ĐỊNH KIẾN TRONG AI VÀ PHƯƠNG PHÁP QUẢN LÝ

Lê Thị Thanh Tuyền[(1)]
(1) Trường Đại học Thủ Dầu Một
Ngày nhận bài 30/10/2024; Chấp nhận đăng 20/11/2024
Liên hệ email: tuyenltt@tdmu.edu.vn

Tóm tắt

Định kiến trong trí tuệ nhân tạo (AI) đang trở thành một vấn đề quan trọng và cần thiết được xem xét nghiêm túc trong quá trình phát triển và ứng dụng các hệ thống thông minh. Với sự gia tăng của AI trong nhiều lĩnh vực từ y tế, giáo dục, tài chính đến pháp lý, khả năng những định kiến xã hội, văn hóa và kỹ thuật tiềm ẩn trong các hệ thống AI gây ra những hệ lụy đáng kể là điều không thể bỏ qua. Định kiến trong AI không đơn thuần là vấn đề kỹ thuật mà còn đặt ra thách thức lớn về tính công bằng, sự minh bạch và bảo vệ quyền lợi của con người, đặc biệt trong bối cảnh các ứng dụng AI đang ngày càng chi phối các quyết định quan trọng trong xã hội. Bài viết này sẽ tập trung vào phân tích các khía cạnh lý thuyết và thực tiễn liên quan đến định kiến trong AI, các quy định và tiêu chuẩn pháp lý hiện hành, cùng các phương pháp quản lý và giảm thiểu định kiến trong quá trình triển khai và ứng dụng AI.

Từ khóa: *các tiêu chuẩn pháp lý, định kiến, phương pháp quản lý*

Abstract

BIAS IN AI: LEGAL STANDARDS AND MANAGEMENT METHODS

Bias in artificial intelligence (AI) is becoming an important and necessary issue to be seriously considered in the development and application of intelligent systems. With the rise of AI in many fields from healthcare, education, finance to law, the possibility of social, cultural, and technical biases hidden in AI systems causing significant consequences cannot be ignored. Bias in AI is not simply a technical issue but also poses a major challenge to fairness, transparency, and protection of human rights, especially in the context of AI applications increasingly dominating important decisions in society. This article will focus on analyzing theoretical and practical aspects related to bias in AI, current regulations and legal standards, and methods of managing and minimizing bias in the process of deploying and applying AI.

1. Đặt vấn đề

Trong thời đại công nghệ phát triển nhanh chóng, trí tuệ nhân tạo (AI) đã và đang trở thành một phần không thể thiếu trong nhiều lĩnh vực của cuộc sống. Từ y tế, giáo dục đến dịch vụ tài chính, thương mại và đặc biệt là lĩnh vực pháp lý, AI không chỉ hỗ trợ tăng cường hiệu quả công việc mà còn mang đến nhiều tiềm năng thay đổi toàn diện cách thức mà con người tương tác với các hệ thống kỹ thuật số. Tuy nhiên, cùng với những tiến bộ vượt bậc của công nghệ, vấn đề về tính công bằng, sự minh bạch và trách nhiệm trong việc sử dụng AI ngày càng trở nên cấp thiết. Một trong những thách thức lớn nhất hiện nay là hiện tượng định kiến trong AI – tức là việc các hệ thống AI đưa ra

các quyết định thiên vị hoặc phân biệt đối xử dựa trên những yếu tố như giới tính, chủng tộc, độ tuổi, tôn giáo và nhiều yếu tố khác.

Định kiến trong AI có thể xuất phát từ nhiều nguyên nhân khác nhau, chẳng hạn như dữ liệu huấn luyện thiếu tính đa dạng, các thuật toán thiếu công bằng hoặc các sai lệch trong thiết kế hệ thống. Hệ quả là các quyết định của AI có thể làm sâu sắc thêm sự bất công xã hội, tạo ra những kết quả bất lợi cho những nhóm người dễ bị tổn thương và làm suy giảm niềm tin của xã hội vào công nghệ. Điều này đặc biệt quan trọng trong các lĩnh vực có tính chất quyết định cao như tư pháp, y tế, tuyển dụng, quyền riêng tư dữ liệu, nơi AI có thể ảnh hưởng trực tiếp đến cuộc sống và quyền lợi của con người.

Bên cạnh đó, các nhà lập pháp và các cơ quan quản lý đang đối mặt với thách thức lớn trong việc xây dựng các tiêu chuẩn pháp lý và quy định nhằm kiểm soát định kiến trong AI. Mặc dù một số khung pháp lý quốc tế đã bắt đầu xuất hiện, như EU AI Act hay Quy định bảo vệ dữ liệu chung (GDPR) của châu Âu, vẫn chưa có một hệ thống quy định toàn diện nào đảm bảo tính công bằng và minh bạch trong việc triển khai AI trên toàn cầu. Các nhà phát triển AI cần phải đối mặt với câu hỏi không chỉ về trách nhiệm pháp lý khi AI đưa ra các quyết định sai lệch, mà còn về cách làm sao để giảm thiểu rủi ro này thông qua các phương pháp quản lý kỹ thuật và pháp lý hiệu quả.

