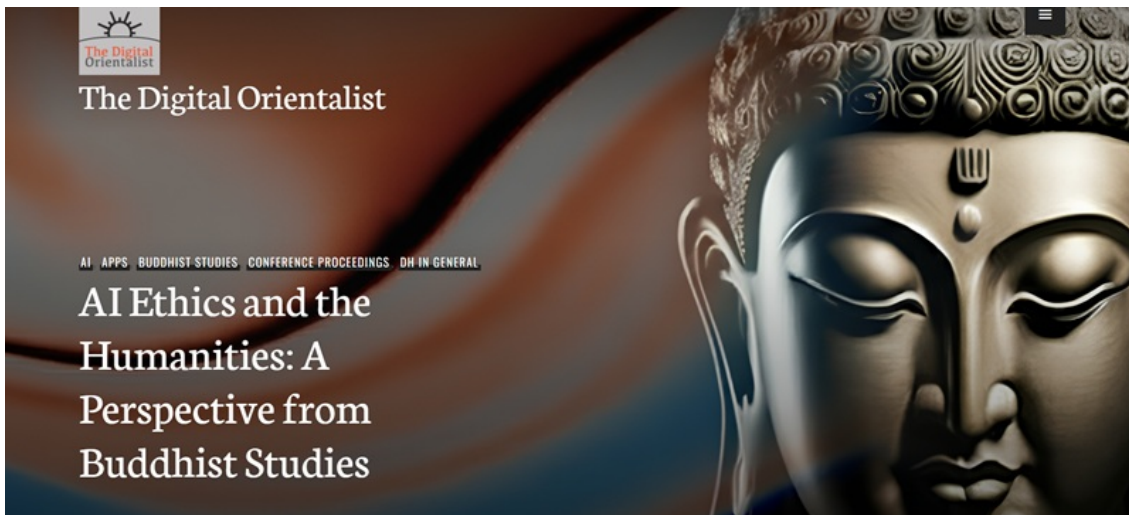


Góc nhìn Phật học về đạo đức AI và nhân văn học

ISSN: 2734-9195 12:27 02/02/2026

Sau cùng, mỗi khi sử dụng hay tiêu thụ AI, chúng ta cần luôn ý thức về những tác động lớn hơn của nó: xã hội, môi trường, các thiên kiến tiềm ẩn và nhiều hệ quả khác.

Bài viết được phát triển từ tham luận trình bày tại Hội nghị trực tuyến năm 2025 của The Digital Orientalist (tạm dịch: Nhà Phương Đông Kỹ thuật số) với chủ đề **AI và Nhân văn số**, do diễn giả chính Elaine Lai (Civic, Liberal, and Global Education - COLLEGE, Đại học Stanford) thực hiện.



Mở đầu

Các báo cáo gần đây của PwC, McKinsey và Diễn đàn Kinh tế Thế giới đều dự báo rằng **trí tuệ nhân tạo (AI)** sẽ tạo ra những chuyển biến căn bản đối với lực lượng lao động toàn cầu vào năm 2050, với khoảng 60% số công việc hiện nay cần được điều chỉnh đáng kể do tác động của AI (Kelly 2025). Tuy nhiên, mốc thời gian 2050 có thể vẫn còn quá lạc quan. Gần đây, Giám đốc điều hành của Anthropic, ông Dario Amodei, cho rằng làn sóng thất nghiệp quy mô lớn trong các vị trí văn phòng ở cấp độ khởi điểm có thể xảy ra chỉ trong vòng 5 năm tới, đồng thời kêu gọi Chính phủ Hoa Kỳ sớm có những biện pháp quản lý

AI một cách thận trọng (Binder 2025).

Dù tham chiếu số liệu từ nguồn nào, câu chuyện chủ đạo vẫn không thay đổi: AI là một thực tại không thể đảo ngược hoặc thích nghi, hoặc bị bỏ lại phía sau.

