

ĐẠO ĐỨC AI - MÔ HÌNH CỦA UNESCO VÀ GIÁ TRỊ THAM KHẢO TRONG XÂY DỰNG CHÍNH SÁCH TẠI VIỆT NAM

TS LÊ THANH DŨNG

Khoa Khoa học cơ bản, Trường Đại học Luật TP. Hồ Chí Minh
Faculty of Fundamental Sciences, Ho Chi Minh City University of Law

Email: ltdung@hcmulaw.edu.vn

PHAN QUÝ KHÁNH NGÂN;⁽¹⁾ NGUYỄN KHÁNH BẢO TRÂN⁽²⁾

Khoa Luật Quốc tế, Trường Đại học Luật TP. Hồ Chí Minh
Faculty of International Law, Ho Chi Minh City University of Law

Email: (1) 2253801015196@email.hcmulaw.edu.vn; (2) 2253801090092@email.hcmulaw.edu.vn

Tóm tắt

Trí tuệ nhân tạo (artificial intelligence, AI) đang là một cơn sốt toàn cầu, tạo nên sự thay đổi từ bên trong mọi hoạt động của đời sống, làm chuyển dịch cấu trúc của toàn xã hội. Trong đó, các quy tắc đạo đức đang bị thách thức để đuổi kịp tốc độ của AI. Chính sách về đạo đức AI đóng vai trò quan trọng nhằm ngăn chặn AI xâm phạm quyền con người, thiên lệch và phân biệt đối xử, thiếu minh bạch, tội phạm... Bài viết này sẽ tập trung nghiên cứu cơ sở lý luận hình thành nên đạo đức AI để đưa đến kết luận về sự cần thiết của nó đối với quá trình sáng tạo và ứng dụng AI bền vững. Đồng thời, bài viết đem lại một số tìm hiểu chung cùng việc đánh giá tính phù hợp trong mô hình đạo đức AI do Tổ chức Giáo dục, Khoa học và Văn hóa Liên Hợp Quốc (United Nations Educational, Scientific and Cultural Organization, UNESCO) thiết lập, từ đó định hướng tham khảo trong xây dựng chính sách về kiểm soát và phát triển AI tại Việt Nam.

Từ khóa: trí tuệ nhân tạo, mô hình đạo đức, UNESCO, xây dựng chính sách, Việt Nam

Abstract

Artificial intelligence (AI) is a global phenomenon, transforming every aspect of life and shifting the structure of society. Ethical norms are being challenged to keep pace with AI. AI ethics policies play a crucial role in preventing AI from infringing on human rights, promoting bias and discrimination, lacking transparency, and contributing to crime. This article will focus on the theoretical foundations of AI ethics to conclude on its necessity for sustainable AI creation and application. Simultaneously, the article will offer some general insights and assess the suitability of the UNESCO-established AI ethics model, providing guidance for policy development regarding the control and development of AI in Vietnam.

Keywords: artificial intelligence, ethical model, UNESCO, policy making, Vietnam

Ngày nhận bài: 01/04/2026

Ngày duyệt đăng: 09/07/2026

Tại Việt Nam, sự phát triển vượt bậc của công nghệ trí tuệ nhân tạo (artificial intelligence, AI) được thúc đẩy bởi động lực to lớn trong các chính sách của Chính phủ, điển hình như Quyết định số 127/QĐ – TTg ngày 26/01/2021 về “Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng AI đến năm 2030”. Chiến lược này được hiện thực hóa trên thực tế với sự nỗ lực của toàn xã hội, đặc biệt đến từ các doanh nghiệp, tổ chức và các nhóm nghiên cứu học thuật, đưa AI vào đời sống nhằm cải thiện mọi mặt trong chất lượng sống của con người. Tuy nhiên, bên cạnh những lợi ích trên, AI cũng đem lại nhiều rủi ro về vấn đề đạo đức khi chúng đạt đến trình độ phát triển tư duy và hành động vượt quá kiểm soát của con người. Với mục đích bảo vệ con người, hỗ trợ con người trong lao động và sáng tạo, AI càng trở nên “giống người” hơn. Mặt khác, khi tham gia sâu rộng trong nhiều lĩnh vực như y tế, giáo dục, kinh doanh, hành chính, giao thông, tín dụng..., xuất phát từ những lập trình mặc định mang tính thiên kiến, AI có thể tạo nên sự bất bình đẳng, đưa ra quyết định sai lầm, gây

ra thiệt hại, đe dọa quyền và tự do cơ bản của con người.¹ Thực tiễn đã chứng minh, tốc độ đột phá của AI đã bỏ xa các quy định và hướng dẫn AI trên toàn thế giới và khung pháp lý về đạo đức AI ở Việt Nam hiện chỉ mới là những bước sơ khai. Trước thời cơ lớn khi các cấp lãnh đạo cao nhất đang đặt sự quan tâm sâu sắc đến tiềm năng của AI, cần nhấn mạnh nghiên cứu và phát triển đạo đức AI. Đây là nhiệm vụ chiến lược cho việc cân bằng giữa phát triển kinh tế thông minh và quản lý rủi ro khi bước vào kỷ nguyên công nghệ số.