Từ thực trạng trên, câu hỏi nghiên cứu quan trọng đặt ra là: Làm thế nào để quản lý và giảm thiểu định kiến trong các hệ thống AI thông qua các tiêu chuẩn pháp lý và phương pháp quản lý phù hợp? Bài báo này sẽ tập trung thảo luận về khía cạnh pháp lý của định kiến trong AI, phân tích các tiêu chuẩn hiện hành và các phương pháp quản lý định kiến, đồng thời đề xuất các hướng đi khả thi nhằm giảm thiểu sự thiên lệch trong quá trình phát triển và triển khai AI. Việc hiểu rõ hơn về định kiến trong AI không chỉ giúp các nhà lập pháp xây dựng khung pháp lý bảo vệ quyền lợi công dân mà còn góp phần làm tăng tính minh bạch và công bằng của công nghệ trong bối cảnh xã hội ngày càng dựa vào trí tuệ nhân tạo để đưa ra các quyết định quan trọng.

2. Phương pháp nghiên cứu

Để đánh giá các vấn đề liên quan đến định kiến trong AI và các tiêu chuẩn pháp lý cũng như phương pháp quản lý, bài nghiên cứu này áp dụng phương pháp tiếp cận đa chiều, bao gồm các bước sau:

Phương pháp nghiên cứu tài liệu sẵn có

Nghiên cứu tài liệu được tiến hành để thu thập và phân tích các công trình đã công bố liên quan đến:

• Định kiến trong AI: Các nghiên cứu về định kiến và thiên lệch trong các hệ thống AI, bao gồm các ví dụ thực tế về cách các hệ thống AI đưa ra các quyết định phân biệt đối xử.

• Khung pháp lý: Các tài liệu về luật pháp và quy định liên quan đến AI và quyền bình đẳng, bao gồm các đạo luật quốc tế, quốc gia và các hướng dẫn từ các tổ chức như Liên minh châu Âu (EU AI Act) và Quy định bảo vệ dữ liệu chung (GDPR).

• Phương pháp giảm thiểu định kiến: Tổng hợp các phương pháp kiểm toán và công nghệ đã được đề xuất và triển khai nhằm giảm thiểu định kiến trong AI.

Nguồn tài liệu bao gồm các bài báo khoa học, báo cáo của các tổ chức chính phủ và phi chính phủ, cũng như các bài báo từ các tạp chí luật pháp và công nghệ. Việc

nghiên cứu tài liệu giúp tạo ra cái nhìn tổng quát về hiện trạng và thách thức của việc quản lý định kiến trong AI.

Phương pháp phân tích trường hợp

Bài báo sử dụng phân tích trường hợp để minh họa cụ thể về những tác động của định kiến trong AI. Các trường hợp này bao gồm:

• Hệ thống AI trong tuyển dụng: Nghiên cứu về các công ty sử dụng AI để lọc hồ sơ tuyển dụng và cách AI tạo ra sự phân biệt đối xử giữa các nhóm giới tính và chủng tộc.

• AI trong tư pháp: Trường hợp nổi bật về các hệ thống đánh giá rủi ro tội phạm dựa trên AI, trong đó AI có xu hướng đánh giá người da màu với nguy cơ cao hơn so với người da trắng trong cùng điều kiện.

• AI trong quảng cáo: Ví dụ về việc các hệ thống quảng cáo tự động ưu tiên hiển thị quảng cáo công việc hoặc sản phẩm cho một nhóm đối tượng cụ thể, gây ra sự bất bình đẳng trong cơ hội tiếp cận.

Các trường hợp này không chỉ minh họa cho các vấn đề lý thuyết về định kiến mà còn cung cấp bằng chứng thực tiễn về cách những thách thức này ảnh hưởng đến cuộc sống con người.

3. Kết quả và thảo luận

3.1. Định kiến trong AI: Hiện trạng và tác động

Định kiến AI đã trở thành một vấn đề nghiêm trọng khi ngày càng nhiều hệ thống AI được áp dụng rộng rãi trong các lĩnh vực như y tế, tài chính, tư pháp và tuyển dụng. Định kiến trong AI thường xuất hiện khi các mô hình AI học tập từ dữ liệu có chứa các thiên kiến xã hội, văn hóa hoặc lịch sử. Điều này có thể dẫn đến các kết quả không công bằng, ảnh hưởng tiêu cực đến những nhóm thiểu số hoặc dễ bị tổn thương. Trong phần này, chúng ta sẽ xem xét chi tiết ba khía cạnh chính: dữ liệu huấn luyện không đại diện, thuật toán thiếu minh bạch và tác động đến các nhóm dễ bị tổn thương.

3.1.1.Dữ liệu huấn luyện không đại diện

Một trong những nguyên nhân chính dẫn đến định kiến trong AI là việc sử dụng dữ liệu huấn luyện không đại diện cho toàn bộ dân số. Các hệ thống AI phụ thuộc vào dữ liệu được thu thập từ các nguồn khác nhau để học cách đưa ra quyết định. Tuy nhiên, nếu dữ liệu này thiên lệch, mô hình AI sẽ kế thừa và tái tạo những thiên kiến đó. Ví dụ, nghiên cứu của Buolamwini và Gebru (2018) về các hệ thống nhận dạng khuôn mặt cho thấy các mô hình này hoạt động kém hơn đáng kể đối với phụ nữ da màu so với nam giới da trắng, vì dữ liệu huấn luyện chủ yếu dựa trên các khuôn mặt của người da trắng.