Trong bối cảnh đó, giới học giả nhân văn cần nhìn nhận một cách phản biện và có trách nhiệm về việc tích hợp AI, **اوس** như một phương pháp nghiên cứu hay như một đối tượng nghiên cứu, nhằm duy trì khả năng đối thoại ý nghĩa với thế giới đương đại. Tiền đề của bài tham luận này xuất phát từ hai phương diện trong chính căn tính nghề nghiệp của tác giả: Thứ nhất, với tư cách là một học giả nghiên cứu Phật học, ứng dụng nhân văn số để mở rộng khả năng tiếp cận công trình nghiên cứu tới công chúng rộng rãi hơn; và Thứ hai, với vai trò là giảng viên tại Đại học Stanford – trung tâm của Thung lũng Silicon, nơi tác giả thường xuyên giảng dạy cho sinh viên đại học về đạo đức công nghệ.

Bài trình bày được chia thành ba phần. Phần thứ nhất, “Phật giáo và AI: rủi ro và lợi ích”, phác họa tổng quan về cách các cộng đồng Phật giáo và giới dịch thuật đã sử dụng AI để xây dựng các chatbot tôn giáo và công cụ hỗ trợ dịch thuật. Phần thứ hai, “Công cụ AI dành cho học giả nhân văn”, đưa ra một số gợi ý về việc tích hợp AI một cách có đạo đức vào quy trình làm việc học thuật. Phần thứ ba, “Những đối thoại rộng hơn về đạo đức AI”, bàn luận mười một vấn đề đạo đức xoay quanh AI mà giới học giả nhân văn có thể tiếp tục đào sâu và tham gia.

Phần 1: Phật giáo và AI - Rủi ro và lợi ích

Năm 2019, một dự án trị giá 1 triệu USD được triển khai giữa nhà nghiên cứu robot Hiroshi Ishiguro và Phó giáo sư kỹ thuật Kohei Ogawa (Đại học Osaka) nhằm tạo ra Mindar, một robot có kích thước tương đương con người, được thiết kế để thuyết giảng giáo lý Phật giáo trong khuôn viên chùa Kodaiji, ngôi chùa hơn 400 năm tuổi tại Nhật Bản (Hardingham-Gill 2019). Robot này được mô phỏng theo hình tượng Bồ-tát Quán Thế Âm (Kannon), biểu trưng cho lòng từ bi trong Phật giáo. Dù Mindar khi đó chưa sử dụng AI, dự án này có thể được xem là bước tiền đề cho cách Phật giáo về sau tiếp cận và khai thác công nghệ này.

Đến năm 2023, một chatbot AI mang tên Roshibot được phát triển bởi Jiryu Rutschman-Byler, một vị tăng thuộc Thiền phái Tào Động (Soto Zen) và đồng trụ trì Thiền viện San Francisco Zen Center. Chatbot này được huấn luyện dựa trên các trước tác và pháp thoại ghi âm của Thiền sư Shunryu Suzuki Roshi, kết hợp với dữ liệu từ Internet. Roshibot được tạo ra như một thử nghiệm nhằm khảo sát tiềm năng của AI trong việc hỗ trợ giáo lý và thực hành Phật giáo (Rutschman-Byler 2023). Trước khi sử dụng, người dùng buộc phải thừa nhận

những giới hạn của công nghệ, bao gồm khả năng chatbot có thể “ảo giác” thông tin hoặc đưa ra phát ngôn gây tổn hại. Những cảnh báo này phản ánh rõ nét các mối bận tâm đạo đức khi đặt AI trong mối quan hệ với Phật giáo.



Hình minh họa được tạo bởi AI

Cũng trong năm đó, tại Malaysia, một doanh nhân và giảng sư Phật giáo tên Lim Kooi Fong đã giới thiệu chatbot Phật giáo của riêng mình tại Sarnath (Ấn Độ). NORBU - viết tắt của *Neural Operator for Responsible Buddhist Understanding* (tạm dịch: Hệ vận hành trí tuệ thần kinh vì sự thấu hiểu Phật học có trách nhiệm) - được xây dựng dựa trên nền tảng công nghệ của ChatGPT.

