quy định đối với nhiều loại sản phẩm như ứng dụng, phần mềm và nền tảng trực tuyến, có thể hoạt động độc lập hoặc kết hợp với các sản phẩm y tế khác.

Cơ quan tư vấn cho chính phủ Anh về đổi mới công nghệ đã công bố báo cáo "Quản lý AI như một thiết bị y tế" vào tháng 11 năm 2022. Báo cáo này xem xét toàn bộ vòng đời sản phẩm của các thiết bị y tế tích hợp AI và nhằm tăng cường sự tham gia của bệnh nhân và công chúng, qua đó cải thiện tính minh bạch trong giao tiếp giữa cơ quan quản lý, nhà sản xuất và người sử dụng.

Vào tháng 9 năm 2021, Cơ quan Quản lý Dược phẩm và Sản phẩm Y tế đã triển khai một chương trình cải cách quản lý "Chương trình thay đổi đối với phần mềm và AI như thiết bị y tế" nhằm xây dựng một khuôn khổ quản lý vững chắc thông qua các hướng dẫn để giám sát thiết bị y tế sử dụng AI.

3.3.2. Tại Việt Nam

Các quy định pháp luật Việt Nam hiện hành về AI, đặc biệt trong lĩnh vực y tế, vẫn còn đang trong giai đoạn phát triển và hoàn thiện.

Có một số điều khoản liên quan đến đảm bảo an ninh và bảo mật thông tin được quy định trong Luật An toàn thông tin mạng 2015 và Luật An ninh mạng 2018. Nghị định 13/2023/NĐ-CP quy định về bảo vệ dữ liệu cá nhân và trách nhiệm bảo vệ dữ liệu cá nhân của cơ quan, tổ chức, cá nhân có liên quan. Luật Trí tuệ nhân tạo 2025 đánh dấu bước ngoặt quan trọng trong quản trị số tại Việt Nam, phân loại hệ thống AI thành ba mức độ: thấp, trung bình và cao. Tại Điều 6 quy định, việc ứng dụng trí tuệ nhân tạo trong lĩnh vực y tế cần bảo đảm an toàn cho người sử dụng dịch vụ y tế; độ tin cậy trong điều kiện thực tế; bảo vệ dữ liệu về sức khỏe theo quy định của pháp luật.

4. Những thách thức đạo đức khi ứng dụng Trí tuệ nhân tạo trong y tế

Qua nghiên cứu các văn liệu quốc tế (WTO (2021), European Parliament & Council (2024), (U.S. FDA, Center for Devices and Radiological Health (CDRH), (2021)) và các sự kiện nổi bật như DeepMind Streams về quyền riêng tư (Powles và Hodson, 2017), Watson Oncology về sai sót (Ross và Swetliz, 2018), thuật toán y tế Mỹ về bất bình đẳng (Joseph, 2025), nghiên cứu lựa chọn 3 thách thức - hộp đen và trách nhiệm, thương mại hóa dữ liệu và quyền tự chủ, thiên kiến thuật toán và bất bình đẳng y tế - làm trọng tâm dựa trên thực tế là chúng xuất hiện xuyên suốt trong các tài liệu nổi bật.

4.1. Vấn đề hộp đen và trách nhiệm giải trình

Mối quan ngại chính là sự bất đối xứng thông tin giữa bác sĩ và thuật toán. AI y tế thường là hệ thống phức tạp mà bác sĩ không hiểu hết được cơ chế ra quyết định (Stogiannos và c.s., 2025), kết hợp với áp lực thời gian trong thực hành lâm sàng và tâm lý mặc định công nghệ cao luôn ưu việt hơn con người, tạo ra thiên kiến tự động hóa (Leslie và c.s., 2024). Lúc này, bác sĩ có xu hướng tin tưởng vào đề xuất của hệ thống mà bỏ qua các bước đánh giá, kiểm chứng độc lập. Hậu quả là khi AI tư vấn sai, việc phân chia trách nhiệm pháp lý giữa nhà cung cấp công nghệ, cơ sở y tế và bác sĩ lâm sàng trở nên mờ nhạt (Stogiannos và c.s., 2025). Như trường hợp IBM Watson Oncology, mặc dù được ứng dụng tại 230 bệnh viện trên thế giới, hệ thống này đã đưa ra khuyến nghị điều trị sai và không an toàn (Ross và Swetliz, 2018). Từ đây, một khoảng trống pháp lý nghiêm trọng lộ rõ. Bản chất AI không phải là một pháp nhân, những sự phức tạp của thuật toán khiến việc truy xuất nguồn gốc sai sót - lỗi lập trình của nhà cung cấp, dữ liệu cơ sở y tế, hay sự thiếu giám sát của các bác sĩ - trở nên gần như bất khả thi. Sự mờ nhạt trong các yếu tố quản trị và quy định pháp lý dẫn đến tình trạng đùn đẩy trách nhiệm khi xảy ra sự cố y khoa, từ đó làm xói mòn cả niềm tin của bệnh nhân lẫn giới chuyên môn vào công nghệ.