1. Cơ sở lý luận về đạo đức AI

Trong bối cảnh cuộc Cách mạng công nghiệp lần thứ tư tiếp tục phát triển trên quy mô toàn cầu và quá trình chuyển đổi số quốc gia diễn ra mạnh mẽ, trí tuệ nhân tạo (AI) đã trở thành lĩnh vực được Đảng và Nhà nước Việt Nam đặc biệt quan tâm, xác định là một trong những động lực chiến lược thúc đẩy tăng trưởng, đổi mới sáng tạo và nâng cao năng lực cạnh tranh quốc gia. Nghị quyết số 57-NQ/TW ngày 22/12/2024 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia đã xác lập các định hướng chiến lược về phát triển, làm chủ và ứng dụng AI như những nhiệm vụ trọng tâm, xuyên suốt trong tiến trình phát triển đất nước. Trên cơ sở đó, Nghị quyết số 66-NQ/TW ngày 30/4/2025 của Bộ Chính trị về đổi mới công tác xây dựng và thi hành pháp luật đáp ứng yêu cầu phát triển đất nước trong kỷ nguyên mới tiếp tục đặt ra yêu cầu hoàn thiện thể chế, chính sách và khuôn khổ pháp lý nhằm bảo đảm phát triển, quản trị và ứng dụng AI theo hướng an toàn, có trách nhiệm và phù hợp với yêu cầu của kỷ nguyên số. Báo cáo chính trị của Ban Chấp hành Trung ương Đảng khóa XIII tại Đại hội Đại biểu toàn quốc lần thứ XIV của Đảng và Dự thảo Chương trình hành động của Ban Chấp hành Trung ương Đảng thực hiện Nghị quyết Đại hội đại biểu toàn quốc lần thứ XIV của Đảng cho thấy, Đảng ta đang định hình chiến lược AI toàn diện, bao trùm các trụ cột về pháp lý, hạ tầng, nhân lực và kinh tế. Chiến lược này thể hiện một bước chuyển tư duy quan trọng, từ tham gia và ứng dụng AI sang chủ trương, chính sách chiến lược để làm chủ và tự quyết về AI, coi đây là yếu tố quyết định, có vai trò quan trọng trong để đất nước ta vững vàng tiến vào kỷ nguyên phát triển mới của dân tộc.

Các nguyên tắc đạo đức AI đầu tiên của Tổ chức Hợp tác và Phát triển Kinh tế (*Organisation de coopération et de développement économiques*, OECD) cho thấy bối cảnh của các nguyên tắc đạo đức AI là rất rộng lớn. Tuy nhiên, các nguyên tắc đạo đức cốt lõi của AI này đều có nguồn gốc từ các quyền cơ bản của con người được công nhận trên toàn cầu, các tuyên bố và công ước hoặc hiệp ước quốc tế.² Sự xuất hiện của AI vào những năm 50 của thế kỷ XX³ đã mở ra một chương mới cho lịch sử nhân loại, thay đổi toàn diện mọi mặt của đời sống xã hội trên cả hai phương diện tích cực và tiêu cực. Sự bất bình đẳng mạnh mẽ và ngày càng trở nên rõ nét khi các nền kinh tế mạnh áp dụng các cơ sở sản xuất tự động hóa, khiến con đường tăng trưởng truyền thống thông qua sản xuất chi phí thấp để xuất khẩu có thể bị cắt đứt, hàng loạt công nhân lâm vào

1 UNESCO, *Recommendation on the ethics of artificial intelligence*, 2021, <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>, truy cập ngày 26/9/2025.

2 “Đạo đức AI - đã đến lúc cần một bộ quy tắc đảm bảo lấy con người làm trung tâm”, <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>, truy cập ngày 07/10/2025.

3 Văn Toàn, “Những cột mốc đánh dấu sự phát triển và hình thành của trí tuệ nhân tạo”, *Báo Nhân dân*, <https://nhandan.vn/nhung-cot-moc-danh-dau-su-phat-trien-va-hinh-thanh-cua-tri-tue-nhan-tao-post742632.html>, truy cập ngày 07/10/2025.

tình cảnh thất nghiệp.⁴ Ngoài ra, một nhóm chủng tộc đã bị phân biệt đối xử bởi các thuật toán của trí tuệ nhân tạo, điển hình là các hệ thống nhận dạng khi quét khuôn mặt người da màu đã gán nhãn họ là khi đột hay hệ thống nhận diện người phía trước của xe tự lái có khả năng nhận biết kém hơn đáng kể nếu phía trước có người da màu, tiềm ẩn rủi ro tai nạn cao.⁵ Điều đầu tiên của Tuyên ngôn quốc tế về quyền con người năm 1948 nêu rõ: “Mọi người sinh ra tự do và bình đẳng về phẩm cách và quyền lợi, có lý trí và lương tri, và phải đối xử với nhau trong tình bác ái”. Chính vì vậy, quyền bình đẳng về phẩm cách và quyền lợi là quyền tự nhiên của con người, bảo đảm con người không bị đối xử khác biệt từ bất kỳ tác động nào, bao gồm cả AI vì “chủng tộc, màu da, giới tính, ngôn ngữ, tôn giáo, chính kiến, nguồn gốc quốc gia hoặc xã hội, tài sản, địa vị khi sinh ra hoặc địa vị khác” (Điều 2 Tuyên ngôn quốc tế về Nhân quyền 1948). Bên cạnh sự phân biệt, AI còn đe dọa tính mạng và sức khỏe của con người khi trong quá trình tự động hóa máy móc, đồng thời là mối nguy cơ đối với danh dự và nhân phẩm, quyền riêng tư trong thu thập, xử lý dữ liệu. AI là một công nghệ mang tính cách mạng làm biến đổi xã hội loài người vô cùng nhanh chóng. Công nghệ dựa trên AI không nên được sử dụng để khai thác con người dưới bất cứ cách thức nào. Thay vào đó, nó phải được sáng tạo, thiết kế và ứng dụng để phục vụ và phát triển con người cùng môi trường sống của họ.

