BÀN VỀ PHÁT TRIỂN QUY PHẠM ĐẠO ĐỨC TRONG TIẾN TRÌNH XÂY DỰNG PHÁP LUẬT VỀ TRÍ TUỆ NHÂN TẠO

TRẦN VIỆT DŨNG
Trường Đại học Luật TP.HCM
Ho Chi Minh City University of Law
Email: tvdung@hcmulaw.edu.vn

Tóm tắt

Trí tuệ nhân tạo (artificial intelligence, AI) đang dần thâm nhập sâu rộng vào đời sống xã hội toàn cầu, với tiềm năng cách mạng hóa nhiều ngành công nghiệp như chăm sóc sức khỏe, tài chính, và giáo dục. Tuy nhiên, việc triển khai AI rộng rãi đặt ra các vấn đề đạo đức như xâm phạm quyền riêng tư, thiên lệch và thông tin giả mạo, đòi hỏi sự điều chỉnh pháp luật. Trong bối cảnh này, các quy tắc đạo đức đóng vai trò quan trọng, tạo nền tảng cho một khung pháp luật linh hoạt và đáng tin cậy về AI, có thể điều chỉnh kịp thời theo sự phát triển của công nghệ.

Từ khóa: *trí tuệ nhân tạo, quy phạm đạo đức AI, quy định pháp luật, bộ quy tắc đạo đức, luật mềm*

Abstract

Artificial intelligence (AI) is gradually permeating global society, with the potential to revolutionize various industries such as healthcare, finance, and education. However, the widespread deployment of AI raises ethical concerns, including privacy invasion, bias, and misinformation, which necessitate legal regulation. In this context, ethical guidelines play a crucial role, forming the foundation for a flexible and reliable legal framework for AI that can be promptly adjusted in line with technological advancements.

Keywords: *artificial intelligence, AI ethical standards, legal regulations, code of ethics, soft law*
Ngày nhận bài: 15/04/2024 **Ngày duyệt đăng: 17/06/2024**

Trí tuệ nhân tạo (*artificial intelligence*, AI) được nhìn nhận như một "cuộc cách mạng" đang diễn ra trong lòng xã hội hiện đại, làm thay đổi về cốt lõi nhiều quan hệ xã hội và trật tự truyền thống. Sự phát triển nhanh chóng của công nghệ AI đang có tác động sâu sắc đến mọi lĩnh vực của đời sống và hoạt động con người, bao gồm giáo dục, khoa học, văn hóa, hệ thống tư pháp và thực hành pháp lý, y học, công nghiệp và giao thông vận tải, hoạt động kinh doanh, truyền thông và thông tin… Điều này đặt ra mối quan tâm về các vấn đề đạo đức và pháp lý liên quan tới sự phát triển của công nghệ AI và tạo ra các cuộc thảo luận khoa học sôi nổi trên phạm vi toàn cầu.[1] Do đó, pháp luật không chỉ cần phải thay đổi về căn bản để điều chỉnh các vấn đề pháp lý phát sinh từ khả năng tự tư duy và hành động của máy móc, mà còn phải có khả năng thích ứng và cập nhật liên tục để đáp ứng những tiến bộ mới nhất của công nghệ AI. Trong bối cảnh đó, cách giải quyết hiệu quả và thiết thực nhất là xây dựng các luật mềm dưới dạng nguyên tắc và tiêu chuẩn đạo đức trong phát triển và ứng dụng AI.[2]

1 K. Siau, Weiyu Wang, "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI", *Journal of Database Management*, No. 31(2), 2020, tr. 74-87; Weiping Sun, "Artificial Intelligence and Ethical Principles", in D. Jin (Ed.), *Reconstructing Our Orders: Artificial Intelligence and Human Society*, Springer, 2018, tr. 29–59.

2 Попова Анна, "Этические принципы взаимодействия с искусственным интеллектом как основа правового регулирования" [trans: Popova Anna, "Sự tương tác của các nguyên tắc đạo đức và trí tuệ nhân tạo như một nền tảng cho quy định pháp luật"], Правовое государство: теория и практика, Vol. 3 (61), 2020, tr. 34–43.

1. Khái quát về trí tuệ nhân tạo và những vấn đề đạo đức trong ứng dụng trí tuệ nhân tạo

AI là một lĩnh vực khoa học máy tính rộng lớn liên quan đến việc tạo ra những cỗ máy thông minh có khả năng thực hiện các hoạt động thường đòi hỏi trí thông minh của con người và thậm chí vượt qua khả năng phân tích trí óc của con người.[3] Cùng với sự bùng nổ của cuộc các mạng công nghệ 4.0, các công nghệ mới của AI như học máy (*machine learning*) và học sâu (*deep learning*) cho phép hệ thống AI tự học hỏi qua xử lý dữ liệu lớn (*big data*), tìm kiếm và lập mô hình cho việc ra quyết định nhanh chóng và thậm chí có thể tự học mà không cần giám sát của con người.[4] Từ ô tô tự hành (*autonomous vehicle*) cho đến các trợ lý thông minh như ChatGPT, từ robot AI phẫu thuật tại bệnh viện cho tới hệ thống phân tích thanh toán điện tử trong ngân hàng… các máy móc thiết bị có tích hợp AI đang từng bước chiếm lĩnh niềm tin của người sử dụng, dần trở thành một phần không thể thiếu trong cuộc sống thường nhật hiện đại.