Điều này không chỉ xảy ra trong nhận diện khuôn mặt mà còn trong nhiều lĩnh vực khác, chẳng hạn như tuyển dụng và tín dụng. Khi các mô hình AI được huấn luyện từ dữ liệu tuyển dụng lịch sử, nơi mà nam giới thường được ưu tiên, các hệ thống này có xu hướng tiếp tục ưu tiên nam giới trong các quyết định tuyển dụng tự động (Binns, 2020). Hậu quả là các ứng viên nữ có thể bị thiệt thòi và không được xem xét công bằng, dù họ có trình độ tương đương.

3.1.2.Thuật toán thiếu minh bạch

Một vấn đề nghiêm trọng khác là thiếu tính minh bạch trong các thuật toán AI. Nhiều hệ thống AI hiện nay, đặc biệt là những hệ thống dựa trên mô hình học sâu (deep learning), hoạt động như một "hộp đen" – người sử dụng không biết chính xác cách hệ thống đưa ra quyết định (Pasquale, 2015). Sự phức tạp của các mô hình này khiến việc kiểm soát và giải thích quyết định đầu ra trở nên khó khăn.

Thuật toán AI có thể tự động hóa những quyết định quan trọng mà không có cơ chế giải thích rõ ràng, dẫn đến những tác động tiêu cực không mong muốn, nhất là trong các lĩnh vực như tư pháp và bảo hiểm. Ví dụ, trong các hệ thống AI dùng để dự đoán rủi ro tội phạm, như hệ thống COMPAS được sử dụng ở Mỹ, đã có những trường hợp hệ thống đánh giá người da màu với nguy cơ tái phạm cao hơn so với người da trắng, ngay cả khi các yếu tố khác tương đương (Angwin và cs., 2016). Những quyết định này ảnh hưởng trực tiếp đến quyền tự do của con người, đặc biệt là đối với các nhóm dân số dễ bị tổn thương.

3.1.3.Tác động lên các nhóm dễ bị tổn thương

Nhiều nghiên cứu đã chỉ ra rằng những nhóm người thiểu số hoặc dễ bị tổn thương là những người chịu ảnh hưởng nhiều nhất từ định kiến trong AI. Trong các hệ thống AI liên quan đến chăm sóc sức khỏe, dữ liệu y tế thường phản ánh tình trạng không đồng đều trong việc tiếp cận dịch vụ y tế giữa các nhóm dân tộc và kinh tế khác nhau. Khi các hệ thống AI dự đoán chẩn đoán hoặc điều trị dựa trên dữ liệu này, các nhóm thiểu số có thể nhận được các quyết định điều trị không chính xác hoặc không đầy đủ (Obermeyer và cs., 2019). Điều này làm gia tăng sự chênh lệch về sức khỏe giữa các nhóm dân cư.

Trong lĩnh vực tài chính, định kiến trong AI cũng có thể dẫn đến việc từ chối tín dụng một cách không công bằng. Các hệ thống đánh giá tín dụng dựa trên lịch sử tài chính có thể phân biệt đối xử với những người thuộc tầng lớp kinh tế thấp, hoặc các cộng đồng dân tộc thiểu số, dẫn đến việc họ khó tiếp cận với các dịch vụ tài chính quan trọng (Mehrabi và cs., 2021). Điều này không chỉ ảnh hưởng đến cuộc sống hàng ngày của họ mà còn làm gia tăng khoảng cách giàu nghèo trong xã hội.

Định kiến trong AI hiện nay không chỉ là một vấn đề kỹ thuật mà còn là một vấn đề xã hội và đạo đức. Nó có thể dẫn đến những quyết định bất công và làm tổn hại đến những nhóm dân cư dễ bị tổn thương. Để giải quyết vấn đề này, cần có sự can thiệp mạnh mẽ từ cả khía cạnh kỹ thuật và pháp lý nhằm đảm bảo rằng các hệ thống AI được phát triển và vận hành một cách công bằng, minh bạch và có trách nhiệm.

3.2. Hiệu quả của các phương pháp quản lý định kiến

Với sự phát triển mạnh mẽ của AI, nhiều biện pháp đã được đề xuất nhằm giảm thiểu định kiến trong các hệ thống AI và đảm bảo tính công bằng trong các quyết định tự động. Những phương pháp này bao gồm kiểm toán thuật toán, cải thiện tính đa dạng của dữ liệu huấn luyện, tăng cường tính minh bạch và giải trình của các hệ thống AI, cũng như sự can thiệp từ các cơ quan pháp lý thông qua các quy định. Mỗi phương pháp mang lại những hiệu quả nhất định trong việc phát hiện và giảm thiểu các định kiến tiềm ẩn, tuy nhiên, chúng cũng đòi hỏi sự kết hợp chặt chẽ giữa các bên liên quan.

3.2.1.Kiểm toán thuật toán

Kiểm toán thuật toán là một trong những phương pháp hiệu quả nhất để phát hiện và quản lý các định kiến trong AI. Việc kiểm toán giúp xác định các yếu tố gây thiên lệch trong cách hệ thống AI đưa ra quyết định, đồng thời cung cấp thông tin chi tiết về các sai lệch có thể xảy ra trong suốt quá trình xử lý dữ liệu và ra quyết định.