Tiền thân của AI đến từ ngành robot học, lý thuyết kiểm soát và mô phỏng các hệ thống sinh học tự nhiên. Trong suốt tiến trình tiến hóa của AI, rất nhiều hội thảo toàn cầu quy tụ chuyên gia trong giới học thuật được tổ chức, nhưng liên quan đến tiền đề của AI phải nhắc đến hội thảo Dartmouth năm 1956 – hội nghị này đưa ra nhiều giải pháp táo bạo để giải quyết các vấn đề phổ biến trong nhiều ngành là phát triển “cỗ máy tư duy”. Từ năm 1981 đến 2009, lượng nghiên cứu về AI và doanh thu kiếm được cho việc kinh doanh từ nó đạt đến đỉnh điểm. Năm 2010, thế giới đã chứng kiến sự bùng nổ trong các lĩnh vực ứng dụng như nhận diện giọng nói, xử lý ngôn ngữ và thị giác máy tính.⁶ Như đã phân tích ở phần trước, những nguy cơ do việc sử dụng trí tuệ nhân tạo mà không thiết lập rào cản gây ra cho nhân loại rất nhiều tổn thất đáng kể. Mối lo ngại này đã đánh động các nhân vật chủ chốt trong giới triết học, đạo đức và trí tuệ nhân tạo. Đặc biệt, khi người ta nghiên cứu, sáng tạo sản phẩm AI tổng quát hiện đại (*Artificial general intelligence*, AGI) nhằm cố gắng làm cho AI có trí thông minh gần với con người và có khả năng tự học mà không cần thông qua đào tạo đã tác động trực tiếp đến các vấn đề đạo đức. Quan điểm phát triển Đạo đức AI đã xuất hiện từ những năm 1985, nhưng gần đây, quan điểm mới trở thành chủ đề được đưa ra bàn và tranh luận sôi nổi. Ý tưởng áp dụng đạo đức vào trí tuệ máy móc là một trong những mục tiêu nghiên cứu chính trong lĩnh vực đạo đức máy móc. Mối đe dọa của AI không chỉ được các chuyên gia trong lĩnh vực này nhận thấy, nhiều nhân vật từ các lĩnh vực khác cũng đã lên tiếng cảnh báo. Nhà viết kịch nổi tiếng Karl Capek, nhà vật lý thiên văn nổi tiếng Stephen Hawking và triết gia có ảnh hưởng Nick Bostrom đều đã cảnh báo về những nguy cơ tiềm ẩn nếu AI có được suy nghĩ riêng của chính mình. Với sự tiến

4 Stuart J. Russell, Peter Norvig, Ming-Wei Chang, *Artificial intelligence: A modern approach* (4th edition, Global Edition), Pearson Education Limited, 2022, tr. 1037.

5 FAQ, “Discrimination and biases”, *Ethics of AI MOOC*, <https://ethics-of-ai.mooc.fi/chapter-6/3-discrimination-and-biases>, truy cập ngày 07/10/2025.

6 Kakembo Aisha Annet, “The ethics of AI: Philosophical perspectives”, *Research Invention Journal of Research in Education*, Vol. 5(2), 2025, tr. 65. DOI: <https://doi.org/10.59298/RIJRE/2025/526571>

bộ nhanh chóng, nhiều người dự đoán tại một thời điểm trong tương lai, AI có thể sánh ngang hoặc vượt qua khả năng của con người, máy móc thông minh hoàn toàn có thể “nổi dậy” chống lại người tạo ra chúng. Thời điểm ấy được Vernor Vinge gọi là “điểm kỳ dị công nghệ” vào năm 1993.⁷ Do đó, theo Nick Bostrom, việc xây dựng AI thân thiện, bị ràng buộc bởi đạo đức là vô cùng quan trọng. Năm 1997, Rosalind Picard nhấn mạnh nhiều lần rằng “cỗ máy càng tự do thì càng cần các tiêu chuẩn đạo đức”.⁸ Với những bản dạng mới trong tiến trình tiến hóa, đạo đức AI đã phát triển mạnh mẽ vượt xa phạm vi học thuật và trở thành làn sóng quan tâm lan rộng trên nhiều diễn đàn quốc tế đồng thời trở thành một hoạt động thường xuyên của các tổ chức liên quốc gia, ví dụ: Lời kêu gọi Rome về đạo đức AI, được đưa ra vào tháng 02/2020, liên kết Vatican với Tổ chức Lương thực và Nông nghiệp Liên hợp quốc (*Food and Agriculture Organization*, FAO), Microsoft, IBM và Bộ Đối mới của Ý. Một ví dụ khác là UNESCO đã bổ nhiệm 24 chuyên gia từ khắp nơi trên thế giới vào tháng 7/2021 và khởi động một cuộc tham vấn trực tuyến toàn cầu về đạo đức AI, song song với tạo điều kiện thuận lợi cho đối thoại với tất cả các quốc gia thành viên UNESCO. Những tranh luận về đạo đức AI cũng thu hút được sự chú ý đáng kể của giới truyền thông, mặc dù một số học giả cho rằng cách truyền thông đề cập đến đạo đức AI vẫn còn rất “hời hợt” (như ý kiến của Ouchchy và cộng sự vào năm 2020). Một trong những vấn đề lớn mà các nhà phát triển về đạo đức AI cũng như các nhà nghiên cứu đạo đức học có thể gặp phải là sự nhập nhằng, thiếu minh bạch về sự chuyển hóa của AI, bởi vì việc nắm rõ bản thân hoạt động của từng hệ thống AI sẽ rất hữu ích trong việc xác định chính xác các vấn đề đạo đức của nó.⁹

Bộ quy tắc đạo đức đầu tiên cho các hệ thống AI được giới thiệu bởi nhà văn khoa học viễn tưởng nổi tiếng Isaac Asimov, người đã trình bày ba định luật về Robot trong cuốn *Runaround* (năm 1942). Ba định luật này sau đó được bổ sung bởi định luật thứ tư ghi nhận trong cuốn “*Robots and Empire*”¹⁰ của cùng tác giả có nội dung cụ thể như sau:

Định luật 1: Một robot không được phép làm hại con người hoặc để con người bị tổn hại do không hành động;

Định luật 2: Một robot phải tuân theo mệnh lệnh của con người, trừ khi mệnh lệnh đó mâu thuẫn với định luật thứ nhất;

Định luật 3: Một robot phải bảo vệ sự tồn tại của chính nó miễn là sự bảo vệ đó không xung đột với định luật thứ nhất hoặc thứ hai;

Định luật 4: Một robot không được gây hại cho loài người hoặc nếu không hành động, sẽ khiến loài người phải chịu tổn hại.¹¹

Bốn định luật trên của Asimov đã đóng một vai trò quan trọng làm nền tảng cho đạo đức AI trong nhiều thập kỷ. Tuy nhiên, nhìn chung bốn định luật này chỉ mới mang tính khái quát các khía cạnh đạo đức đơn thuần, chưa giải quyết được những vấn

7 Vernor Vinge, “The coming technological singularity: How to survive in the post-human era”, *Whole Earth Review*, Winter Issue, 1993, tr. 88-95.