Tuy nhiên, việc cho phép máy móc có thể tự tư duy và tự phát triển tư duy thông qua học hỏi hệ thống dữ liệu lớn, và việc con người quá phụ thuộc vào nó sẽ có thể làm phát sinh những hậu quả nghiêm trọng ngoài tầm kiểm soát. Ví dụ điển hình là một số người đã bị thiệt mạng trong những vụ tai nạn ô tô tự hành, khi ô tự lái gặp phải tình huống không thể đưa ra quyết định an toàn.[5] Tương tự, trường hợp hệ thống AI được công ty môi giới chứng khoán sử dụng để phân tích và thực hiện giao dịch trên sàn tiền mã hóa (*crypto*) trực tuyến đã không biết ngắt giao dịch khi phát sinh lỗi kỹ thuật làm thị trường bị gián đoạn dẫn tới những thiệt hại to lớn cho khách hàng.[6]

Trong bối cảnh hệ thống pháp luật về AI chưa được ban hành hoặc chưa đầy đủ, cơ sở pháp lý cơ bản để xác định trách nhiệm pháp lý do các tai nạn

3 "Trí tuệ nhân tạo" một nhánh của khoa học máy tính tập trung vào trí tuệ do máy móc điều khiển (tức là trí thông minh không phải của con người). Trong lịch sử hiện đại, thuật ngữ "Trí tuệ nhân tạo" được lần đầu tiên sử dụng vào năm 1955 tại Hội nghị Dartmouth bởi John McCarthy – nhà khoa học máy tính và khoa học nhận thức của Hoa Kỳ. Ông và các đồng nghiệp đã phát triển Lisp, ngôn ngữ lập trình Lisp phổ biến nhất và được ưa chuộng để nghiên cứu trí tuệ nhân tạo. Chương trình máy tính AI đầu tiên được giới thiệu bởi nhóm các nhà khoa học Allen Newell, Hebert Simon và Cliff Shaw vào năm 1959, sau đó đã đưa AI gắn với ngành robot và các máy móc tự động hóa. Xem thêm: A Brief History of AI, "AI Topics", 2018, https://aitopics.org/misc/brief-history, truy cập ngày 29/04/2024.

4 Matthias Artzt & Tran Viet Dung, "Artificial Intelligence and Data Protection: How to Reconcile Both Areas from the European Law Perspective", *Vietnames Journal of Legal Sciences*, Vol. 7, No. 2, 2022, tr. 39-58, https://intapi.sciendo.com/pdf/10.2478/vjls-2022-0007, truy cập ngày 29/04/2024.

5 Gregory Yalt, "Inside the final seconds of a deadly Tesla Autopilot crash", *Washington Post*, https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis, truy cập ngày 22/05/2024; Andrew J. Hawkins, "Tesla's Autopilot and Full Self-Driving linked to hundreds of crashes, dozens of deaths", https://www.theverge.com/2024/4/26/24141361/tesla-autopilot-fsd-nhtsa-investigation-report-crash-death, truy cập ngày 22/05/2024.

6 Nguyễn Hoàng Thái Hy & Lê Tấn Phát, "Rủi ro từ giao dịch tự động trên sàn giao dịch tiền mã hóa nằm ngoài ý chí của các bên – "lẽ công bằng" có phải là giải pháp?", *Tạp chí Khoa học pháp lý Việt Nam*, số 10(170), 2023, tr. 27-37.

của AI là hợp đồng giữa nhà sản xuất AI hoặc các nhà vận hành hệ thống AI với người sử dụng, khách hàng. Tuy nhiên, thông thường các hợp đồng đó lại sẽ do nhà sản xuất hoặc bên vận hành hệ thống AI soạn thảo, vì vậy sẽ chứa các điều khoản loại trừ hoặc hạn chế trách nhiệm pháp lý của họ trong trường hợp xảy ra tai nạn… Nhưng từ quan điểm đạo đức, cho dù có điều khoản miễn trách trong hợp đồng thì trách nhiệm sản phẩm rõ ràng cũng có thể phần nào đó vẫn thuộc về bên cung cấp sản phẩm AI.

Mặc dù AI có tốc độ và khả năng xử lý cao hơn nhiều so với con người, nhưng không phải lúc nào công nghệ này cũng được tin cậy là công bằng và trung lập. Hệ thống AI vận hành trên cở sở các dữ liệu đầu vào và các mẫu phân tích do con người cung cấp. Năm 2015, Google Photos đã đối mặt với sự chỉ trích nghiêm trọng do một sự cố liên quan đến phần mềm nhận diện hình ảnh. Hệ thống AI được sử dụng bởi Google Photos đã gắn nhầm ảnh của người Mỹ gốc Phi là "*gorillas*" (khỉ đột).[7] Sai sót này đã làm nổi bật một thiếu sót nghiêm trọng trong việc đào tạo và xử lý dữ liệu của AI. Google nhanh chóng xin lỗi về sai sót này và đã thực hiện các bước ngay lập tức để khắc phục. Sự cố này đã là một lời cảnh tỉnh cho ngành công nghệ về khả năng giải quyết các thiên vị trong hệ thống AI, nhấn mạnh sự cần thiết phải có dữ liệu đào tạo đa dạng và toàn diện hơn để ngăn chặn các lỗi tương tự và làm nổi bật trách nhiệm đạo đức và xã hội của các công ty phát triển công nghệ AI. Cần lưu ý rằng các mẫu dữ liệu của hệ thống AI được thực hiện bởi các cá nhân, và họ có thể là những người có thể thiên vị và hay phán xét. Rõ ràng nếu công nghệ được sử dụng đúng cách bởi những người tìm kiếm sự tiến bộ xã hội, AI có thể đóng vai trò là chất xúc tác cho những thay đổi tích cực, và ngược lại. Do đó, cần phải đảm bảo rằng hệ thống AI được thiết kế và hoạt động như dự định và không thể bị chi phối bởi những người có động cơ ích kỷ.