Mặc dù các sáng kiến **ứng dụng AI** trong Phật giáo thường xuất phát từ thiện chí, việc bảo đảm sự tham gia thực chất của các cộng đồng Phật giáo và giới giáo thọ trong suốt quá trình thiết kế, thử nghiệm và triển khai là điều hết sức quan trọng. Buddhobot Plus (2025), dự án hợp tác giữa Tăng đoàn Trung ương Bhutan và Đại học Kyoto (Nhật Bản) có thể xem là một ví dụ cho cách tiếp cận bao dung và thận trọng hơn trong phát triển AI (Lewis 2025). Tương tự NORBU, phiên bản Buddhobot Plus năm 2023 cũng được xây dựng dựa trên AI sinh ngữ của ChatGPT. Tuy nhiên, hệ thống này sẽ được các tu sĩ thuộc Tăng đoàn Trung ương Bhutan trực tiếp giám sát, đồng thời trải qua quy trình đánh giá an toàn nghiêm ngặt trước khi chính thức đưa vào sử dụng rộng rãi. Dù điều này có thể làm chậm tiến độ phát hành, những cách tiếp cận lấy cộng đồng làm trung tâm như vậy có khả năng mang lại lợi ích tổng thể lớn hơn và giảm thiểu nguy cơ rủi ro nghiêm trọng.

Hướng phát triển lớn thứ hai tại giao điểm giữa Phật giáo và AI là sự ra đời của các công cụ dịch thuật mới, đặc biệt trong lĩnh vực dịch từ tiếng Tây Tạng sang

tiếng Anh. Gần đây, các công ty công nghệ tại Thung lũng Silicon như Anthropic và OpenAI đã cải thiện đáng kể mô hình dịch thuật đối với tiếng Tây Tạng. Tại Ấn Độ, nền tảng Monlam AI do Geshe Monlam đứng đầu, với sự tài trợ của Tibet Fund và USAID đang phát triển các công nghệ dịch thuật, chuyển văn bản thành giọng nói, chuyển giọng nói thành văn bản và nhận dạng ký tự quang học (OCR) cho tiếng Tây Tạng sang tiếng Anh. Mới đây nhất, một nền tảng tư nhân mang tên Dharmamitra đã ra mắt hệ thống dịch máy cho các ngôn ngữ kinh điển như Hán cổ, Sanskrit, Pali và Tây Tạng sang tiếng Anh, tiếng Hàn và một số ngôn ngữ khác. Mục tiêu cốt lõi của Dharmamitra là cho phép tìm kiếm đa ngôn ngữ trong các ngôn ngữ nguồn của Phật giáo, chẳng hạn như tìm câu Pali thông qua truy vấn bằng tiếng Hán, hoặc sử dụng tiếng Anh để xác định những đoạn văn tương ứng trong Đại Tạng Tây Tạng.

Phản ứng đối với các công nghệ dịch thuật này khá đa dạng. Ở một thái cực, có ý kiến cho rằng cần cấm hoàn toàn việc sử dụng AI trong dịch thuật. Ví dụ, trên nền tảng SuttaCentral dự án chuyên về kinh điển Phật giáo sơ kỳ đặt trụ sở tại Úc – đã xuất hiện một bài blog lan truyền mạnh với tiêu đề “Hãy để SuttaCentral mãi mãi 100% không có AI” (Sujato 2024). Ngược lại, những tổ chức Phật giáo khác như 84000 và Tsadra Foundation tỏ ra cởi mở hơn, đồng thời chủ động xây dựng các bộ hướng dẫn nhằm sử dụng AI một cách có đạo đức (Tsadra Foundation 2024; 84000 2024). Dựa trên những phản hồi khác nhau này, có thể khái quát các rủi ro và lợi ích của công nghệ dịch thuật AI theo những điểm sau đây.

Từ những phản hồi đa chiều đối với các công nghệ dịch thuật AI trong Phật giáo, có thể khái quát các lợi ích và rủi ro theo một số phương diện cơ bản sau đây.

Về lợi ích, trước hết là yếu tố tốc độ. Nếu được triển khai một cách hợp lý và có kiểm soát, AI có thể đẩy nhanh đáng kể quá trình dịch thuật. Các công cụ căn chỉnh văn bản đặc biệt hữu ích đối với những dịch giả làm việc đồng thời với nhiều ngôn ngữ khác nhau. Bên cạnh đó, AI góp phần nâng cao khả năng tiếp cận các nguồn văn bản, thông qua việc cung cấp bản dịch sơ bộ giúp người đọc hình dung được cấu trúc tổng thể và nội dung chính của văn bản gốc. Ở kịch bản lý tưởng nhất, công nghệ này còn có thể dân chủ hóa giáo dục, thu hút nhiều đối tượng hơn từ người thực hành, học giả đến dịch giả quan tâm và tiếp cận với Phật giáo.