8 John-Stewart Gordon và Sven Nyholm, “Ethics of artificial intelligence”, *Internet Encyclopedia of Philosophy*, 2021, <https://iep.utm.edu/ethics-of-artificial-intelligence/>, truy cập ngày 08/10/2025..

9 Bernd Carsten Stahl, Doris Schroeder, Rowena Rodrigues, *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges*, Springer Briefs in Research and Innovation Governance, 2023, tr. 2.

10 Isaac Asimov, *Robots and Empire*, Doubleday, New York, 1985, tr. 291.

11 Isaac Asimov, “Runaround”, in The Isaac Asimov Collection (ed.), *I, Robot*, New York City: Doubleday, 1950, tr. 40

đề phức tạp liên quan đến đạo đức của trí tuệ nhân tạo. Ngày nay, nhiều tiêu chuẩn toàn cầu về đạo đức AI được thiết lập trên thế giới, nổi bật là nghiên cứu của các tổ chức quốc tế OECD, EU và UNESCO. Các tiêu chuẩn này tập trung chủ yếu vào tính minh bạch, công bằng và đảm bảo cao nhất các quyền con người, dân khắc phục được những hạn chế trong định luật của Asimov, định hướng xây dựng một Bộ quy tắc Đạo đức về AI hoàn thiện cho toàn thế giới nhằm định khung cho hành vi của AI, bảo vệ tương lai bền vững của nhân loại.

2. Mô hình đạo đức AI của UNESCO

Theo UNESCO đạo đức AI được định nghĩa là “Sự phản ánh có tính hệ thống và tính quy phạm một cách toàn diện, tổng thể và theo khung phát triển các giá trị, nguyên tắc và hành động phụ thuộc lẫn nhau nhằm hướng dẫn cộng đồng, xã hội đối phó một cách có trách nhiệm với các tác động đã biết và chưa biết của công nghệ AI”.¹² UNESCO đã hiện thực hóa sự phản ánh này thông qua mô hình đạo đức AI của mình. Mô hình được ghi nhận tại Khuyến nghị của UNESCO về đạo đức của AI (*Recommendation on the Ethics of Artificial Intelligence*). Khuyến nghị này ra đời ngày 23/11/2021 với sự thông qua của toàn bộ 193 quốc gia thành viên tại Kỳ họp Đại hội đồng UNESCO lần thứ 41. Đây là văn bản đầu tiên mang tính chuẩn mực toàn cầu do một tổ chức quốc tế ban hành nhằm đưa ra những khuyến nghị về đạo đức trong AI.¹³ Khuyến nghị này gồm 8 Chương với 141 điểm nội dung trình bày rõ ràng và toàn diện các vấn đề về đạo đức AI. Trong đó, nội dung về mô hình đạo đức AI được thể hiện trọng tâm ở các chương III – Giá trị và nguyên tắc, IV – Các lĩnh vực hoạt động chính sách và V – Giám sát và đánh giá. Cấu trúc của mô hình đạo đức AI của UNESCO được hình thành từ bốn tầng chính: (i) các giá trị nền tảng; (ii) các nguyên tắc đạo đức (iii) các công cụ thực thi và (iv) các cơ chế giám sát.

Thứ nhất, theo mô hình đạo đức AI của UNESCO, bốn giá trị cốt lõi làm nên đạo đức AI trong suốt vòng đời tồn tại của chúng là phải đảm bảo: tôn trọng, bảo vệ và thúc đẩy nhân quyền, các quyền tự do cơ bản và phẩm giá con người; môi trường và hệ sinh thái phát triển mạnh mẽ; đảm bảo tính đa dạng và hòa nhập; sống trong xã hội hòa bình, công bằng và gắn kết. Đây là phần đạo đức gốc của AI, đặt nền móng cho hệ thống AI hoạt động, tồn tại để đem lại lợi ích cho con người, xã hội và môi trường.

Thứ hai, để đạt được những giá trị cốt lõi ấy, mô hình của UNESCO đã đặt ra 10 nguyên tắc mà những nhà sáng lập, phát triển, người sử dụng, nhà quản lý và tất cả những hoạt động liên quan đến AI phải tuân theo gồm: tính tương 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; sự giám sát và quyết định của con người; tính minh bạch và khả năng giải thích; Trách nhiệm và giải trình; nhận thức và hiểu biết; quản trị và hợp tác một cách đa phương và thích ứng. Các nguyên tắc này trước đây đã từng được nhắc đến nhiều lần trong phát triển AI, tiêu biểu là năm 2010, Vương quốc Anh đã xây dựng

12 Nguyễn Đức Thủy “Đạo đức trong ứng dụng trí tuệ nhân tạo”, *Tạp chí Thông tin và Truyền thông*, số 4, 2023, <https://vjst.vn/dao-duc-trong-ung-dung-tri-tue-nhan-tao-57064.html>, truy cập ngày 9/10/2025.

13 Hà Quang Thủy, 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”, https://uet.vnu.edu.vn/~thuyhq/PPNCKH/TTNT_dao-duc.pdf, truy cập ngày 9/10/2025.

một bộ Nguyên tắc về Robot với nội dung tương tự.¹⁴ Tuy nhiên so với các diễn đạt một cách mơ hồ trước đây, UNESCO đã có sự diễn giải, hướng dẫn chi tiết cho từng quy tắc. Nếu các giá trị đạo đức AI là cốt lõi thì các nguyên tắc đạo đức AI là sự cụ thể hóa, vạch ra đường hướng rõ ràng cho việc phát triển AI một cách bền vững.