Với công nghệ học máy và học sâu của AI, có thể thấy máy móc có khả năng tự nhân rộng các suy luận và học hỏi để đưa ra các hành vi, không cần sự can thiệp của con người. Nhiều người hoài nghi liệu con người đã thực sự hiểu cách thức công nghệ hoạt động và thu thập dữ liệu nào đó khi hệ thống đang hoạt động không? Phân tích đạo đức luôn bắt đầu bằng việc thu thập thông tin, thế nhưng khi máy móc tự thực hiện việc lấy thông tin đầu vào và tự đánh giá thông tin đó theo những logic mà nó tích lũy qua việc tự học thì những giá trị đạo đức của con người có thể bị máy bỏ qua. Kỹ thuật học máy và học sâu trong mạng lưới "thần kinh" của hệ thống AI phức tạp và to lớn sẽ gây khó khăn cho con người trong việc tìm hiểu lý do tại sao hệ thống lại đưa ra lựa chọn như vậy ở một số tình huống. Những trường hợp khác, máy móc có thể đưa ra một số lời giải thích nhưng lại quá phức tạp để con người

7 Daniel Howley, "Google Photos Mislabels 2 Black Americans as Gorillas", *Yahoo Tech*, https://www.yahoo.com/tech/google-photos-mislabels-two-black-americans-as-122793782784.html, truy cập ngày 22/05/2024.

có thể hiểu được. Điều này khiến cho tính minh bạch của những quyết định do máy móc tạo ra bị giảm sút.[8]

Trong những năm gần đây, chúng ta chứng kiến các chatbot AI ngày càng trở nên thành thạo trong việc mô phỏng sự tương tác và các mối quan hệ của con người. Hệ thống ChatGPT, Alexa, Siri, Gemini… đã xuất sắc trong việc tương tác với con người một cách hiệu quả, thậm chí không khác gì một người thật, mang lại các phân tích nhanh chóng và sự hài lòng cho người dùng. Nhiều chuyên gia tin rằng đây là bước khởi đầu của một kỷ nguyên mới, khi con người sẽ sử dụng máy móc như người, đặc biệt là trong các lĩnh vực như dịch vụ chăm sóc khách hàng và bán hàng. Trong khi con người có hạn chế về thời gian và sự chú ý dành cho người khác, *robot* AI có thể dùng nguồn lực gần như không giới hạn để xây dựng và duy trì các mối quan hệ. Thế nhưng, cũng có nguy cơ rằng các *chatbot* AI có thể không thực sự hiểu vấn đề (hoặc thậm chí là tạo ra các thông tin không có thật để có lợi cho ai đó), không đáp ứng được mọi khía cạnh của tâm trạng và nhu cầu của người dùng, điều có thể dẫn đến sự thất vọng hoặc nhận thức sai lầm về vấn đề của người dùng trong quá trình tương tác. Cũng có sẽ xảy ra việc sử dụng AI để thay thế hoặc giả mạo sự tương tác giữa con người nhằm mục đích bất chính (ví dụ như sử dụng công nghệ *deep fake* để lừa đảo). Từ góc độ đạo đức, vấn đề được đặt ra là liệu việc thay thế giao tiếp giữa con người bằng AI có góp phần vào việc tạo ra một môi trường giao tiếp và tương tác xã hội lành mạnh và tích cực hay không.

Cuối cùng, một vấn đề mang tính đạo đức nổi bật liên quan tới sự tiến hóa của các máy móc – vốn đã luôn xảy ra trong suốt chiều dài lịch sử văn minh nhân loại, đó là việc sử dụng công nghệ AI sẽ loại trừ hoặc giảm bớt công ăn việc làm cho lực lượng lao động cao người. Ứng dụng công nghệ AI trong sản xuất kinh doanh một cách rộng khắp sẽ có thể dẫn tới tình trạng những người có quyền sở hữu trong các doanh nghiệp ứng dụng AI sẽ thu được nguồn lợi nhuận lớn, trong người lao động thất nghiệp do không thể cạnh tranh với máy móc; các doanh nghiệp không có quyền sở hữu công nghệ AI sẽ lụi tàn và phá sản… khi đó sẽ phát sinh nhiều vấn đề xã hội phức tạp. Xã hội văn minh đề cao nguyên tắc công bằng và bình đẳng, như một giá trị đạo đức quan trọng. Vì vậy, việc sử dụng công nghệ AI có thể khiến cho lực lượng lao động truyền thống bị loại trừ và mất cơ hội việc làm, làm tăng sự bất bình đẳng trong xã hội cần phải được xem xét. Công bằng trong cơ hội tiếp cận công nghệ và sự minh bạch trong việc chia sẻ lợi ích từ sự phát triển công nghệ AI cần được đảm bảo. Đạo đức yêu cầu doanh nghiệp phải chịu trách nhiệm với tác động của công nghệ AI đến xã hội.