Nghiên cứu của Raji và cs. (2020) đã chỉ ra rằng các hệ thống AI được kiểm toán định kỳ có xu hướng giảm thiểu các sai lệch so với các hệ thống không được kiểm tra. Quá trình kiểm toán bao gồm việc phân tích cách thuật toán đưa ra quyết định, kiểm tra sự phân biệt đối xử đối với các nhóm thiểu số và đảm bảo rằng các mô hình không làm trầm trọng thêm các định kiến xã hội. Ngoài ra, kiểm toán còn giúp đảm bảo rằng các hệ thống AI tuân thủ các tiêu chuẩn đạo đức và pháp lý hiện hành.

Mặc dù kiểm toán có hiệu quả trong việc phát hiện các định kiến, nhưng nó cũng đòi hỏi nguồn lực lớn và sự hợp tác giữa các nhà phát triển AI và cơ quan quản lý. Các hệ thống AI phức tạp như deep learning, cần được kiểm tra kỹ lưỡng để đảm bảo rằng mọi quy trình ra quyết định đều được giám sát và kiểm soát. Tuy nhiên, một thách thức lớn là các công ty công nghệ thường miễn cưỡng chia sẻ thông tin chi tiết về các thuật toán của họ, làm cho việc kiểm toán trở nên khó khăn hơn.

3.2.2. Tính đa dạng trong dữ liệu huấn luyện

Một trong những nguyên nhân chính dẫn đến định kiến trong AI là dữ liệu huấn luyện không đại diện. Do đó, việc cải thiện tính đa dạng trong dữ liệu huấn luyện là một biện pháp thiết yếu để giảm thiểu định kiến. Nếu dữ liệu huấn luyện không bao gồm các đại diện đầy đủ và cân bằng của các nhóm dân số khác nhau, hệ thống AI có thể đưa ra các quyết định thiên lệch và gây phân biệt đối xử.

Nghiên cứu của Mehrabi và cs. (2021) đã chỉ ra rằng khi các hệ thống AI được huấn luyện với các bộ dữ liệu đa dạng, chúng có xu hướng hoạt động công bằng hơn và giảm thiểu nguy cơ thiên lệch. Ví dụ, các hệ thống nhận diện khuôn mặt, khi được huấn luyện với dữ liệu bao gồm đầy đủ các sắc tộc và giới tính, đã cho thấy sự cải thiện đáng kể về độ chính xác và tính công bằng.

Tuy nhiên, việc thu thập và sử dụng dữ liệu đa dạng không phải lúc nào cũng dễ dàng. Nhiều quốc gia có quy định nghiêm ngặt về quyền riêng tư, điều này làm hạn chế khả năng thu thập dữ liệu về chủng tộc, giới tính, hoặc các đặc điểm nhạy cảm khác. Hơn nữa, trong nhiều trường hợp, dữ liệu đại diện không tồn tại hoặc rất khó tiếp cận, khiến việc huấn luyện mô hình AI trở nên phức tạp và tốn kém hơn.

3.2.3.Cải thiện tính minh bạch và giải trình

Minh bạch trong các hệ thống AI là yếu tố quan trọng để phát hiện và ngăn chặn định kiến. Các hệ thống AI có tính minh bạch cao sẽ cung cấp thông tin rõ ràng về cách thức hoạt động của thuật toán, từ dữ liệu đầu vào, quá trình ra quyết định đến kết quả đầu ra. Điều này cho phép các nhà phát triển, người dùng và cơ quan quản lý có thể theo dõi và kiểm tra sự công bằng của hệ thống.

Nghiên cứu của Doshi-Velez và Kim (2017) nhấn mạnh rằng các hệ thống AI có tính giải trình tốt giúp người dùng hiểu rõ hơn về quá trình ra quyết định, từ đó có thể can thiệp khi nhận thấy dấu hiệu của sự thiên lệch. Các mô hình AI có khả năng giải thích rõ ràng và dễ hiểu về quyết định của chúng sẽ giảm thiểu rủi ro của các quyết định thiên lệch và bất công. Minh bạch và giải trình không chỉ giúp cải thiện tính công bằng mà còn tăng cường niềm tin của người dùng đối với các hệ thống AI.

Tuy nhiên, một thách thức lớn là các hệ thống AI hiện đại, đặc biệt là deep learning, thường rất phức tạp và khó giải thích. Nhiều hệ thống AI hoạt động như một "hộp đen", khiến việc hiểu và giải thích quá trình ra quyết định trở nên khó khăn đối với người ngoài cuộc. Do đó, việc cải thiện tính minh bạch đòi hỏi các nhà phát triển AI không chỉ cung cấp kết quả đầu ra mà còn phải giải thích rõ ràng quá trình ra quyết định và các yếu tố ảnh hưởng đến kết quả.

3.2.4. Hướng dẫn từ chính phủ và các tiêu chuẩn pháp lý

Các quy định pháp lý đóng vai trò quan trọng trong việc quản lý định kiến trong AI. Các tiêu chuẩn pháp lý không chỉ thiết lập các nguyên tắc chung về quyền con người, tính công bằng và đạo đức mà còn buộc các nhà phát triển AI phải tuân thủ các quy định nghiêm ngặt về việc sử dụng dữ liệu và ra quyết định tự động. Ví dụ, Quy định Bảo vệ Dữ liệu Chung (GDPR) của Liên minh châu Âu đã đưa ra các hướng dẫn rõ ràng về cách các tổ chức sử dụng dữ liệu cá nhân trong các hệ thống tự động.