Thứ ba, khác với các bộ nguyên tắc thông thường, mô hình đạo đức AI của UNESCO không chỉ đặt ra những nguyên tắc lý thuyết mà còn lập ra một tầng các công cụ thực thi với 11 nhóm lĩnh vực chính sách cụ thể: đánh giá tác động đạo đức; quản trị và quản lý 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. Dựa trên những nguyên tắc đạo đức AI, khi áp dụng các công cụ thực thi này vào mỗi lĩnh vực khác nhau sẽ mang tính đặc thù, chuyên biệt và việc quy định các lĩnh vực chính sách này đảm bảo sự hướng dẫn đủ chi tiết, là hành động cụ thể trên đường hướng đã vạch.

Thứ tư, khi đã có nền tảng cốt lõi, có đường hướng rõ ràng, hành động cụ thể thì cái cần thiết cuối cùng chính là công cụ, cơ chế giám sát nhằm đảm bảo, thúc đẩy thực thi và đánh giá tác động của chính sách. UNESCO trao quyền cho các quốc gia thành viên được tự đặt ra các điều kiện, cơ cấu quản lý, các điều khoản pháp luật của mình để đánh giá các chính sách một cách minh bạch và đáng tin cậy, kết hợp giữa định lượng và định tính. Tuy nhiên, UNESCO cũng hỗ trợ các quốc gia bằng cách đặt trước các nền móng ban đầu như: xây dựng phương pháp luận của UNESCO về Đánh giá tác động đạo đức AI (*Ethical Impact Assessment, EIA*); xây dựng phương pháp đánh giá mức độ sẵn sàng của các thành viên (*Readiness assessment methodology, RAM*); phát triển phương pháp luận của UNESCO để đánh giá trước và sau về hiệu quả của các chính sách; tăng cường nghiên cứu và phân tích dựa trên bằng chứng và báo cáo về chính sách; thu thập và phổ biến các thông tin về chính sách để hỗ trợ chia sẻ và học hỏi lẫn nhau. Với nền tảng và những tài nguyên đó, UNESCO mong muốn sẽ thúc đẩy để quá trình xây dựng chính sách riêng của mỗi quốc gia được nhanh chóng, đồng bộ và thiết thực.

Đồng thời, UNESCO cũng đưa ra những lưu ý về quá trình giám sát và đánh giá chính sách. Một là quá trình này phải đảm bảo sự tham gia rộng rãi của tất cả các bên liên quan, bao gồm và không giới hạn những người dễ bị tổn thương hoặc trong hoàn cảnh dễ bị tổn thương nhằm đem lại sự đa dạng, tính bao trùm và dân chủ trong giám sát chính sách đạo đức AI. Hai là, cần xây dựng các công cụ và chỉ số phù hợp để đánh giá hiệu quả chính sách so với các tiêu chuẩn đã thống nhất bao gồm các mục tiêu cụ thể dành cho nhóm yếu thế, nhóm thiểu số và nhóm dễ bị tổn thương. Ba là, việc đánh giá này phải thực hiện liên tục, có hệ thống, bao gồm việc tự đánh giá, đánh giá các tổ chức, nhà cung cấp và chương trình tư nhân công cộng.

Đặc biệt, khuyến nghị theo mô hình của UNESCO cũng đề xuất các cơ chế giám sát và đánh giá khả thi cho các quốc gia gồm: ủy ban đạo đức, đài quan sát toàn cầu về đạo đức AI, kho lưu trữ về sự phát triển tuân thủ đạo đức AI và sáng kiến giải quyết trong các lĩnh vực, cơ chế chia sẻ kinh nghiệm, khuôn khổ thử nghiệm quy định về AI (*sandbox*) và hướng dẫn đánh giá việc tuân thủ các khuyến nghị cho tất cả các bên liên quan.

Khuyến nghị về đạo đức AI của UNESCO đã thể hiện nỗ lực hợp tác quốc tế, thảo luận, đàm phán, tham vấn rộng rãi về các quan điểm khác nhau giữa các quốc gia

14 Stuart J. Russell, Peter Norvig, Ming-Wei Chang, *ltd.*

để cuối cùng dẫn đến sự thống nhất, thỏa hiệp cao độ nhằm thông qua khuyến nghị cuối cùng.¹⁵ Với mục tiêu thúc đẩy các giá trị cơ bản trong đó cốt lõi là nhân phẩm con người – nền tảng cho mọi quyền con người, cho công lý xã hội, mô hình đạo đức AI của UNESCO là một giải pháp pháp lý và đạo đức mới mang tính chuyển đổi trong thời đại công nghệ hiện nay.¹⁶ Điểm đặc biệt của mô hình đạo đức AI theo UNESCO là bên cạnh những nguyên tắc, phương hướng chiến lược và hành động, mô hình này còn cung cấp công cụ để thực thi, giám sát và điều tiết có hiệu quả các nguyên tắc đó. Bản khuyến nghị này tồn tại như một luật mềm, cho phép các quốc gia tự nguyện tiếp thu để xây dựng chính sách về AI.¹⁷ Tuy nhiên, tính chất “mềm” này cũng kèm theo một số cơ chế như khuyến nghị các quốc gia thực hiện báo cáo, đánh giá đạo đức, giám sát quốc tế giúp mở rộng đối tượng tiếp nhận nhưng vẫn giữ được tính hiệu lực và định hướng thống nhất.