Với hàng loạt những vấn đề nêu trên, vấn đề đạo đức AI thực sự đóng vai trò quan trọng đối với sự phát triển của xã hội trong tương lai. Ở đây, bên

8 Matthias Artzt & Tran Viet Dung, *tldd*.

cạnh việc đảm bảo rằng việc phát triển và triển khai công nghệ AI, các bên liên quan phải được thực hiện hoạt động này với sự cân nhắc kỹ lưỡng và tuân thủ các nguyên tắc đạo đức, đảm bảo rằng nó mang lại giá trị tích cực và không gây hại cho người dùng và xã hội.

2. Sự phát triển của các tiêu chuẩn đạo đức trí tuệ nhân tạo trên thế giới

Cách đây không lâu, các công nghệ AI đã phát triển trong điều kiện gần như hoàn toàn tự do, không được điều chỉnh hoặc chi phối bởi bất cứ quy phạm đạo đức, pháp lý hay tiêu chuẩn kỹ thuật nào. Phải tới năm 2017, vấn đề thiết lập các tiêu chuẩn đạo đức đối với hoạt động phát triển và sử dụng AI mới lần đầu tiên được giới thiệu thông qua các Nguyên tắc Asilomar về AI,[9] đánh dấu sự khởi đầu của việc tiếp cận có trách nhiệm của nhân loại đối với sự phát triển của AI và *robot*. Ngay sau đó, hầu như tất cả các quốc gia phát triển đã xây dựng các nguyên tắc đạo đức cho AI: Khuyến nghị về xe tự hành ở Đức, Tuyên bố về Phát triển AI có Trách nhiệm ở Canada, Nguyên tắc Hướng dẫn Đạo đức cho Hiệp hội AI Nhật Bản, Hướng dẫn Đạo đức cho AI Đáng tin cậy của Châu Âu, Mô hình Công ước về Robot và AI ở Nga; Nguyên tắc Đạo đức cho AI tại Hoa Kỳ… cũng như các chiến lược quốc gia để phát triển trí tuệ nhân tạo.

Một phân tích về "cảnh quan toàn cầu" về các nguyên tắc đạo đức AI, được thực hiện bởi một nhóm các nhà nghiên cứu quốc tế do Anna Jobin đứng đầu,[10] cho thấy sự hội tụ của các bên liên quan trong việc thúc đẩy các nguyên tắc riêng lẻ. Cụ thể, hầu hết các chuyên gia và nhà nghiên cứu AI đều thống nhất về một số nguyên tắc liên quan tới sự minh bạch, công bằng, không gây hại, trách nhiệm giải trình và bảo mật.[11] Các học giả K. Burle và D. Cortiz, sau khi phân tích các sáng kiến của Ủy ban châu Âu, Bộ Quốc phòng Hoa Kỳ, Tổ chức Hợp tác và Phát triển Kinh tế, Học viện AI Bắc Kinh, Google và Microsoft liên quan tới AI, đã xác định một số nguyên tắc chung về công bằng, chuẩn mực về độ tin cậy và an toàn, trách nhiệm giải trình, đồng thời cũng đưa ra các nguyên tắc tác động xã hội và minh bạch.[12] Quan điểm này được hoan nghênh bởi nhiều chuyên gia khác trên trường quốc tế.[13] A.V. Popova trong nghiên cứu của mình cũng đã xác định các nguyên tắc đạo đức AI cơ bản, bao gồm: an toàn, độ tin cậy, minh bạch, bảo vệ nhân quyền, khách quan, công

9 Bộ nguyên tắc Asilomar (Asilomar Principles) gồm 23 nguyên tắc trong nghiên cứu và phát triển AI, được công bố tại Asilomar Conference on Beneficial AI do Future of Life Institute tổ chức tại California vào năm 2017, với sự tham gia của hơn 100 chuyên gia chính sách, nhà khoa học kinh tế, luật, đạo đức, robot học và triết học có uy tín quốc tế.

10 Jobin Anna, Ienca Marcello, Vayena Effy, "The global landscape of AI ethics guidelines", *Natural Machine Intelligence*, Vol. 01, 2019, tr. 389–399.

11 *Như trên*.

12 Burle Caroline & Cortiz Diogo, "Mapping principles of Artificial Intelligence", *Sao Paulo: Nucleo de Informacao eCoordenacao do Ponto BR*, 2019, https://www.researchgate.net/publication/344357163_Mapping_principles_of_Artificial_Intelligence, truy cập ngày 12/05/2024.

13 Hagendorff Thilo, "The Ethics of AI Ethics: An Evaluation of Guidelines", Minds & Machines, No. 30, 2020, tr. 99–120; Boddington Paula, *Towards a Code of Ethics for Artificial Intelligence*, Springer International Publishing, 2017; Luciano Floridi & Josh Cowls, "A unified framework of five principles for AI in society", *Harvard Data Science Review,* Vol. 01(01), 2019, tr. 2–17.

bằng, kiểm soát, bảo mật, trách nhiệm...[14] Có thể nói tại thời điểm hiện nay một hệ thống các nguyên tắc điều chỉnh đạo đức của AI đang được hình thành trên phạm vi toàn cầu. Chúng tạo ra một nền tảng lý luận quan trọng cho các học thuyết pháp lý cũng như việc xây dựng các quy phạm pháp luật về AI.