Nghiên cứu của Veale và Borgesius (2021) về Đạo luật AI của Liên minh châu Âu đã chỉ ra rằng các quy định pháp lý có thể tạo ra một khung quản lý mạnh mẽ, giúp giảm thiểu rủi ro của định kiến trong AI. Đạo luật này yêu cầu các hệ thống AI có rủi ro cao phải tuân thủ các tiêu chuẩn về tính minh bạch, công bằng và trách nhiệm giải trình. Những hướng dẫn từ chính phủ không chỉ bảo vệ quyền lợi của cá nhân mà còn khuyến khích các tổ chức phát triển các hệ thống AI một cách có trách nhiệm hơn.

Mặc dù các quy định pháp lý mang lại hiệu quả rõ ràng, nhưng việc thực thi chúng cũng đối mặt với nhiều thách thức. Tốc độ phát triển nhanh chóng của công nghệ AI đòi hỏi các nhà lập pháp phải liên tục cập nhật và điều chỉnh các quy định để theo kịp những thay đổi. Đồng thời, việc đảm bảo tuân thủ pháp lý yêu cầu sự hợp tác chặt chẽ giữa các cơ quan quản lý, tổ chức phát triển công nghệ và cộng đồng người dùng.

Các phương pháp quản lý định kiến trong AI, từ kiểm toán thuật toán đến các hướng dẫn pháp lý, đều đóng vai trò quan trọng trong việc đảm bảo rằng các hệ thống AI hoạt động một cách công bằng và không thiên lệch. Tuy nhiên, hiệu quả của các phương pháp này phụ thuộc vào sự kết hợp chặt chẽ giữa các bên liên quan, từ nhà phát triển công nghệ, cơ quan quản lý đến người dùng. Việc quản lý định kiến trong AI không chỉ là một thách thức kỹ thuật mà còn là một trách nhiệm xã hội và đạo đức, đòi hỏi sự chú trọng đặc biệt đến các yếu tố công bằng, minh bạch và trách nhiệm giải trình.

3.3. Phương pháp quản lý và giảm thiểu định kiến trong AI

Các phương pháp quản lý và giảm thiểu định kiến trong AI đóng vai trò then chốt trong việc đảm bảo rằng hệ thống AI hoạt động một cách công bằng và không gây ra hậu quả tiêu cực. Những phương pháp này tập trung vào từng giai đoạn của quá trình phát triển AI, từ thu thập dữ liệu, xây dựng thuật toán, đến giám sát hiệu suất trong quá trình sử dụng. Sau đây là một số phương pháp quan trọng để giảm thiểu và quản lý định kiến trong các hệ thống AI.

3.3.1. Lựa chọn và xử lý dữ liệu đa dạng

Dữ liệu là nền tảng của bất kỳ hệ thống AI nào và do đó, việc đảm bảo dữ liệu đầu vào đa dạng, đại diện và không chứa định kiến là một bước cực kỳ quan trọng. Nhiều nghiên cứu chỉ ra rằng định kiến trong dữ liệu đầu vào sẽ trực tiếp ảnh hưởng đến kết quả của AI (Buolamwini & Gebru, 2018). Ví dụ, các hệ thống nhận diện khuôn mặt trước đây thường mắc lỗi cao hơn đối với phụ nữ và người da màu do dữ liệu huấn luyện thiếu đa dạng về chủng tộc và giới tính. Để giảm thiểu tình trạng này, các nhà phát triển cần đảm bảo rằng dữ liệu được thu thập từ nhiều nguồn khác nhau, phản ánh đầy đủ các nhóm dân số mà hệ thống AI sẽ phục vụ.

Xử lý và làm sạch dữ liệu là một công đoạn không thể thiếu nhằm loại bỏ các yếu tố có khả năng gây định kiến. Obermeyer và cs. (2019), trong lĩnh vực y tế, khi dữ liệu đào tạo được xây dựng trên cơ sở chi phí y tế trong quá khứ, hệ thống AI có thể đưa ra các dự đoán thiên lệch vì chi phí y tế không phải lúc nào cũng phản ánh đúng nhu cầu

chăm sóc sức khỏe. Do đó, việc chuẩn hóa dữ liệu và loại bỏ các biến thiên lệch là cần thiết để giảm thiểu định kiến.

3.3.2. Xây dựng thuật toán công bằng

Các kỹ thuật học máy công bằng (fair machine learning) đã được phát triển nhằm đối phó với định kiến trong các mô hình AI. Một số phương pháp như việc sử dụng thuật toán điều chỉnh các biến độc lập với đặc điểm cá nhân (như giới tính, chủng tộc) giúp mô hình giảm thiểu các kết quả thiếu công bằng (Barocas, Hardt, & Narayanan, 2019). Chẳng hạn, kỹ thuật "fair representation learning" tập trung vào việc xây dựng các biểu diễn của dữ liệu mà không chứa đựng những thông tin dễ gây định kiến, giúp hệ thống đưa ra các quyết định ít thiên lệch hơn.