3. Vận dụng mô hình đạo đức AI của UNESCO trong xây dựng chính sách ở Việt Nam hiện nay

Hiện nay, Việt Nam đang trong quá trình rà soát, hoàn thiện thể chế, chính sách nhằm tạo hành lang thông thoáng, vững chắc cho mục tiêu phát triển AI bền vững. Nghị quyết số 57-NQ/TW ngày 22/12/2024 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia đã thể hiện tầm nhìn chiến lược và quyết tâm đột phá về chủ trương và chính sách để phát triển công nghệ, chuyển đổi số nhằm mục tiêu đến năm 2030, Việt Nam trở thành nhóm ba nước dẫn đầu khu vực Đông Nam Á về nghiên cứu và phát triển trí tuệ nhân tạo. Ngày 11/6/2024, Bộ Khoa học và Công nghệ cũng đã ban hành Quyết định số 1290/QĐ-BKHCN về việc hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các hệ thống AI có trách nhiệm. Tuy nhiên, công cụ đánh giá mức độ sẵn sàng (*Readiness Assessment Methodology*, RAM) của UNESCO cho thấy, tại Việt Nam, khía cạnh đạo đức trong hoạt động AI chưa được đề cập đúng mức trong các chiến lược cũng như văn bản pháp luật nào.¹⁸ Mặc dù sự ra đời của Ủy ban Đạo đức AI thuộc Hiệp hội Phần mềm và dịch vụ công nghệ thông tin Việt Nam ngày 5/12/2024 đã thể hiện nỗ lực rõ ràng nhưng đây là một hoạt động của khu vực tư nhân.¹⁹ Về mặt chính sách quốc gia, chúng ta dường như đang tập trung vào mặt khuyến khích mà chưa có văn bản nào thật sự trở thành nền tảng cho đạo đức AI. Do đó, với tư cách là thành viên tích cực trong UNESCO, cùng với lợi thế hạ tầng kỹ thuật, nhân lực, tài chính, thị trường, Việt Nam có thể tiếp thu khuyến nghị của tổ chức này một cách có hệ thống, chọn lọc và sáng tạo phù hợp với tình hình trong nước. Một số nội dung có thể vận dụng từ mô hình đạo đức AI của UNESCO là:

Một là, xây dựng một bộ nguyên tắc đạo đức AI chung, thống nhất, chi tiết, phù hợp với đặc trưng kinh tế – văn hóa – xã hội Việt Nam. Một trong những nguyên tắc của đạo đức AI là lấy con người làm trung tâm, triển khai để cải thiện nhân loại, nâng

15 Michal Naturski, “Compromise in multilateral negotiations and the global regulation of artificial intelligence”, *Democratization*, Vol. 31(5), 2024, tr. 1092. DOI: <https://doi.org/10.1080/13510347.2024.2351583>

16 UNESCO, “Shaping the Future: IFAP webinar explores ethical AI and inclusive data governance”, <https://www.unesco.org/en/articles/shaping-future-ifap-webinar-explores-ethical-ai-and-inclusive-data-governance?hub=32618>, truy cập ngày 9/10/2025.

17 Trần Việt Dũng, “Bàn về phát triển quy phạm đạo đức trong tiến trình xây dựng pháp luật về trí tuệ nhân tạo”, *Tạp chí Khoa học pháp lý Việt Nam*, số 9, 2024, tr. 38.

18 Ngọc Diệp, “Đặt đạo đức và tính bao trùm làm trọng tâm trong phát triển AI”, *VJST*, <https://vjst.vn/dat-dao-duc-va-tinh-bao-trum-lam-trong-tam-trong-phat-trien-ai-69736.html>, truy cập ngày 9/10/2025.

19 Lê Lâm, “Ra mắt ủy ban Đạo đức trí tuệ nhân tạo”, *Báo Nhân Dân*, <https://nhandan.vn/ra-mat-uy-ban-dao-duc-tri-tue-nhan-tao-post848838.html>, truy cập ngày 9/10/2025.

cao phúc lợi (hoặc ít nhất là không gây ảnh hưởng tiêu cực đến phúc lợi), phẩm giá con người là sự phát triển thịnh vượng của con người.²⁰ Vì vậy, xây dựng những nguyên tắc cốt lõi dựa trên các nguyên tắc cơ bản về nhân quyền là rất cần thiết. Có thể học hỏi 10 nguyên tắc của mô hình UNESCO như tính bảo mật, an toàn, không biệt đối xử, trách nhiệm... AI dần trở thành một ngành nghề hiện đại, và tương tự như các ngành nghề khác, việc ban hành một đạo đức nghề nghiệp có tính định vị sẽ là kim chỉ nam cho những người hành nghề và khách hàng của họ.²¹ Đồng thời, các nguyên tắc này phải được điều chỉnh cho phù hợp với thực tiễn pháp luật, thể chế chính trị của Việt Nam.

Hai là, xây dựng các quy tắc trong từng lĩnh vực, bao hàm cả quy tắc riêng của lĩnh vực đó. Tương tự như mô hình của UNESCO với 11 lĩnh vực hành động cụ thể, bộ quy tắc AI Việt Nam cần xác định các lĩnh vực nòng cốt trong hoạt động AI như Đánh giá tác động đạo đức; quản trị đạo đức; chính sách dữ liệu; hợp tác quốc tế; môi trường; giới tính; văn hóa; giáo dục; truyền thông; kinh tế; lao động; y tế; khoa học công nghiệp... Từ đó, xây dựng các nguyên tắc đặc trưng phù hợp về sử dụng AI trong từng lĩnh vực. Tuy nhiên, một điểm yếu lớn của nguyên tắc đạo đức AI là sự thiếu hụt tương đối về lịch sử thực nghiệm trong từng ngành nghề và các chuẩn mực rõ ràng về hành vi được coi là “tốt”.²² Trong khi các ngành nghề khác, điển hình là y học có lịch sử lâu đời và hình thành nên văn hóa chuyên biệt với các tiêu chuẩn toàn diện về Đạo đức Y khoa. Do đó, khi xây dựng nguyên tắc AI trong từng lĩnh vực, cần kế thừa các chuẩn mực đạo đức đã được thừa nhận trong từng lĩnh vực (như đạo đức công vụ, đạo đức y tế, đạo đức giáo dục, đạo đức khoa học và công nghệ...), đồng thời bổ sung các yêu cầu đặc thù đối với việc phát triển và sử dụng AI để đảm bảo tính thống nhất của hệ thống pháp luật.