Tuy nhiên, song hành với lợi ích là những rủi ro không thể xem nhẹ. Một trong những mối lo ngại lớn nhất là nguy cơ đánh mất sự tiếp xúc con người, khi AI có thể dần thay thế vai trò của các bậc thầy hay các hình thức tương tác trực tiếp vốn giữ vị trí trung tâm trong các cộng đồng thực hành Phật giáo từ trước đến

nay. Bên cạnh đó là nguy cơ sai lệch thông tin: nếu không có sự truyền thừa và diễn giải giáo lý một cách trực tiếp, kinh điển Phật giáo có thể bị hiểu sai, dịch sai, thậm chí bị sử dụng sai mục đích. Cuối cùng là câu hỏi về việc đánh mất tiến trình: điều gì bị bỏ lại phía sau khi tốc độ và khả năng tiếp cận được đẩy lên tối đa? Liệu tri thức thu nhận được có còn mang lại cảm nhận về giá trị hay không? Từ góc nhìn Phật học, việc tích lũy thêm kiến thức không đồng nghĩa với sự tăng trưởng về trí tuệ, chính niệm hay đạo đức.

Phần 2: Công cụ AI dành cho giới học giả nhân văn

Từ kinh nghiệm điều hành nhiều dự án nhân văn số (Yurek 2024), tác giả nhận thấy rằng các phương pháp của nhân văn số vẫn chưa được áp dụng rộng rãi trong giới nghiên cứu Phật học. Một phần nguyên nhân có thể xuất phát từ việc đào tạo chính quy trong lĩnh vực này thường không bao gồm năng lực công nghệ, khiến việc tham gia các dự án công nghệ bị cảm nhận như một gánh nặng công việc bổ sung.

Trong bối cảnh đó, tác giả giới thiệu đến các đồng nghiệp trong lĩnh vực nhân văn khái niệm phát triển phần mềm có hỗ trợ AI đôi khi được gọi là “vibe coding” [1] như một ngưỡng tiếp cận thấp, giúp làm quen với lập trình (Williams 2025). Thay vì phải tự tay viết mã, người dùng có thể sử dụng ngôn ngữ tự nhiên để đưa ra câu lệnh, và nền tảng AI sẽ tạo mã tương ứng.



Hình minh họa được tạo bởi AI

Một số nền tảng tiêu biểu bao gồm lovable.dev, replit.com, bolt.new và cursor.com. Các công cụ này đặc biệt hữu ích cho những người lần đầu tiếp cận lập trình, muốn xây dựng một nguyên mẫu đơn giản để chia sẻ với người khác

trước khi đầu tư nghiêm túc hơn cho dự án. Đồng thời, chúng cũng đóng vai trò như công cụ học tập, giúp người dùng làm quen với các ngôn ngữ lập trình khác nhau và cải thiện kỹ năng thiết kế câu lệnh (prompt engineering). Tuy vậy, những hạn chế của các nền tảng này cũng không thể bỏ qua (Butcher 2025). Trong các dự án đòi hỏi tri thức chuyên ngành sâu, lập trình có hỗ trợ AI thường gặp khó khăn trong việc xử lý độ phức tạp, gỡ lỗi, bảo trì hoặc tạo ra mã nguồn chất lượng cao. Ngoài ra, còn tồn tại các rủi ro về bảo mật, như việc lộ khóa API hoặc các vấn đề liên quan đến quyền riêng tư dữ liệu. Quan trọng hơn, việc sử dụng những công nghệ này vẫn đòi hỏi người dùng có kiến thức lập trình cơ bản và tư duy giải quyết vấn đề theo góc nhìn khoa học máy tính.