Liên minh châu Âu (*European Union*, EU) đã thông qua Đạo luật AI vào tháng 5 năm 2024,[15] đánh dấu sự xuất hiện của một đạo luật lớn đầu tiên trên thế giới để điều chỉnh AI. Đây là một nỗ lực của EU trong việc định hình và điều chỉnh sự phát triển của AI nhằm đảm bảo rằng công nghệ AI được sử dụng một cách đúng đắn và có ích cho xã hội. Đạo luật AI được thiết kế theo phương pháp đánh giá rủi ro đối với AI, có nghĩa là các ứng dụng khác nhau của công nghệ này sẽ được xử lý theo cách khác nhau, tùy thuộc vào các mối đe dọa được nhận thức mà chúng gây ra đối với xã hội.[16] Đạo luật AI đặt ra các tiêu chí đánh giá rủi ro căn cứ theo các tiêu chuẩn đạo đức AI đã được công nhận tại EU, như an toàn và bảo mật thông tin, minh bạch, tôn trọng quyền cá nhân và đảm bảo tính công bằng và trách nhiệm xã hội…[17] Dù tầm ảnh hưởng của Đạo luật AI của EU cũng có thể sẽ mang lại hiệu ứng Brussel như Bộ Quy định chung về Bảo vệ dữ liệu cá nhân (*General Data Protection Regulation*), nghiên cứu của Joblin cũng chỉ ra rằng các nguyên tắc đạo đức AI đã và đang được phát triển riêng lẻ sẽ không thể cung cấp một cơ sở tham chiếu chung cho mọi quốc gia với từng bối cảnh xã hội và thể chế khác biệt.[18]

Tại Việt Nam, mọi thứ mới chỉ ở những bước đầu tiên trong tiếp cận công nghệ AI và đến nay hầu như không có những tài liệu hay hướng dẫn về đạo đức AI. Tuy nhiên, vấn đề quy định đạo đức AI rõ ràng liên quan đến Việt Nam, vì Chính phủ đã đặt mục tiêu đưa AI trở thành lĩnh vực công nghệ quan trọng của Việt Nam trong cuộc Cách mạng công nghiệp lần thứ tư. Đến năm 2030, Việt Nam mong muốn trở thành trung tâm đổi mới sáng tạo, phát triển các giải pháp và ứng dụng AI trong khu vực ASEAN và trên thế giới.[19] Chiến lược Quốc gia về Phát triển Trí tuệ Nhân tạo cho đến năm 2030 nhấn mạnh việc điều chỉnh các quy định liên quan đến sự tương tác của con người với AI, sự cần thiết phải phát triển các tiêu chuẩn đạo đức phù hợp cũng được ghi nhận. Bên cạnh việc nghiên cứu các quy tắc pháp luật về AI, Việt Nam cũng cần quan tâm phát triển các luật mềm dưới hình thức các hướng dẫn hoặc bộ quy tắc đạo đức về AI.

14 Попова Анна [Popova Anna], *tlđd*.

15 Ryan Browne, "World's first major law for artificial intelligence gets final EU green light", *CNBC News*, 21/05/2024, https://www.cnbc.com/2024/05/21/worlds-first-major-law-for-artificial-intelligence-gets-final-eu-green-light.html, truy cập ngày 26/05/2024.

16 European Parliament, "Artificial Interligence Act (Briefing)", https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf, truy cập ngày 26/05/2024.

17 Như trên; Xem Bản thảo nội dung của Đạo luật AI, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf, truy cập ngày 26/05/2024.

18 Jobin Anna, Ienca Marcello, Vayena Effy, *tlđd*.

19 Nghị quyết số 127/QĐ-TTg của Thủ tướng ngày 26/01/2021, Ban hành "Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo đến năm 2030".

Các doanh nghiệp, hiệp hội doanh nghiệp, hiệp hội ngành nghề, các tổ chức nghiên cứu và phát triển AI tại Việt Nam cần cùng bắt tay xây dựng bộ quy tắc đạo đức về AI vì lợi chung của sự phát triển. Cần lưu ý rằng mục tiêu ở đây không chỉ là xác định danh sách các nguyên tắc chung và phát triển một cách tiếp cận thống nhất để hiểu chúng, mà còn là tạo ra một hệ thống toàn diện để điều chỉnh đạo đức dựa trên chúng, bao gồm cả việc phát triển và triển khai các cơ chế thực hiện và kiểm soát AI. Các bên liên quan có thể tham khảo Khuyến nghị về Đạo đức của Trí tuệ Nhân tạo của Tổ chức Giáo dục, Khoa học và Văn hóa của Liên hợp quốc (*United Nations Educational, Scientific and Cultural Organization*, UNESCO) ban hành ngày 25/6/2021 – một thỏa thuận xác định các giá trị và nguyên tắc chung cần thiết để đảm bảo sự phát triển của AI.[20]