Một phương pháp khác là tối ưu hóa có ràng buộc (constrained optimization), trong đó hệ thống AI được thiết kế để đáp ứng các tiêu chí công bằng xác định trước trong quá trình tối ưu hóa (Hardt, Price, & Srebro, 2016). Điều này giúp đảm bảo rằng các mô hình không bị lệch lạc trong việc đánh giá hoặc đưa ra quyết định đối với các nhóm dân số khác nhau. Ví dụ, trong tuyển dụng, hệ thống có thể được lập trình để đảm bảo rằng tỷ lệ giới tính hoặc chủng tộc trong các quyết định tuyển dụng phản ánh công bằng xã hội.

3.3.3. Giám sát và kiểm tra định kỳ

Giám sát và đánh giá định kỳ hiệu suất của hệ thống AI là một phần không thể thiếu nhằm duy trì tính công bằng trong suốt vòng đời của mô hình. Nhiều tổ chức đã thiết lập các quy trình kiểm tra định kỳ để phát hiện sớm và điều chỉnh các dấu hiệu định kiến trong hệ thống AI (Mitchell và cs., 2019). Kiểm tra định kỳ bao gồm việc đánh giá hiệu suất của hệ thống với các chỉ số công bằng, độ chính xác và tính minh bạch. Chẳng hạn, trong hệ thống tín dụng, các mô hình AI có thể được kiểm tra xem có xu hướng từ chối tín dụng không công bằng đối với các nhóm sắc tộc hoặc giới tính cụ thể hay không.

Một số công cụ kiểm tra công bằng đã được phát triển để hỗ trợ quá trình này, chẳng hạn như công cụ "Fairness Indicators" của Google AI, giúp các tổ chức theo dõi hiệu suất công bằng của mô hình qua các nhóm nhân khẩu học khác nhau. Phân tích định kỳ như vậy đảm bảo rằng các mô hình AI không tự động tích lũy các sai lệch trong thời gian dài và giúp duy trì sự công bằng ngay cả khi các điều kiện môi trường thay đổi.

3.3.4. Đào tạo và tăng cường nhận thức về AI công bằng

Giáo dục và đào tạo về các nguyên tắc công bằng trong AI là một yếu tố quan trọng để phát triển AI có trách nhiệm. Việc tăng cường nhận thức về tính công bằng giúp các nhà phát triển, nhà quản lý và người sử dụng nhận thức rõ ràng về những ảnh hưởng tiêu cực tiềm ẩn của định kiến trong AI. Buolamwini và Gebru (2018) nhấn mạnh rằng các nhà phát triển có thể không nhận ra tác động định kiến của hệ thống mà họ xây dựng nếu không được trang bị kiến thức về công bằng và đạo đức AI.

Đào tạo về công bằng trong AI cũng giúp các nhà phát triển làm quen với các công cụ và phương pháp kiểm tra định kiến, từ đó nâng cao khả năng phát hiện và giải quyết các vấn đề định kiến. Nhiều tổ chức và trường đại học hiện đang tích hợp các khóa học về AI có trách nhiệm trong chương trình đào tạo, giúp chuẩn bị một thế hệ chuyên gia AI với ý thức sâu sắc về trách nhiệm xã hội và đạo đức.

3.4. Thảo luận

Trong phần này, chúng ta sẽ thảo luận về những kết quả chính từ nghiên cứu, đồng thời đánh giá những thành công và hạn chế của các phương pháp quản lý định kiến trong AI. Phần thảo luận cũng sẽ tập trung vào vai trò của các bên liên quan, bao gồm các nhà phát triển công nghệ, các cơ quan quản lý pháp luật và xã hội trong việc quản lý và giảm thiểu định kiến. Ngoài ra, những thách thức và hướng đi tương lai sẽ được đề xuất nhằm đảm bảo các hệ thống AI phát triển theo hướng công bằng và minh bạch hơn.

3.4.1.Hiệu quả của các phương pháp kỹ thuật

Các phương pháp kỹ thuật để giảm thiểu định kiến, như đã được trình bày, bao gồm kiểm toán thuật toán, cải thiện tính đa dạng trong dữ liệu huấn luyện và tăng cường tính minh bạch của hệ thống AI. Những phương pháp này đã cho thấy nhiều tín hiệu tích cực trong việc giảm thiểu các tác động tiêu cực của định kiến trong AI.

Kiểm toán thuật toán, khi được thực hiện đúng cách, giúp phát hiện ra những điểm yếu của hệ thống và cung cấp các biện pháp khắc phục. Một ví dụ điển hình là nghiên cứu của Raji và cs. (2020), nơi kiểm toán thuật toán đã phát hiện ra các điểm yếu trong các hệ thống nhận diện khuôn mặt và đề xuất các phương pháp cải tiến. Tuy nhiên, một trong những hạn chế lớn nhất của việc kiểm toán là yêu cầu rất cao về nguồn lực, bao gồm cả tài chính lẫn nhân lực. Đồng thời, do tính chất phức tạp và "hộp đen" của nhiều hệ thống AI hiện đại, đặc biệt là các mô hình học sâu, việc giải thích và hiểu các quyết định của chúng cũng trở thành một thách thức lớn (Doshi-Velez & Kim, 2017).

