Báo cáo Nghiên cứu Khoa học: Đạo đức trong Trí tuệ Nhân tạo

Tác giả: Trịnh Minh Khánh

Tóm tắt

Trí tuệ Nhân tạo (AI) đang tạo ra những thay đổi sâu rộng trong nhiều lĩnh vực của đời sống. Tuy nhiên, sự phát triển nhanh chóng này cũng đặt ra những thách thức đạo đức đáng kể. Báo cáo này khám phá các khái niệm cơ bản về đạo đức AI, các nguyên tắc cốt lõi, những thách thức hiện tại và các khung quản trị được đề xuất để đảm bảo sự phát triển và ứng dụng AI có trách nhiệm. Chúng tôi tập trung vào các khuyến nghị của UNESCO và quan điểm của IBM, cùng với các nghiên cứu khoa học thuật, để cung cấp một cái nhìn toàn diện về lĩnh vực quan trọng này.

1. Giới thiệu về Đạo đức AI

Sự phát triển của AI mang lại tiềm năng to lớn nhưng cũng đặt ra nhiều lo ngại về mặt đạo đức, đặc biệt trong các hệ thống tự động hóa và thuật toán. Các vấn đề như thiên vị thuật toán, xâm phạm quyền riêng tư, và trách nhiệm giải trình đòi hỏi việc xây dựng các khuôn khổ đạo đức cho AI để đảm bảo AI phục vụ lợi ích nhân loại một cách công bằng và có trách nhiệm.

2. Định nghĩa và Các Nguyên tắc Cốt lõi của Đạo đức AI

Đạo đức AI là lĩnh vực nghiên cứu tập trung vào việc tối ưu hóa các tác động tích cực của AI đồng thời giảm thiểu rủi ro tiềm ẩn. Nó bao gồm việc thiết kế, phát triển, triển khai và sử dụng các hệ thống AI một cách có trách nhiệm, tập trung vào tính minh bạch, công bằng, khả năng giải thích, quyền riêng tư, an toàn và trách nhiệm giải trình.

Các Nguyên tắc Cốt lõi

UNESCO đã đưa ra **Mười nguyên tắc chính** cho đạo đức AI, bao gồm tôn trọng quyền con người và phẩm giá, xây dựng xã hội hòa bình, công bằng và kết nối, thúc đẩy đa

dạng và hòa nhập, bảo vệ môi trường và hệ sinh thái bền vững, đảm bảo tính tương xứng và không gây hại, an toàn và bảo mật, quyền riêng tư và bảo vệ dữ liệu, quản trị đa bên và hợp tác, trách nhiệm giải trình, minh bạch và khả năng giải thích. IBM và Báo cáo Belmont cũng nhấn mạnh các nguyên tắc tương tự như tôn trọng con người, lợi ích và công lý.

3. Các Thách thức Đạo đức Chính trong AI

Sự phát triển của AI đang đối mặt với nhiều thách thức đạo đức cần được giải quyết:

3.1. Thiên vị và Phân biệt đối xử

Nguy cơ AI khuếch đại và củng cố các thành kiến hiện có trong xã hội là một thách thức lớn. Các thuật toán AI học từ dữ liệu có thể tái tạo và khuếch đại thành kiến nếu dữ liệu đào tạo không công bằng. Ví dụ, hệ thống tuyển dụng AI của Amazon từng thiên vị ứng viên nam giới [8].

3.2. Quyền riêng tư và Giám sát

AI yêu cầu lượng lớn dữ liệu, làm tăng lo ngại về quyền riêng tư. Việc thu thập, lưu trữ và xử lý dữ liệu cá nhân có thể dẫn đến lạm dụng và vi phạm quyền riêng tư. Các quy định như GDPR và CCPA đã được ban hành, nhưng việc đảm bảo quyền riêng tư vẫn là một thách thức liên tục [8].

3.3. Khả năng giải thích và Minh bạch

Nhiều hệ thống AI, đặc biệt là mô hình học sâu, hoạt động như “hộp đen”, khiến việc hiểu cách chúng đưa ra quyết định trở nên khó khăn. Sự thiếu minh bạch và khả năng giải thích này gây khó khăn cho việc xác định nguyên nhân lỗi hoặc thành kiến, làm suy yếu niềm tin vào AI [8]. Điều này đặc biệt quan trọng trong các lĩnh vực nhạy cảm như y tế hoặc pháp luật.

3.4. Trách nhiệm giải trình

Việc xác định trách nhiệm khi AI gây hậu quả là một vấn đề phức tạp. Hiện tại, không có luật pháp phổ quát nào điều chỉnh các hoạt động AI, dẫn đến sự phân tán trách nhiệm [8].

3.5. Tác động đến việc làm

AI có khả năng tự động hóa nhiều công việc, gây lo ngại về mất việc làm. Tuy nhiên, AI cũng tạo ra công việc mới và nâng cao năng suất, đòi hỏi các chính sách và chương trình đào tạo để người lao động thích nghi [8].