Ba là, ban hành những bộ tiêu chuẩn kỹ thuật, bộ công cụ đo lường, những chỉ số cụ thể để kiểm tra, kiểm toán về đạo đức AI. Mặc dù xây dựng bộ nguyên tắc là bước đầu cốt lõi nhưng nguyên tắc không thể tự động trở thành thực tiễn mà phải được chuyển hóa thành các yêu cầu thực tế và chuẩn mực thực hành. AI tồn tại dưới dạng những thiết kế kỹ thuật, do đó, cần có những chỉ số, tiêu chuẩn kỹ thuật để hướng dẫn và đánh giá các sản phẩm AI một cách rõ ràng, thông qua nhiều hình thức ở từng giai đoạn thiết kế, phát triển và triển khai. Cần quan tâm rằng, cùng với việc xác định các chuẩn mực, chúng ta cũng xác định thứ tự ưu tiên, phân cấp theo tính chất quan trọng của từng chuẩn mực (ví dụ: tiêu chuẩn cấm gây hại cho con người thường có giá trị hơn lệnh cấm nói dối).²³ Bên cạnh đó, cần cung cấp bộ công cụ đo lường, đánh giá mức độ tuân thủ, là cơ sở để kiểm toán, kiểm định những thiên lệch tiêu chuẩn đạo đức của AI. Bộ công cụ này có thể gồm các chức năng: cung cấp chương trình kiểm soát các thuật toán, cung cấp phần mềm ngắt mạch để ngăn chặn vi phạm đạo đức, thiết lập ranh giới sử dụng, xây dựng lòng tin giữa các bên, chứng minh trách nhiệm của nhà phát triển hệ thống AI.²⁴ Và cũng như UNESCO đã khuyến nghị, cần có sự

20 Emre Kazim, Adriano Soares Koshiyama, “A high-level overview of AI ethics”, *Patterns*, Vol. 2(9), 2021, tr. 8. DOI: <https://doi.org/10.1016/j.patter.2021.100314>

21 Alexander Filipović, Christopher Koska, Claudia Paganini, *Developing a professional ethics for algorithmists: Learning from the examples of established ethics*, Bertelsmann Stiftung, 2018, tr. 42.

22 Brent Mittelstadt, “Principles alone cannot guarantee ethical AI”, *Nature Machine Intelligence*, Vol. 1, 2019, tr. 6. DOI: <https://doi.org/10.1038/s42256-019-0114-4>

23 IEEE, *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (Version 2)*, IEEE, 2019, tr. 40.

24 Emre Kazim, Adriano Koshiyama, *No AI regulator: An analysis of artificial intelligence and public standards report*, UK, 2020, tr. 3.

tham khảo các tiêu chuẩn quốc tế, trao đổi, hợp tác quốc tế để có thể xây dựng được những chỉ số, tiêu chuẩn thống nhất.

Bốn là, thực hiện đánh giá tác động đạo đức AI (*Ethical Impact Assessment, EIA*) trong cộng đồng. Đây là việc xem xét ảnh hưởng của đạo đức AI lên môi trường số, các thói quen sinh hoạt của cộng đồng và môi trường sinh thái, từ đó tiến hành giảm thiểu những tác hại tiềm ẩn của hệ thống AI. Khác với các kiểm toán công nghệ mang tính kỹ thuật ở trên, đánh giá tác động thiên về xã hội, bao gồm đánh giá về quyền công dân, quyền con người, sức khỏe tâm lý và xã hội, tác động đến kinh tế và môi trường. Có thể kể đến các cách thức đánh giá như khảo sát, thu thập phản ánh, tổng hợp và phân tích các thông tin về các hiện tượng, hành vi liên quan đến đạo đức của các hệ thống, nền tảng, sản phẩm, dịch vụ có ứng dụng công nghệ AI như mạng xã hội, nền tảng thương mại điện tử, tuyển dụng, các ứng dụng cho vay tín dụng, các ứng dụng cung cấp dịch vụ công, ... So với đánh giá tác động đạo đức AI thì đánh giá tác động bảo vệ dữ liệu (đạo đức dữ liệu) hiện đang phát triển và có khung pháp lý rõ ràng hơn và là mô hình lý tưởng để tham khảo.²⁵

Năm là, tiến hành hoạt động giám sát thực thi với sự tham gia của tất cả các bên liên quan đến phát triển AI. Theo như UNESCO khuyến nghị, việc giám sát, đánh giá thực thi phải được thực hiện một cách toàn diện, có hệ thống thông qua việc tự đánh giá, đánh giá các tổ chức công cộng và tư nhân, nhà cung cấp, các bên trung gian, người tiêu dùng, ... Các mô hình cụ thể đã được gợi ý theo UNESCO như ủy ban đạo đức, đài quan sát đạo đức, kho lưu trữ sáng kiến, khuôn khổ thử nghiệm quy định về AI, hướng dẫn đánh giá đều là những cơ chế hữu ích nhằm đảm bảo tuân thủ đạo đức AI Việt Nam. Hoạt động này gắn liền với quản trị AI, đặc biệt là ở các doanh nghiệp. Bên cạnh người sáng lập, vai trò giám sát của nhiều bên trong suốt quá trình AI hoạt động cũng là cơ sở truy cứu trách nhiệm đối với các thiệt hại.²⁶ Đây là một cơ chế quản trị thực tiễn giúp chuyển hóa các nguyên tắc đạo đức AI từ các định hướng mang tính lý thuyết thành hệ thống giám sát và thực thi bởi nhiều chủ thể trong quá trình vận hành hệ thống AI.