Cải thiện tính đa dạng trong dữ liệu huấn luyện là một trong những biện pháp hiệu quả để đảm bảo các mô hình AI không bị thiên lệch. Khi hệ thống được huấn luyện trên các dữ liệu đa dạng và cân bằng, nó có khả năng đưa ra các quyết định công bằng hơn cho mọi nhóm dân số. Tuy nhiên, một vấn đề nổi lên là việc thu thập dữ liệu đa dạng không phải lúc nào cũng khả thi do các quy định pháp lý về quyền riêng tư và bảo mật. Ví dụ, Quy định Bảo vệ Dữ liệu Chung (GDPR) của Liên minh châu Âu giới hạn việc thu thập và sử dụng dữ liệu cá nhân, làm giảm khả năng tiếp cận dữ liệu về chủng tộc, giới tính, hoặc các yếu tố nhạy cảm khác (Veale & Borgesius, 2021).

Ngoài ra, tính minh bạch và giải trình của hệ thống AI cũng đóng vai trò quan trọng trong việc phát hiện và ngăn chặn định kiến. Khi các hệ thống AI minh bạch, người dùng và cơ quan giám sát có thể theo dõi và hiểu rõ cách các quyết định được đưa ra, từ đó dễ dàng phát hiện các thiên lệch tiềm ẩn. Tuy nhiên, các hệ thống AI hiện đại thường quá phức tạp để người dùng thông thường có thể hiểu và đánh giá. Đây là một vấn đề đáng lưu tâm, khi các mô hình AI cần phải vừa mạnh mẽ vừa dễ tiếp cận trong việc giải thích các kết quả của mình.

3.4.2. Vai trò của quy định pháp lý và chính phủ

Quy định pháp lý đóng một vai trò không thể thiếu trong việc đảm bảo rằng các hệ thống AI được phát triển và sử dụng một cách công bằng và có trách nhiệm. Ví dụ, Đạo luật AI của Liên minh châu Âu (EU AI Act) đưa ra các tiêu chuẩn nghiêm ngặt cho các hệ thống AI có rủi ro cao, yêu cầu chúng phải tuân thủ các tiêu chuẩn về tính công bằng, minh bạch và giải trình (Veale & Borgesius, 2021). Những quy định này tạo ra một khung pháp lý rõ ràng, giúp quản lý rủi ro từ các hệ thống AI và đảm bảo quyền lợi của người dân.

Tuy nhiên, quy định pháp lý cũng đối mặt với nhiều thách thức. Đầu tiên, tốc độ phát triển của công nghệ AI thường vượt qua tốc độ ra quyết định và triển khai các quy

định. Các nhà lập pháp thường gặp khó khăn trong việc theo kịp những thay đổi nhanh chóng của công nghệ, đặc biệt khi các hệ thống AI ngày càng phức tạp và tích hợp sâu vào nhiều lĩnh vực của xã hội. Điều này đòi hỏi các cơ quan pháp luật phải không ngừng điều chỉnh và cập nhật quy định để đảm bảo rằng luật pháp luôn phù hợp và hiệu quả.

Hơn nữa, mặc dù các quy định như GDPR và EU AI Act giúp bảo vệ quyền lợi của cá nhân, chúng cũng có thể làm giảm tính sáng tạo và tốc độ phát triển của công nghệ. Các công ty công nghệ có thể gặp khó khăn trong việc tuân thủ các quy định nghiêm ngặt, dẫn đến sự trì hoãn trong việc triển khai các sản phẩm mới. Do đó, một sự cân bằng giữa quản lý và phát triển là điều cần thiết để đảm bảo rằng AI phát triển theo hướng có lợi cho xã hội mà không bị cản trở quá mức bởi các rào cản pháp lý.

3.4.3. Sự cần thiết của hợp tác đa chiều

Để giải quyết hiệu quả vấn đề định kiến trong AI, cần có sự hợp tác chặt chẽ giữa các bên liên quan, bao gồm các nhà phát triển AI, cơ quan quản lý và xã hội. Sự hợp tác này cần phải mang tính toàn diện, trong đó mỗi bên có vai trò và trách nhiệm riêng trong việc giám sát, phát hiện và giải quyết các vấn đề về định kiến trong AI.

Các nhà phát triển công nghệ cần có trách nhiệm đạo đức và xã hội trong quá trình phát triển các hệ thống AI. Họ phải đảm bảo rằng các mô hình AI không chỉ tối ưu hóa về mặt kỹ thuật mà còn tuân thủ các tiêu chuẩn về tính công bằng và minh bạch. Điều này đòi hỏi các nhà phát triển phải áp dụng các phương pháp kiểm toán và đánh giá thường xuyên, cũng như hợp tác với các chuyên gia về đạo đức và pháp lý để đảm bảo rằng các hệ thống AI không gây ra các hậu quả không mong muốn.

Các cơ quan quản lý cần tiếp tục đóng vai trò quan trọng trong việc giám sát và thực thi các quy định pháp lý liên quan đến AI. Đồng thời, họ cần phải thúc đẩy việc giáo dục và tăng cường nhận thức cho công chúng về các nguy cơ và lợi ích của AI, nhằm xây dựng một môi trường mà người dân có thể tự tin sử dụng và hưởng lợi từ các hệ thống này mà không lo ngại về các vấn đề thiên lệch.