3. Khuyến nghị về đạo đức của trí tuệ nhân tạo được ban hành bởi Tổ chức Giáo dục, Khoa học và Văn hóa của Liên hợp quốc

Khuyến nghị về Đạo đức của Trí tuệ Nhân tạo của UNESCO (Bản khuyến nghị UNESCO) là một thỏa thuận quốc tế nhằm mục đích tạo ra một nền tảng cho hoạt động của hệ thống AI vì lợi ích của nhân loại, các cá nhân, xã hội, môi trường và hệ sinh thái, ngăn ngừa tác hại, thúc đẩy việc sử dụng AI một cách hòa bình, cũng như tạo ra một công cụ quy định được quốc tế công nhận, tập trung vào việc đưa các vấn đề bình đẳng giới, bảo vệ môi trường và hệ sinh thái vào, và như được nêu trong văn bản, là sự bổ sung cho các nguyên tắc đạo đức hiện có liên quan đến AI trên toàn thế giới.

Bản khuyến nghị UNESCO cung cấp một khuôn khổ phổ quát về các giá trị, nguyên tắc và hành động mà các quốc gia có thể được hướng dẫn khi xây dựng luật pháp, chính sách hoặc các công cụ khác liên quan đến AI. Đồng thời, do tính chất phức tạp của các vấn đề đạo đức liên quan đến AI, văn bản nhấn mạnh sự cần thiết của đối thoại toàn cầu và liên văn hóa, cũng như trách nhiệm chung của các bên liên quan.

Trong số các giá trị cần thiết để đảm bảo sự phát triển lành mạnh của AI, tài liệu này nêu rõ các giá trị sau: tôn trọng, bảo vệ và thúc đẩy quyền con người và các quyền tự do cơ bản cũng như phẩm giá con người; sự phát triển thịnh vượng của môi trường và các hệ sinh thái; đảm bảo tính đa dạng và hòa nhập; sống trong các xã hội hòa bình, công bằng và kết nối. Các nguyên tắc được xác định bao gồm: tính cân xứng và không gây hại; an toàn và bảo mật; công bằng và không phân biệt đối xử; tính bền vững; quyền riêng tư và bảo vệ dữ liệu; kiểm soát và quyết định của con người; minh bạch và có thể giải thích; trách nhiệm và trách nhiệm giải trình; nhận thức và hiểu biết; quản trị và hợp tác đa phương và thích ứng.

Thỏa thuận không chỉ tập trung vào việc xây dựng các giá trị và nguyên tắc mà còn đặt trọng tâm vào việc thực hiện chúng trong thực tế và đưa ra các

20 UNESCO, Recommendation on the Ethics of Artificial Intelligence, SHS/BIO/PI/2021/1, https://unesdoc.unesco.org/ark:/48223/pf0000381137, truy cập ngày 10/04/2024.

khuyến nghị cụ thể cho các quốc gia trong các lĩnh vực như đánh giá tác động đạo đức, quản trị đạo đức, chính sách dữ liệu, phát triển và hợp tác quốc tế, môi trường và hệ sinh thái, giới tính, văn hóa, giáo dục và nghiên cứu, truyền thông và thông tin, kinh tế và lao động, y tế và phúc lợi xã hội.

Bản khuyến nghị UNESCO, nêu ra một số điểm quan trọng. Đầu tiên, UNESCO đề xuất áp dụng các biện pháp hiệu quả nhằm phát triển các khuôn khổ chính sách và cơ chế, đảm bảo sự tuân thủ của các bên liên quan trong lĩnh vực AI. Các khuôn khổ để đánh giá tác động đạo đức cũng được triển khai nhằm xác định, đánh giá lợi ích, thách thức và rủi ro của các hệ thống AI, cũng như các biện pháp ngăn ngừa và giảm thiểu rủi ro. Đồng thời, Bản khuyến nghị này cũng nhấn mạnh tầm quan trọng của việc phát triển cơ chế thẩm định và giám sát để bảo đảm việc tuân thủ quyền con người, pháp quyền và sự hòa nhập xã hội. Việc thực thi các cơ chế nghiêm ngặt nhằm điều tra và khắc phục các thiệt hại do AI gây ra cũng được đề cao. Ngoài ra, UNESCO cũng quan tâm đến việcxây dựng chiến lược AI quốc gia và khu vực, cùng với các cơ chế chứng nhận và đánh giá các hệ thống AI hiện có, nhằm đảm bảo tính khả thi và phù hợp của việc triển khai AI. Bản khuyến nghị về đạo đức của AI cũng khuyến khích tạo ra cơ chế để đảm bảo sự tham gia của tất cả các quốc gia thành viên trong các cuộc thảo luận quốc tế về AI, đồng thời thông qua các khung pháp lý để bảo vệ quyền riêng tư và an toàn dữ liệu cá nhân. Việc xem xét lại chính sách và khung pháp lý để phản ánh các yêu cầu cụ thể của AI, thúc đẩy hợp tác quốc tế trong lĩnh vực này cũng là một điểm quan trọng. Cuối cùng, cơ chế để ngăn chặn lạm dụng vị thế thống trị trên thị trường và tuân thủ các tiêu chuẩn đạo đức khi xuất khẩu và phát triển hệ thống AI tại các quốc gia khác cũng được xem trọng.