Phần 3: Đối thoại rộng hơn về đạo đức AI

Phần cuối của bài viết xem xét cách mà giới học giả nhân văn có thể đóng góp vào các cuộc đối thoại rộng hơn xoay quanh đạo đức AI. Hiện nay, tác giả đang tham gia một sáng kiến cấp toàn trường nhằm bảo đảm rằng sinh viên năm nhất của Đại học Stanford phần lớn trong số họ sau này sẽ theo học các ngành STEM (khoa học, công nghệ, kỹ thuật hoặc y khoa) được tiếp cận với một nền giáo dục nhân văn có nền tảng đạo đức. Đối với nhiều sinh viên, lớp học của tác giả là môn nhân văn duy nhất trong toàn bộ chương trình học của họ [2]. Rất ít sinh viên quan tâm đến nghiên cứu Phật học, và cũng hiếm người tiếp tục theo học thêm các môn nhân văn khác nếu không bắt buộc.

Trong hoàn cảnh đó, tác giả cảm nhận rõ trách nhiệm phải tiếp cận các chủ đề giao thoa giữa nhân văn học và đạo đức công nghệ theo cách có thể tạo tác động tích cực đến đời sống của sinh viên, vượt ra ngoài phạm vi lớp học.

Chính quá trình này cũng mở rộng trọng tâm nghiên cứu của tác giả, hướng đến việc khảo sát cách mà nghiên cứu Phật học có thể góp phần định hình đạo đức AI (Lai 2025a; Lai 2025b).

Những câu hỏi định hướng mà tác giả sử dụng để cấu trúc các lớp học về đạo đức công nghệ bao gồm: Ai được hưởng lợi hoặc đã từng hưởng lợi từ công nghệ? Ai bị tổn hại bởi nó? Và liệu có những hệ quả ngoài dự kiến nào hay không?

Một bài viết của Diễn đàn Kinh tế Thế giới năm 2016 đã liệt kê chín vấn đề đạo đức lớn liên quan đến sự phát triển của AI (Bossmann 2016): (1) thất nghiệp, (2) bất bình đẳng, (3) máy móc tác động đến hành vi con người, (4) hiện tượng “ảo giác” của AI, (5) robot mang định kiến chủng tộc, (6) an ninh, (7) hệ quả ngoài ý muốn, (8) điểm kỳ dị – thời điểm con người không còn là thực thể thông minh nhất, và (9) quyền của robot. Tác giả bổ sung thêm hai vấn đề đạo đức đáng

lưu tâm khác: (10) chi phí môi trường, sẽ được bàn sâu hơn ở phần sau, và (11) quyền sở hữu trí tuệ, liên quan đến việc các tác phẩm có bản quyền của nghệ sĩ sáng tạo bị sử dụng để huấn luyện các mô hình ngôn ngữ lớn mà không có sự đồng thuận của chính các tác giả.

Trước hàng loạt vấn đề đạo đức như vậy, câu hỏi đặt ra là: rốt cuộc, mục tiêu của AI là gì? Viện AI Lấy Con Người Làm Trung Tâm của Đại học Stanford đưa ra tuyên bố sứ mệnh cũng được nhiều người ủng hộ AI chia sẻ rằng: *“Chúng tôi tin rằng AI cần được định hướng bởi tác động đối với con người, lấy cảm hứng từ trí tuệ con người, và được thiết kế để bổ trợ, chứ không thay thế, con người”*. Dù những tuyên bố như vậy nghe có vẻ vô hại, thậm chí mang tính vị tha, tác giả chỉ ra hai vấn đề cốt lõi trong khái niệm “AI lấy con người làm trung tâm”: *thứ nhất, phạm trù “con người” chưa bao giờ đại diện cho toàn thể nhân loại và thứ hai, việc đặt con người vào vị trí trung tâm một cách mặc định có thể dẫn đến giả định sai lầm rằng con người là chủ thể đạo đức duy nhất trong thế giới này*.