4.2. Thương mại hóa dữ liệu và quyền tự chủ

Sự phụ thuộc của AI y tế vào nguồn dữ liệu khổng lồ đang tạo ra mâu thuẫn sâu sắc giữa mục tiêu cải tiến điều trị và nguy cơ thương mại hóa dữ liệu cá nhân (Lewin và c.s., 2024). Nhiều trường hợp (ví dụ Google DeepMind với dự án Streams tại Anh (Powles và Hodson, 2017)) cho thấy dữ liệu nhạy cảm của hàng triệu bệnh nhân được chuyển cho công ty tư nhân mà không xin đồng thuận đầy đủ. Mâu thuẫn cấu trúc này biến quyền được đồng thuận thành một thủ tục mang tính hình thức (Powles và Hodson, 2017); bởi lẽ, trên thực tế, bệnh nhân thường không hiểu rõ mức độ AI can thiệp vào quá trình điều trị và không lường trước được mọi kịch bản khai thác dữ liệu của mình trong tương lai (Yu, Lee và Hwang, 2024). Ở Việt Nam, bất cập này càng nổi bật khi các yếu tố quản trị và quy định pháp lý đối với dữ liệu y tế AI còn nhiều khoảng trống. Mặc dù Luật Trí tuệ nhân tạo 2025 (số 134/2025/QH15) đã yêu cầu bảo vệ dữ liệu sức khỏe, nhưng việc thiếu hụt các cơ chế minh bạch về mục đích và phạm vi sử dụng dữ liệu khiến việc bảo vệ quyền tự chủ của bệnh nhân trở nên rất khó.

4.3. Thiên kiến thuật toán và bất bình đẳng y tế

AI học từ dữ liệu lịch sử do con người tạo ra, do đó nó không chỉ phản ánh mà còn tạo ra vòng lặp phản hồi, làm trầm trọng thêm các bất công xã hội (Obermeyer và c.s., 2019).

Quá trình này diễn ra theo một vòng tròn khép kín: hệ thống học tập từ dữ liệu đầu vào bị thiên lệch, dẫn đến đưa ra các quyết định y khoa và phân bổ nguồn lực sai lệch, thực tiễn sai lệch tạo ra các hồ sơ y tế mới bị ô nhiễm, và rồi lại tiếp tục được nạp lại vào quá trình huấn luyện khiến thiên kiến ngày càng phình to theo thời gian. Biểu hiện ở mắt xích đầu tiên là sự khuyết thiếu đại diện trong dữ liệu. Các mô hình dự đoán nhiễm trùng huyết thường sai số cao với nhóm dân tộc thiểu số, hoặc các ứng dụng y tế như Aarogya Sety ở Ấn Độ đã loại trừ hoàn toàn nhóm dân cư không có thiết bị thông minh (Joseph, 2025). Ở mắt xích tiếp theo, theo nghiên cứu của Obermeyer và c.s. (2019) trên hàng triệu bệnh nhân cho thấy cho thấy thuật toán đã đánh giá sai rủi ro sức khỏe, khiến tỷ lệ bệnh nhân da màu được chỉ định vào chương trình chăm sóc đặc biệt chỉ đạt 17,7%; sau khi sửa lại thuật toán, con số này đã tăng lên 46,5%. Ở mắt xích cuối, khi AI tự động tối ưu hóa các chỉ số thống kê, hệ thống có xu hướng thiên vị ưu tiên lịch khám cho nhóm bệnh nhân giàu có, khỏe mạnh hơn, trong khi bỏ sót hoặc phân loại thấp các trường hợp phức tạp nhằm cải thiện tỷ lệ thành công thống kê của bệnh viện (Stogiannos và c.s., 2025). Do đó, thiên kiến được xem là trọng tâm vì nó trực tiếp đe dọa công bằng xã hội và mục tiêu y tế toàn dân.