Khuyến nghị cuối cùng đáng chú ý ở chỗ hướng tới bảo vệ các nước đang phát triển khỏi việc trở thành sân thử nghiệm cho các công ty công nghệ. Điều này cũng giúp thúc đẩy tăng cường hỗ trợ trao đổi kiến thức, thiết lập tiêu chuẩn, đối thoại chính sách, nâng cao năng lực và phát triển mạng lưới liên quan đến AI tại các nước đang phát triển để vượt qua sự phân bổ không đồng đều về lợi ích và rủi ro tiềm tàng của công nghệ AI trên toàn cầu.

Để thực hiện đánh giá tác động của các hệ thống AI, cũng như đánh giá các chính sách, chương trình và cơ chế liên quan đến đạo đức AI, Bản khuyến nghị UNESCO đề xuất triển khai hệ thống các cơ chế giám sát. Các cơ chế có thể được đề xuất bao gồm ủy ban đạo đức, trung tâm quan sát đạo đức AI, kho lưu trữ phát triển các hệ thống AI phù hợp với nhân quyền và đạo đức, cơ chế trao đổi kinh nghiệm, cơ chế thử nghiệm (*sandbox*) quy định trong lĩnh vực AI và hướng dẫn đánh giá cho tất cả các đối tượng AI để đánh giá sự tuân thủ các khuyến nghị chính sách.

Nhìn chung các khuyến nghị nêu trên của UNESCO sẽ giúp thiết lập bộ khung cho hệ thống quản lý đạo đức AI, và nếu chúng được thực hiện

đúng cách sẽ có thể trở thành nền tảng đạo đức phổ quát cho AI tại quốc gia. Tuy nhiên, cũng cần lưu ý rằng các quy phạm đạo đức về AI vẫn là những quy tắc không mang tính ràng buộc, và sự cam kết với các nguyên tắc và giá trị đạo đức dễ dàng bị lấn át bởi các động cơ kinh tế. Vì vậy, việc điều chỉnh đạo đức của AI vẫn cần phải được bổ sung bởi các quy định pháp luật, nhằm tạo ra một môi trường an toàn cho sự phát triển của AI và giảm thiểu những rủi ro có thể xảy ra.[21] Nói cách khác, các nguyên tắc đạo đức sẽ đóng vai trò là "hướng dẫn cho việc sử dụng tích cực AI",[22] trong khi pháp luật cần ngăn chặn, chấm dứt và trừng phạt việc sử dụng công nghệ AI tiêu cực.[23]

Bản Khuyến nghị của UNESCO là một luật mềm về AI rất hữu ích cho các quốc gia đang tìm hiểu xây dựng khung chính sách pháp luật về AI. Chúng không chỉ đề xuất các quy tắc về mà còn cung cấp các công cụ cho việc giám sát điều tiết và dân sự, cho phép đưa các nguyên tắc đạo đức về AI vào thực tiễn và làm cho chúng "có thể đo lường được". Trong hầu hết các tài liệu khoa học, tiêu chuẩn hóa sẽ đóng vai trò quyết định trong việc điều chỉnh các hệ thống AI trong tương lai.[24] Ví dụ, có thể kể đến gói IEEE-P7000 và các chương trình chứng nhận, cũng như các tiêu chuẩn ISO/IEC 20546:2019 về Công nghệ thông tin. Dữ liệu lớn, ISO/IEC 20547-3:2020 về Công nghệ thông tin. Việc áp dụng luật mềm cho phép tạo ra một nền tảng pháp lý đơn giản, vẫn đáng tin cậy và linh hoạt, có thể được thay đổi kịp thời theo sự phát triển của công nghệ.

Kết luận

AI không phải là một kỹ thuật thuần túy mà là một công nghệ mang tính cách mạng sẽ biến đổi đáng kể xã hội loài người và thế giới trong tương lai, có nhiều chỗ để phát triển và triển vọng ứng dụng rộng rãi. Tuy nhiên, sự phức tạp và mức độ giải thích không rõ ràng, sai lệch dữ liệu, bảo mật dữ liệu, bảo vệ quyền riêng tư dữ liệu, trách nhiệm pháp lý khi có sự cố… hiện đang là những vấn đề của AI chưa được điều chỉnh đầy đủ bởi pháp luật. Điều này tiềm ẩn nhiều rủi ro cho người dùng, nhà phát triển, nhân loại và xã hội.

21 Hoffmann-Riem Wolfgang, "Artificial Intelligence as a Challenge for Law and Regulation", in Wischmeyer, T., Rademacher, T. (eds), *Regulating Artificial Intelligence*, Springer; Mihaela Constantinescu, "AI, moral externalities, and soft regulation", 2021, https://www.researchgate.net/publication/356612427_AI_moral_externalities_and_soft_regulation, truy cập ngày 10/01/2024; Jiabao Wang, Xiaoyu Yu, Jie Li, Xiaoling Jin, "Artificial Intelligence and International Norms", in Donghan Jin (ed.), *Reconstructing Our Orders*, Springer, 2018, tr. 195-229*; Камалова Гульфия Гафиятовна,* "Правовые и этические принципы регулирования искусственного интеллекта и робототехники" [trans: Kamalova Gulfia Gafiatova, "Những nguyên tắc pháp luật và đạo đức điều chỉnh AI và kỹ thuật robot", Право и государство: теория и практика, Vol. 202(10), 2021, tr. 181–184.