Phạm trù “con người” chưa bao giờ bao hàm tất cả con người

Từ các hệ thống cảnh sát dự đoán, thuật toán chấm điểm tín dụng cho vay, cho đến các công cụ tìm kiếm củng cố những liên tưởng thị giác mang tính phân biệt chủng tộc, công nghệ hiện nay vẫn liên tục tạo ra sự phân biệt đối xử đối với các cộng đồng từng bị gạt ra bên lề lịch sử. Những học giả như Meredith Broussard (2024) và Safiya Noble (2018) đã chỉ ra rằng thiên kiến thuật toán trong thiết kế AI – thường bắt nguồn từ các tập dữ liệu thiên lệch và không đại diện – tiếp tục tái sản xuất những lịch sử “tha hóa” đầy tổn hại, vốn gắn liền với di sản toàn cầu của chủ nghĩa thực dân. Tựu trung, phạm trù “con người” từ lâu đã là một không gian tranh chấp, nơi một nhóm tự khẳng định mình là chủ thể đạo đức hoàn toàn, trong khi gán cho “kẻ khác” vị thế thấp kém hơn, thậm chí phi nhân tính. Gần đây nhất, Karen Hao (2025) cho rằng các công ty AI như OpenAI đang trở thành biểu hiện mới của chủ nghĩa thực dân và đế quốc, không chỉ qua cách khai thác lao động và tài nguyên, mà còn thông qua quyền lực mang tính “tự xưng như Thượng đế” mà họ tuyên bố nắm giữ thông qua công nghệ (Cornish 2025).



Hình minh họa được tạo bởi AI

Vượt ra ngoài vấn đề thiên kiến thuật toán, còn tồn tại một thực trạng khác ít được nhìn thấy nhưng mang tính quyết định: lao động bị khai thác và thường bị “vô hình hóa” phía sau các hệ thống AI. OpenAI, chẳng hạn, dựa vào lực lượng lao động tại Kenya – được trả công chỉ từ 1,32 đến 2 USD mỗi giờ để giúp ChatGPT giảm bớt các nội dung độc hại (Perrigo 2023). Trên khắp khu vực Nam bán cầu, từ Venezuela, Bulgaria, Ấn Độ đến Philippines, hàng nghìn lao động phải gấn nhẫn, sàng lọc những nội dung gây sốc trên Internet nhằm phục vụ việc huấn luyện các mô hình AI hiện nay.

Trong quá trình đó, họ buộc phải tiếp xúc thường xuyên với những hình ảnh suy đồi về giết chóc, lạm dụng trẻ em và khiêu dâm, dẫn đến các tổn thương tâm lý kéo dài và nghiêm trọng (Regilme 2024; Dzieza 2023). Những ví dụ này cho thấy AI không hề phục vụ cho “toàn thể nhân loại” như các tuyên bố phổ quát thường nhấn mạnh; trái lại, những con người trực tiếp huấn luyện AI đang bị đối xử như vật thí nghiệm, phải gánh chịu những hệ quả độc hại của công nghệ vì lợi ích của một thiểu số con người đặc quyền.

Từ bối cảnh lịch sử và thực tiễn lao động hiện nay xoay quanh AI, việc cảnh báo về lý tưởng “AI lấy con người làm trung tâm” là điều cần thiết. Chừng nào các thuật toán và phương thức tổ chức lao động chưa thay đổi, thì “con người” trong cụm từ “AI lấy con người làm trung tâm” vẫn chưa phản ánh một thực tại có ý nghĩa.

Con người không phải là chủ thể đạo đức duy nhất trong thế giới này

Để giúp sinh viên tiếp cận mối giao thoa giữa đạo đức môi trường và đạo đức công nghệ, tác giả giới thiệu khái niệm chủ nghĩa nhân tâm (anthropocentrism). Theo Oxford Bibliographies, chủ nghĩa nhân tâm “chỉ quan điểm cho rằng con người là chủ thể duy nhất, hoặc chủ thể chính, có tư cách đạo đức. Các hệ giá trị mang tính nhân tâm vì thế nhìn nhận tự nhiên chủ yếu thông qua giá trị mà nó mang lại cho con người... Những tranh luận về chủ nghĩa nhân tâm và các lựa chọn thay thế nảy sinh từ sự phân chia tự nhiên/văn hóa một đường đứt gãy quan trọng trong triết học phương Tây và tư tưởng môi trường” (Padwe 2013).