4. Giải pháp và Khung quản trị Đạo đức AI

Để giải quyết các thách thức đạo đức của AI, cần có cách tiếp cận đa diện, bao gồm khung quản trị, chính sách và giáo dục. UNESCO và IBM đã đề xuất các giải pháp sau:

4.1. Quản trị AI

Quản trị AI bao gồm giám sát vòng đời AI, xác định vai trò, trách nhiệm, và thiết lập quy trình xây dựng, quản lý, giám sát hệ thống AI [8]. Các hội đồng đạo đức AI có thể cung cấp quản trị tập trung và ra quyết định cho các chính sách đạo đức [8]

4.2. Lĩnh vực Chính sách Hành động

UNESCO khuyến nghị mười một lĩnh vực chính sách hành động để phát triển AI có trách nhiệm, bao gồm quản trị dữ liệu, môi trường, giới tính, giáo dục, nghiên cứu, và sức khỏe [7]. Điều này nhấn mạnh việc chuyển đổi nguyên tắc đạo đức thành chiến lược thực tế.

4.3. Nâng cao Nhận thức và Hiểu biết

Thúc đẩy sự hiểu biết của công chúng về AI và dữ liệu là rất quan trọng, thông qua giáo dục mở, sự tham gia của công dân, kỹ năng số và đào tạo đạo đức AI.

4.4. Hợp tác Đa bên

Giải quyết thách thức đạo đức AI đòi hỏi hợp tác giữa chính phủ, ngành công nghiệp, học viện và xã hội dân sự. UNESCO nhấn mạnh quản trị đa bên để đảm bảo cách tiếp cận toàn diện [7].

5. Kết luận

Đạo đức AI là yêu cầu cấp bách để đảm bảo AI phục vụ lợi ích nhân loại. Tuân thủ các nguyên tắc cốt lõi như quyền con người, công bằng, minh bạch và trách nhiệm giải trình sẽ định hình sự phát triển AI có trách nhiệm và bền vững. Các khung quản trị toàn cầu và nỗ lực từ các tập đoàn đang đặt nền móng cho một tương lai AI phát huy tối đa tiềm năng mà không gây hại. Nghiên cứu, đối thoại và hợp tác là chìa khóa để xây dựng hệ sinh thái AI đạo đức.

Bảng Tổng hợp và Đánh giá Độ tin cậy của Nguồn thông tin

STT	Tên nguồn tài liệu	Tác giả/Cơ quan xuất bản	Năm xuất bản	Loại nguồn	Độ tin cậy	Ưu điểm	Nhược điểm
1	Khuyến nghị UNESCO về Đạo đức Trí tuệ Nhân tạo	Hà Quang Thụy et al. (Đại học Quốc gia Hà Nội)	Không rõ	Bài báo khoa học	Rất cao	Tác giả là chuyên gia đầu ngành, cơ quan xuất bản uy tín, phân tích sâu về khuyến nghị UNESCO và liên hệ Việt Nam.	Có thể mang tính chất nghiên cứu lý thuyết, ít ví dụ thực tiễn cụ thể.
2	The ethics of artificial intelligence	Nick Bostrom & Eliezer Yudkowsky	2018	Bài báo khoa học	Rất cao	Tác giả nổi tiếng thế giới về AI, trích dẫn lớn (>2500 lần), phân tích sâu về các vấn đề đạo đức khi tạo ra máy móc có khả năng tư duy.	Có thể tập trung vào các khía cạnh triết học, ít đi sâu vào ứng dụng thực tiễn.

STT	Tên nguồn tài liệu	Tác giả/Cơ quan xuất bản	Năm xuất bản	Loại nguồn	Độ tin cậy	Ưu điểm	Nhược điểm
3	The ethics of artificial intelligence: Principles, challenges, and opportunities	Luciano Floridi	2023	Sách/Bài báo	Rất cao	Tác giả là chuyên gia hàng đầu, cập nhật mới nhất, phân tích toàn diện các thách thức, nguyên tắc và cơ hội trong AI.	Có thể đòi hỏi kiến thức nền tảng nhất định để hiểu sâu.
4	Incorporating ethics into artificial intelligence	Amitai Etzioni & Oren Etzioni	2017	Bài báo khoa học	Cao	Tập trung vào cách tích hợp đạo đức vào hệ thống AI tự hành.	Có thể hơi cũ so với sự phát triển nhanh chóng của AI.

Tài liệu tham khảo

[7] Hà Quang Thụy, Phan Xuân Hiếu, Nguyễn Trí Thành, Trần Trọng Hiếu, Trần Mai Vũ, Lê Đức Trọng, Lê Hoàng Quỳnh. “Khuyến nghị đạo đức trí tuệ nhân tạo của UNESCO và liên hệ tới Việt Nam”. Đại học Quốc gia Hà Nội. [8] IBM. “What is AI Ethics?”. URL: https://www.ibm.com/thought/ai-ethics [9] Harvard Gazette. “Ethical concerns mount as AI takes bigger decision-making role”. 2020. URL: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/ [10] Coursera. “AI Ethics: What It Is, Why It Matters, and More”. 2025. URL: https://www.coursera.org/articles/ai-ethics [11] Tạp chí An toàn thông tin. “Đạo đức Trí tuệ nhân tạo: Thực trạng, thách thức và giải pháp về an toàn thông tin”. 2025. [12] Wikipedia. “Ethics of artificial intelligence”. URL: https://en.wikipedia.org/wiki/Ethics_of_artificial_intelligence