22 Margarita Robles Carrillo, "Artificial intelligence: From ethics to law", *Telecommunications Policy*, Vol. 44, No. 6, 2020, https://www.sciencedirect.com/science/article/abs/pii/S030859612030029X, truy cập ngày 20/5/2024.

23 Попова Анна [Popova Anna], *tlđd*.

24 Stix Charlotte, "Foundations for the future: institution building for the purpose of artificial intelligence governance", *AI Ethics*, Vol. 2, 2022, tr. 463–476; Kamalova Gulfia Gafiatova [Камалова Гульфия Гафиятовна], *tlđd*; Hoffmann-Riem Wolfgang, *tlđd*.

Chính vì vậy, các vấn đề đạo đức và pháp lý trong phát triển AI mang tính cấp bách, và trong bối cảnh hiện nay, việc hình thành nền tảng đạo đức chung cho AI cùng với quá trình xây dựng quy định pháp lý quốc tế, khu vực và quốc gia cần phải được phát triển song song. Việc xây dựng và áp dụng các nguyên tắc và tiêu chuẩn đạo đức AI sẽ giúp tạp nên một hệ thống luật mềm để giải quyết tức thời những vấn đề mới và liên tục thay đổi do tác động của sự tiến hóa của công nghệ. Đây sẽ là nền tảng để các nhà làm luật thiết lập các quy phạm pháp luật khái quát và các chế tài phù hợp đối với các vấn đề pháp lý phát sinh từ ứng dụng công nghệ AI. ●

Tài liệu tham khảo

[1] Jobin Anna, Ienca Marcello, Vayena Effy, "The global landscape of AI ethics guidelines", *Natural Machine Intelligence*, Vol. 1, 2019

[2] Попова Анна, "Этические принципы взаимодействия с искусственным интеллектом как основа правового регулирования" [trans: Popova Anna, "The Interplay of Ethics and Artificial Intelligence as a Foundation for Legal Regulation"], Правовое государство: теория и практика, Vol. 61(3), 2020

[3] Matthias Artzt & Tran Viet Dung, "Artificial Intelligence and Data Protection: How to Reconcile Both Areas from the European Law Perspective", *Vietnames Journal of Legal Sciences*, Vol. 7(2), 2022

[4] Ryan Browne, "World's first major law for artificial intelligence gets final EU green light", *CNBC News*, 21/05/2024

[5] Burle Caroline & Cortiz Diogo, "Mapping principles of Artificial Intelligence", *Sao Paulo: Nucleo de Informacao eCoordenacao do Ponto BR*, 2019

[6] Margarita Robles Carrillo, "Artificial intelligence: From ethics to law", *Telecommunications Policy*, Vol. 44(6), 2020

[7] Stix Charlotte, "Foundations for the future: institution building for the purpose of artificial intelligence governance", *AI Ethics*, Vol. 2, 2022

[8] Luciano Floridi & Josh Cowls, "A unified framework of five principles for AI in society", *Harvard Data Science Review,* Vol. 01(01), 2019

[9] Камалова Гульфия Гафиятовна, "Правовые и этические принципы регулирования искусственного интеллекта и робототехники" [trans: Kamalova Gulfia Gafiatova, "Legal and ethical principles governing AI and robotics", Право и государство: теория и практика, Vol. 202(10), 2021

[10] Andrew J. Hawkins, "Tesla's Autopilot and Full Self-Driving linked to hundreds of crashes, dozens of deaths"

[11] Daniel Howley, "Google Photos Mislabels 2 Black Americans as Gorillas", *Yahoo Tech*

[12] Nguyễn Hoàng Thái Hy & Lê Tấn Phát, "Rủi ro từ giao dịch tự động trên sàn giao dịch tiền mã hóa nằm ngoài ý chí của các bên – "lẽ công bằng" có phải là giải pháp?", *Tạp chí Khoa học pháp lý Việt Nam*, số 10(170), 2023 [trans: Nguyen Hoang Thai Hy & Le Tan Phat, "Risks from Automatic Transactions on Cryptocurrency Exchanges Beyond the Will of the Parties – Is 'Equity' the Solution?", *Vietnamese Journal of Legal Sciences*, Vol. 170(10), 2023]

[13] D. Jin (Ed.), *Reconstructing Our Orders: Artificial Intelligence and Human Society*, Springer, 2018

[14] European Parliament, "Artificial Interligence Act (Briefing)"

[15] Boddington Paula, *Towards a Code of Ethics for Artificial Intelligence*, Springer International Publishing, 2017

[16] Hagendorff Thilo, "The Ethics of AI Ethics: An Evaluation of Guidelines", *Minds & Machines*, No. 30, 2020

[17] Jiabao Wang, Xiaoyu Yu, Jie Li, Xiaoling Jin, "Artificial Intelligence and International Norms", in Donghan Jin (ed.), *Reconstructing Our Orders*, Springer, 2018

[18] Wischmeyer, T., Rademacher, T. (eds), *Regulating Artificial Intelligence*, Springer; Mihaela Constantinescu, "AI, moral externalities, and soft regulation", 2021

[19] Gregory Yalt, "Inside the final seconds of a deadly Tesla Autopilot crash", *Washington Post*