4.4. Phân tích trường hợp Dự án Streams của Google DeepMind

Vụ việc DeepMind Streams tại Anh (2015-2017) đã minh chứng rõ thách thức đạo đức khi AI y tế xử lý dữ liệu. ICO kết luận Bệnh viện Royal Free London NHS Trust đã trao quyền truy cập 1,6 triệu hồ sơ bệnh án (nhập viện, xuất viện, xét nghiệm từ nhiều năm) cho DeepMind là không tuân thủ luật bảo vệ dữ liệu vì bệnh nhân không hề được thông báo (Powles và Hodson, 2017).

Vấn đề cốt lõi là đồng thuận và minh bạch. Mặc dù Royal Free và DeepMind viện dẫn cơ chế điều trị trực tiếp và đồng ý ngầm để biện minh, nhưng trong thực tế không một bệnh nhân nào được thông báo hay ký đồng ý cụ thể. Bài viên của Powles và Hodson mô tả hậu quả là "những sai sót không thể chấp nhận" và cảnh báo quyền lực bất đối xứng giữa công dân và công ty công nghệ. Đôi bên đã phủ nhận sai sót, nhưng rõ ràng quy trình này đã xâm phạm quyền tự chủ của bệnh nhân và gây tổn thất niềm tin.

Về khía cạnh pháp lý, ICO đánh giá Royal Free giữ vai trò người quyết định cách dùng dữ liệu, còn DeepMind chỉ là người dùng dữ liệu. Theo đó, trách nhiệm tuân thủ luật pháp y tế thuộc về bệnh viện. ICO kết luận việc thử nghiệm ứng dụng với dữ liệu thực đã vượt quá quyền hạn của Royal Free, tuy không phạt tiền nhưng buộc bệnh viện xem xét lại toàn bộ cơ sở pháp lý và công khai kết quả kiểm toán nội bộ. Theo Quy định Anh (DPA 1998 áp dụng năm 2017), nếu không ở trong mối quan hệ điều trị trực tiếp với từng bệnh nhân thì phải có sự đồng thuận rõ ràng hoặc ủy quyền nghiên cứu đặc biệt để xử lý dữ liệu nhạy cảm. DeepMind thừa nhận chưa xin phép nghiên cứu và chưa tiếp cận cơ quan chuyên trách (NDG, NHS, CAG) cho kế hoặc trước đó (Powles và Hodson, 2017).

Qua đó thấy được bài toán quyền riêng tư và thương mại hóa dữ liệu khi mà giá trị của 1,6 triệu hồ sơ bệnh nhân bị khai thác mà bệnh nhân không kiểm soát. Nó cũng làm nổi bật khủng hoảng niềm tin và rủi ro thương hiệu, công chúng cảm thấy tổn thương khi phát hiện dữ liệu y tế bị dùng trái dự kiến. Vụ việc còn liên quan đến nguyên tắc công bằng, nhiều

chuyên gia lưu ý dòng trao đổi dữ liệu rời rạc chỉ phục vụ cho nhóm bệnh nhân AKI (chiếm 1/6), còn 5/6 bệnh nhân còn lại không hưởng lợi gì mà không được biết hoặc đồng ý. Như vậy, nguyên tắc tự chủ của bệnh nhân đã bị xâm phạm nghiêm trọng (Powles và Hodson, 2017).