Sáu là, đào tạo đội ngũ đạo đức AI. Với mục tiêu đến năm 2030, Việt Nam tiến đến nhóm quốc gia dẫn đầu có thu nhập trung bình cao, công tác giáo dục và đào tạo, bồi dưỡng nhân tài về AI là một yêu cầu thiết yếu.²⁷ Đặc biệt, về đạo đức AI là một nội dung chủ đạo, nền tảng định hướng mọi hoạt động phát triển AI. Các nhà thiết kế AI hiện nay đang được đào tạo về khoa học máy tính, kỹ thuật hơn là hiểu biết về những tác động xã hội liên quan đến AI. Trong khi, để xây dựng một hệ thống AI theo chuẩn đạo đức thì vấn đề không chỉ nằm ở một hoặc vài tiêu chí kỹ thuật mà còn phức tạp hơn nhiều khiến AI cũng có thể mang thiên kiến của chính người tạo ra nó.²⁸ Ví dụ, một AI được lập trình là không thiên vị về mặt giới tính có thể bảo vệ được phụ nữ nhưng không có nghĩa là bảo vệ tốt quyền lợi của phụ nữ da đen hoặc châu

25 Emre Kazim, Adriano Koshiyama, “The interrelation between data and AI ethics in the context of impact assessments”, *AI and Ethics*, Vol. 1, 2020, tr. 222. DOI: <https://doi.org/10.1007/s43681-020-00029-w>

26 “Đạo đức AI – nền tảng xây dựng niềm tin trong kỷ nguyên trí tuệ nhân tạo”, *VietnamNet*, https://vietnamnet.vn/dao-duc-ai-nen-tang-xay-dung-niem-tin-trong-ky-nguyen-tri-tue-nhan-tao-2407517.html?gidl=Tx-lAZ2ul7LseQibDg-xUd2ggImh-Rb_RAsfUlcjkovfghvuAF-xB6xwhob-gEijQVskUMHrDoPdCBcyU0, truy cập ngày 9/10/2025.

27 Quyết định số 127/QĐ-TTg của Thủ tướng Chính phủ: Ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng Trí tuệ nhân tạo đến năm 2030.

28 Nguyễn Thị Quế Anh, Vũ Công Giao, Mai Văn Thắng, *Trí tuệ nhân tạo với pháp luật và quyền con người*, Nxb. Tư pháp, 2019, tr. 358.

Á. Do đó, đạo đức AI đòi hỏi nhà phát triển không chỉ loại bỏ thiên kiến đối với từng nhóm riêng biệt mà còn phải đánh giá tác động của AI đối với các nhóm mang nhiều đặc điểm nhận dạng khác nhau nhằm bảo đảm tính công bằng và không phân biệt đối xử.²⁹ Việc lồng ghép giáo dục về đạo đức AI và tác động xã hội của nó vào đào tạo kỹ thuật là vô cùng quan trọng. Bên cạnh đó, việc đào tạo các ngành nghề, lĩnh vực khác về đạo đức AI cũng góp phần chuẩn bị cho một tương lai phổ biến AI an toàn, có trách nhiệm trong tương lai. ●

Tài liệu tham khảo

- [1] Nguyễn Thị Quế Anh, Vũ Công Giao, Mai Văn Thắng, *Trí tuệ nhân tạo với pháp luật và quyền con người*, Nxb. Tư pháp, 2019 [trans: Nguyen Thi Que Anh, Vu Cong Giao & Mai Van Thang, *Artificial intelligence, law and human rights*, Justice Publishing House, 2019]
- [2] Kakembo Aisha Annet, “The ethics of AI: Philosophical perspectives”, *Research Invention Journal of Research in Education*, Vol. 5(2), 2025. DOI: <https://doi.org/10.59298/RIJRE/2025/526571>
- [3] Jason Borenstein, “AI ethics: A long history and a recent burst of attention”, *Computer*, Vol. 54(1), 2021. DOI: <https://doi.org/10.1109/MC.2020.3037520>
- [4] Trần Việt Dũng, “Bàn về phát triển quy phạm đạo đức trong tiến trình xây dựng pháp luật về trí tuệ nhân tạo”, *Tạp chí Khoa học pháp lý Việt Nam*, số 9, 2024 [trans: Tran Viet Dung, “On the development of ethical norms in the process of building law on artificial intelligence”, *Vietnam Journal of Legal Sciences*, No. 9, 2024]
- [5] Alexander Filipović, Christopher Koska, Claudia Paganini, *Developing a professional ethics for algorithmists: Learning from the examples of established ethics*, Bertelsmann Stiftung, 2018
- [6] John-Stewart Gordon và Sven Nyholm, “Ethics of artificial intelligence”, *Internet Encyclopedia of Philosophy*, 2021
- [7] IEEE, *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2), IEEE, 2019
- [8] Emre Kazim, Adriano Koshiyama, “The interrelation between data and AI ethics in the context of impact assessments”, *AI and Ethics*, Vol. 1, 2020. DOI: <https://doi.org/10.1007/s43681-020-00029-w>
- [9] Emre Kazim, Adriano Soares Koshiyama, “A high-level overview of AI ethics”, *Patterns*, Vol. 2(9), 2021. DOI: <https://doi.org/10.1016/j.patter.2021.100314>
- [10] Emre Kazim, Adriano Koshiyama, *No AI regulator: An analysis of artificial intelligence and public standards report*, UK, 2020
- [11] Brent Mittelstadt, “Principles alone cannot guarantee ethical AI”, *Nature Machine Intelligence*, Vol. 1, 2019. DOI: <https://doi.org/10.1038/s42256-019-0114-4>
- [12] Michal Natonski, “Compromise in multilateral negotiations and the global regulation of artificial intelligence”, *Democratization*, Vol. 31(5), 2024. DOI: <https://doi.org/10.1080/13510347.2024.2351583>
- [13] Stuart J. Russell, Peter Norvig, Ming-Wei Chang, *Artificial intelligence: A modern approach* (4th edition, Global Edition), Pearson Education Limited, 2022
- [14] Bernd Carsten Stahl, Doris Schroeder, Rowena Rodrigues, *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges*, Springer Briefs in Research and Innovation Governance, 2023
- [15] Nguyễn Đức Thủy “Đạo đức trong ứng dụng trí tuệ nhân tạo”, *Tạp chí Thông tin và Truyền thông*, số 4, 2023 [trans: Nguyen Duc Thuy “Ethics in the application of artificial intelligence”, *Information and Communication Magazine*, No. 4, 2023]
- [16] Vernor Vinge, “The coming technological singularity: How to survive in the post-human era”, *Whole Earth Review*, Winter Issue, 1993

29 Jason Borenstein, “AI ethics: A long history and a recent burst of attention”, *Computer*, Vol. 54(1), 2021, tr. 98. DOI: <https://doi.org/10.1109/MC.2020.3037520>