Xã hội và người dùng cũng có vai trò quan trọng trong việc giám sát và phản hồi về các hệ thống AI mà họ sử dụng. Sự tham gia của công chúng trong quá trình phát triển và giám sát AI sẽ giúp đảm bảo rằng các hệ thống này đáp ứng được nhu cầu thực tế của xã hội, đồng thời phát hiện và ngăn chặn các vấn đề tiềm ẩn về định kiến ngay từ giai đoạn sớm.

3.4.4.Thách thức và hướng đi tương lai

Mặc dù đã có nhiều tiến bộ trong việc quản lý định kiến trong AI, vẫn còn nhiều thách thức mà chúng ta cần phải giải quyết. Các hệ thống AI hiện đại, đặc biệt là những hệ thống dựa trên mô hình học sâu (deep learning), vẫn còn rất phức tạp và khó giải thích. Điều này gây khó khăn cho việc đảm bảo tính minh bạch và giải trình, dẫn đến những lo ngại về việc hệ thống có thể đưa ra các quyết định thiên lệch mà không thể dễ dàng giải thích được.

Bên cạnh đó, tốc độ phát triển nhanh chóng của AI đòi hỏi các nhà lập pháp phải linh hoạt và sáng tạo hơn trong việc thiết lập các quy định mới. Việc xây dựng một khung pháp lý toàn cầu cũng là một thách thức lớn, vì mỗi quốc gia có thể có những quy định và tiêu chuẩn khác nhau về việc sử dụng AI. Điều này đặt ra câu hỏi về sự cần thiết của một bộ quy tắc quốc tế thống nhất, nhằm đảm bảo rằng AI phát triển một cách công bằng và bền vững trên toàn cầu.

Trong tương lai, việc tích hợp AI vào mọi khía cạnh của cuộc sống sẽ đòi hỏi một cách tiếp cận đa chiều và toàn diện hơn. Điều này bao gồm việc phát triển các công nghệ AI có khả năng giải thích tốt hơn, cải thiện dữ liệu huấn luyện và tăng cường tính minh bạch trong quy trình phát triển AI. Đồng thời, cần tiếp tục thúc đẩy sự hợp tác giữa các tổ chức công nghệ, cơ quan quản lý và xã hội để đảm bảo rằng AI phát triển theo hướng mang lại lợi ích lớn nhất cho tất cả các nhóm dân cư, đặc biệt là những nhóm dễ bị tổn thương.

4. Kết luận

Trong bối cảnh AI ngày càng trở thành một phần không thể thiếu trong nhiều lĩnh vực của cuộc sống, vấn đề định kiến trong AI nổi lên như một thách thức đáng kể về mặt đạo đức, xã hội và pháp lý. Định kiến trong AI, khi không được kiểm soát, có thể dẫn đến các quyết định bất công, ảnh hưởng nghiêm trọng đến các nhóm dễ bị tổn thương và làm gia tăng sự bất bình đẳng trong xã hội. Do đó, việc nhận diện, ngăn chặn và giảm thiểu các định kiến trong AI là một nhiệm vụ cấp thiết đối với các nhà phát triển công nghệ, cơ quan quản lý và toàn xã hội.

Các phương pháp quản lý định kiến, từ kiểm toán thuật toán, cải thiện tính đa dạng trong dữ liệu huấn luyện, đến việc tăng cường tính minh bạch và giải trình của hệ thống AI, đã cho thấy nhiều hiệu quả đáng kể trong việc giảm thiểu các tác động tiêu cực. Đồng thời, sự can thiệp từ các quy định pháp lý, chẳng hạn như Đạo luật AI của Liên minh châu Âu, đóng vai trò quan trọng trong việc xây dựng một khuôn khổ pháp lý vững chắc để kiểm soát định kiến trong AI. Tuy nhiên, những biện pháp này không phải là giải pháp hoàn hảo, bởi vẫn còn nhiều thách thức cần phải vượt qua, bao gồm việc giám sát các hệ thống AI phức tạp và đảm bảo quyền riêng tư cho người dùng.

Tương lai của AI sẽ phụ thuộc vào cách chúng ta xử lý và quản lý các định kiến tồn tại trong các hệ thống này. Một sự kết hợp hài hòa giữa các biện pháp kỹ thuật và quy định pháp lý, cùng với việc hợp tác chặt chẽ giữa các bên liên quan, sẽ là chìa khóa để xây dựng một môi trường AI công bằng và có trách nhiệm. Bằng cách đảm bảo rằng các hệ thống AI hoạt động minh bạch, giải trình và không thiên lệch, chúng ta có thể tối ưu hóa lợi ích của AI đồng thời bảo vệ quyền lợi của tất cả các nhóm dân số, đặc biệt là những nhóm dễ bị tổn thương nhất trong xã hội.

TÀI LIỆU THAM KHẢO

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it’s biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 514-524. Association for Computing Machinery (ACM). https://doi.org/10.1145/3287560.3287572

[3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1-15. http://proceedings.mlr.press/v81/buolamwini18a.html

[4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

[5] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35. https://doi.org/10.1145/3457607

[6] Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

[7] Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2020). You can't sit with us: Exclusionary pedagogy in AI ethics education. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 61-71. Association for Computing Machinery (ACM). https://doi.org/10.1145/3351095.3372866

[8] Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International*, *22*(4), 97-112. https://doi.org/10.9785/cri-2021-220402

[9] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77-91.

[10] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315-3323.

[11] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229.

[12] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.