Trường hợp DeepMind Streams minh họa rõ cơ chế rủi ro đạo đức phát sinh khi dữ liệu y tế được chia sẻ trên quy mô lớn trong khi bệnh nhân không được bảo đảm đầy đủ quyền được biết và quyền đồng thuận. Khi các tổ chức y tế ưu tiên hiệu quả công nghệ và hợp tác đổi mới hơn tính minh bạch trong quản trị dữ liệu, nguy cơ xâm phạm quyền riêng tư và suy giảm niềm tin công chúng sẽ gia tăng. Vụ việc cho thấy rủi ro khi khung pháp lý và đạo đức vẫn chưa theo kịp tốc độ phát triển của AI y tế, đặc biệt trong việc xác định ranh giới giữa mục tiêu chăm sóc sức khỏe, nghiên cứu và thương mại hóa dữ liệu bệnh nhân.

4.5. Nghiên cứu trường hợp của phần mềm AI RAPID

Tại Việt Nam, rủi ro đạo đức về thiên kiến tự động hóa xảy ra qua thực tiễn triển khai phần mềm trí tuệ nhân tạo RAPID trong cấp cứu đột quỵ. Từ năm 2019, Bệnh viện Nhân dân 115 và Bệnh viện Gia An 115 đã tiên phong ứng dụng hệ thống này nhằm phân tích hình ảnh quét não, giúp mở rộng giờ vàng can thiệp từ 6 giờ lên tới 24 giờ.

Trong thực tế cấp cứu, nếu hệ thống AI phân tích và hiển thị kết quả "NO" (không có chỉ định can thiệp), các bác sĩ gần như sẽ từ chối phẫu thuật cho bệnh nhân đó. Tuy nhiên, thuật toán của RAPID có thể bị lỗi và đưa ra đánh giá sai lệch do các yếu tố khách quan.

Dưới áp lực của một hệ thống công nghệ hiện đại, các bác sĩ dễ dàng rơi vào trạng thái chấp nhận AI một cách thiếu phản biện. Khi một bệnh nhân mất đi cơ hội can thiệp tái thông mạch máu cuối cùng vì một lỗi nhiễu ảnh khiến AI báo "NO", câu hỏi về việc ai sẽ phải chịu trách nhiệm vẫn là một điểm nghẽn pháp lý và đạo đức lớn trong việc triển khai AI y tế.

Trường hợp của phần mềm RAPID cho thấy thiên kiến tự động hóa có thể hình thành khi các bác sĩ phụ thuộc vào khuyến nghị của hệ thống AI trong bối cảnh áp lực thời gian và tính cấp bách của quyết định lâm sàng. Rủi ro về việc phân định trách nhiệm giữa bác sĩ, bệnh viện và nhà phát triển công nghệ khi AI tham gia trực tiếp vào quá trình ra quyết định cho thấy sự cần thiết của việc xây dựng các cơ chế kiểm chứng thuật toán, khung thử nghiệm có kiểm soát và đào tạo nhân lực y tế nhằm hạn chế sự phụ thuộc quá mức vào hệ thống AI trong thực hành lâm sàng.

5. Hàm ý cho cơ sở y tế

Từ việc nhận diện các rủi ro đạo đức, các giải pháp và kiến nghị do nhóm tác giả đề xuất hướng đến thiết lập một hệ sinh thái y tế số an toàn, minh bạch. Việc thực hiện đồng bộ các biện pháp này giúp cơ sở y tế tối ưu hóa hiệu quả điều trị, đồng thời bảo vệ quyền lợi và củng cố niềm tin của người sử dụng dịch vụ y tế.

5.1. Tích hợp đạo đức trí tuệ nhân tạo vào chiến lược trách nhiệm xã hội và quản trị doanh nghiệp

Việc kết hợp các nguyên tắc đạo đức vào chiến lược trách nhiệm xã hội và các tiêu chuẩn ESG là giải pháp quan trọng để bảo vệ tài sản thương hiệu của các bệnh viện. Trong bối cảnh dữ liệu y tế tồn tại rủi ro thương mại hóa làm gia tăng nguy cơ xâm phạm quyền riêng tư và suy giảm quyền tự chủ của bệnh nhân trong việc kiểm soát thông tin cá nhân, cách tiếp cận này giúp củng cố niềm tin của công chúng. Thực hiện bảo mật dữ liệu và đảm bảo tính minh bạch của thuật toán như một cam kết chiến lược sẽ giúp các cơ sở y tế khẳng định được việc tuân thủ các giá trị đạo đức.

5.2. Xây dựng tiêu chuẩn đạo đức và quy định quản lý thiết bị y tế sử dụng AI

Nhằm đảm bảo an toàn cho người sử dụng dịch vụ y tế, ban giám đốc các cơ sở y tế có thể thiết lập bộ tiêu chuẩn cơ sở riêng biệt cho các thiết bị tích hợp trí tuệ nhân tạo như hệ thống chẩn đoán hình ảnh hay thiết bị phẫu thuật tự động. Các quy định này cần yêu cầu tính minh bạch cao, đảm bảo thuật toán đằng sau các đề xuất lâm sàng được giải thích rõ ràng để bác sĩ có thể kiểm chứng. Khi nhiều hệ thống AI hoạt động như một hộp đen gây khó khăn trong việc xác định trách nhiệm khi xảy ra sai sót chuyên môn, bộ tiêu chuẩn này sẽ góp phần nâng cao tính minh bạch, củng cố trách nhiệm giải trình trong quá trình ứng dụng AI vào khám chữa bệnh.

5.3. Xây dựng khung thử nghiệm và phương án quản trị rủi ro

Các cơ sở y tế cần chủ động lập ra các khung thử nghiệm có kiểm soát ngay tại đơn vị trước khi đưa công nghệ mới vào quy trình khám chữa bệnh thực tế. Điều này cho phép hội đồng chuyên môn đánh giá toàn diện các tác động đạo đức và nhận diện sớm sai số kỹ thuật. Việc thử nghiệm trước triển khai giúp phát hiện nguy cơ thiên kiến thuật toán và hạn chế khả năng tạo ra sai lệch kết quả giữa các nhóm bệnh nhân khác nhau. Bên cạnh đó, cơ sở y tế cần chú trọng đào tạo nhân lực để y bác sĩ hiểu rõ giới hạn của thuật toán nhằm bảo vệ quyền lợi chính đáng của người sử dụng dịch vụ y tế và nhân viên y tế khi phát sinh sự cố.

6. Kết luận

Việc ứng dụng AI trong y tế góp phần nâng cao chất lượng chăm sóc, nhưng đã làm phát sinh một hệ thống rủi ro đạo đức mới. Nghiên cứu cho thấy các thách thức chủ yếu bao gồm tính minh bạch của hệ thống, bảo mật và thương mại hóa dữ liệu bệnh nhân, thiên kiến trong quyết định y tế và hậu quả về niềm tin với bệnh viện.

Từ việc nhận diện các thách thức đạo đức trong ứng dụng trí tuệ nhân tạo, nghiên cứu đề xuất một số hàm ý quản trị dành cho các cơ sở y tế nhằm bảo đảm việc triển khai AI diễn ra có đạo đức và minh bạch. Khung đạo đức AI đề xuất - tích hợp trực tiếp vào CSR - nhằm đảm bảo rằng đổi mới công nghệ song hành cùng bảo vệ an toàn người sử dụng dịch vụ y tế. Các bệnh viện cần chủ động xây dựng quy tắc rõ ràng cho từng sản phẩm AI và tổ chức giám sát liên tục. Cùng với đó là hoàn thiện khung quy định đặc thù cho AI y tế, đẩy mạnh giám

sát định tính và định lượng thông qua thử nghiệm có kiểm soát. Chỉ khi đó AI mới thực sự là công cụ thúc đẩy tiến bộ y học mà không đánh mất các giá trị nhân bản vốn có trong y đức.

Đồng thời, cần khuyến khích đối thoại liên ngành giữa nhà đạo đức, bác sĩ, nhà phát triển và bệnh nhân để đảm bảo rằng các khuyến nghị chính sách không chỉ khớp lý thuyết mà còn phù hợp với thực tế.

TÀI LIỆU THAM KHẢO

Beauchamp, T.L. và Childress, J.F. (2009) *Principles of Biomedical Ethics*. Oxford University Press.

European Parliament & Council (2024) *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. Truy cập tại: http://data.europa.eu/eli/reg/2024/1689/oj (Truy cập: 22 Tháng Ba 2026).

Hamet, P. và Tremblay, J. (2017) "Artificial intelligence in medicine", *Metabolism*, 69, tr S36–S40. Truy cập tại: https://doi.org/10.1016/j.metabol.2017.01.011.

Joseph, J. (2025) "Algorithmic bias in public health AI: a silent threat to equity in low-resource settings", *Frontiers in Public Health*, 13, tr 1643180. Truy cập tại: https://doi.org/10.3389/fpubh.2025.1643180.

Leslie, D. và c.s. (2024) "AI Fairness in Practice", *SSRN Electronic Journal* [Preprint]. Truy cập tại: https://doi.org/10.2139/ssrn.4731838.

Lewin, S. và c.s. (2024) "Ethical Challenges and Opportunities in Applying Artificial Intelligence to Cardiovascular Medicine", *Canadian Journal of Cardiology*, 40(10), tr 1897–1906. Truy cập tại: https://doi.org/10.1016/j.cjca.2024.06.029.

Li, Y. và c.s. (2025) "Exploring Factors Influencing Patients' Intention to Adopt Generative AI on Online Healthcare Platforms", *Journal of Theoretical and Applied Electronic Commerce Research*, 20(4), tr 287. Truy cập tại: https://doi.org/10.3390/jtaer20040287.

Limna, P., Siripipatthanakul, S. và Phayaphrom, B. (2021) "The Role of Big Data Analytics in Influencing Artificial Intelligence (AI) Adoption for Coffee Shops in Krabi, Thailand". Rochester, NY: Social Science Research Network. Truy cập tại: https://papers.ssrn.com/abstract=3944075.

Luật Trí tuệ nhân tạo 2025 số 134/2025/QH15 (2025). Truy cập tại: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Luat-Tri-tue-nhan-tao-2025-so-134-2025-QH15-679013.aspx (Truy cập: 22 Tháng Ba 2026).

Obermeyer, Z. và c.s. (2019) “Dissecting racial bias in an algorithm used to manage the health of populations”, *Science*, 366(6464), tr 447–453. Truy cập tại: https://doi.org/10.1126/science.aax2342.

Powles, J. và Hodson, H. (2017) “Google DeepMind and healthcare in an age of algorithms”, *Health and Technology*, 7(4), tr 351–367. Truy cập tại: https://doi.org/10.1007/s12553-017-0179-1.

Reddy, S., Fox, J. và Purohit, M.P. (2019) “Artificial intelligence-enabled healthcare delivery”, *Journal of the Royal Society of Medicine*, 112(1), tr 22–28. Truy cập tại: https://doi.org/10.1177/0141076818815510.

Ross, C. và Swetliz, I. (2018) *IBM’s Watson recommended “unsafe and incorrect” cancer treatments | STAT*. STAT .Truy cập tại: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (Truy cập: 22 Tháng Ba 2026).

Stogiannos, N. và c.s. (2025) “Ethical AI: A qualitative study exploring ethical challenges and solutions on the use of AI in medical imaging”, *European Journal of Radiology Artificial Intelligence*, 1, tr 100006. Available at: https://doi.org/10.1016/j.ejrai.2025.100006.

Tang, L., Li, J. và Fantus, S. (2023) “Medical artificial intelligence ethics: A systematic review of empirical studies - Lu Tang, Jinxu Li, Sophia Fantus, 2023”, *DIGITAL HEALTH* [Preprint].Truy cập tại: https://journals.sagepub.com/doi/full/10.1177/20552076231186064.

U.S. FDA, Center for Devices and Radiological Health (CDRH) (2021) “Artificial Intelligence/Machine Learning (AI/ML)-Based.:Jf/<X Software as a Medical Device (SaMD) Action Plan”.

WTO (2021) *Ethics and governance of artificial intelligence for health*. Truy cập tại: https://www.who.int/publications/i/item/9789240029200 (Truy cập: 22 Tháng Ba 2026).

Yu, S., Lee, S.-S. và Hwang, H. (2024) “The ethics of using artificial intelligence in medical research”, *Kosin Medical Journal*, 39(4), tr 229–237. Truy cập tại: https://doi.org/10.7180/kmj.24.140.

Zeb, S. và c.s. (2024) “AI in Healthcare: Revolutionizing Diagnosis and Therapy”, *International Journal of Multidisciplinary Sciences and Arts*, 3(3), tr 118–128. Truy cập tại: https://doi.org/10.47709/ijmdsa.v3i3.4546.