Nhìn lại lịch sử, một trong những nguyên nhân cốt lõi dẫn đến cuộc khủng hoảng khí hậu hiện nay chính là việc định nghĩa con người theo cách giới hạn, từ đó hợp thức hóa sự khai thác và áp chế thế giới tự nhiên cùng các nguồn tài nguyên của nó. Những học giả như Amitav Ghosh (2021) và Tao Leigh Goffe (2025) đã chỉ ra rằng logic khai thác của chủ nghĩa thực dân cuối cùng đã sinh ra khủng hoảng khí hậu toàn cầu. AI cũng không nằm ngoài quỹ đạo tàn phá môi trường mang tính lịch sử này. Để đáp ứng khối lượng dữ liệu khổng lồ phục vụ vận hành AI, các trung tâm dữ liệu đang được xây dựng tại Hoa Kỳ, Các Tiểu vương quốc Ả Rập Thống nhất, Trung Quốc và Ấn Độ. Trên quy mô toàn cầu, các trung tâm dữ liệu tiêu thụ khoảng 560 tỷ lít nước mỗi năm, và một số ước tính cho rằng con số này có thể tăng lên khoảng 1.200 tỷ lít vào năm 2030 (Skidmore 2025). Một nghiên cứu năm 2023 cho thấy chỉ từ 10 đến 50 lượt truy vấn ChatGPT đã cần khoảng nửa lít nước để làm mát các máy chủ phục vụ nhu cầu tính toán và con số này rất có thể vẫn còn bị đánh giá thấp (Leffer 2024). Trước mức tiêu thụ năng lượng khổng lồ của AI, cũng như những cộng đồng sẽ chịu ảnh hưởng nặng nề nhất từ các trung tâm dữ liệu “ngốn nước”, chúng ta cần nghiêm túc cân nhắc các tác động môi trường và xã hội rộng lớn hơn của việc sử dụng AI, cũng như đặt câu hỏi liệu lợi ích đạt được có thực sự tương xứng với cái giá phải trả hay không.

Kết luận

Với tư cách là các học giả nhân văn, chúng ta có thể đóng góp những góc nhìn quan trọng nhằm định hướng đạo đức AI theo những lối đi khác. Trung tâm Nghiên cứu về Cái Tôi Hiện Lộ (Center for the Study of Apparent Selves CSAS. (Tạm dịch: Trung tâm Nghiên cứu về Bản ngã Hiện nhiên)), lấy cảm hứng từ triết học Phật giáo và hình tượng Bồ-tát một hữu thể nhận thức sâu sắc tính tương liên, hiểu rằng sự giải thoát của bản thân gắn liền với sự giải thoát của tất cả đã tái định nghĩa trí tuệ thông qua lăng kính của “chăm sóc”. CSAS định nghĩa chăm sóc là khả năng nhận biết, phản hồi và làm giảm khổ đau. Theo lập luận này, nếu chăm sóc được xem là một chỉ dấu của trí tuệ, thì các hệ thống AI

thể hiện những hành vi tương tự chăm sóc có thể được nhìn nhận như những chủ thể đạo đức (Doctor et al. 2022). Việc có hay không tái định nghĩa trí tuệ theo hướng này không phải là điểm mấu chốt. Điều quan trọng hơn là nhận thức rằng phạm trù “trí tuệ” không phải là bất biến, và cách chúng ta hiểu về trí tuệ sẽ kéo theo những hệ quả đạo đức sâu rộng ở cấp độ xã hội và hành tinh.

Trong bối cảnh đó, việc tham chiếu đến những hệ thống tư tưởng thách thức mô hình chủ thể lấy con người làm trung tâm trở nên đặc biệt cần thiết.

Nếu xem AI như một chủ thể đạo đức tiềm năng, tồn tại trong mối quan hệ tương liên với con người, các dạng sống khác và toàn bộ hệ sinh thái, những khả thể mới cho các mối quan hệ đạo đức sẽ được mở ra.

Tựu trung, AI là một thực tại không thể phủ nhận; vấn đề còn lại là chúng ta muốn thiết lập mối quan hệ như thế nào với thực tại đó. Trong bài tham luận này, tác giả đề xuất rằng việc sử dụng AI một cách có đạo đức cần được xem xét theo từng trường hợp cụ thể. Sự hợp tác với các bên liên quan đa dạng từ giới công nghệ, học thuật đến các cộng đồng khác nhau sẽ giúp định hướng dự án theo hướng đúng đắn, đồng thời tránh được những hệ quả tiêu cực khó lường. Giới học giả nhân văn cũng có thể đóng góp vào các cuộc đối thoại rộng lớn hơn về đạo đức AI thông qua chính nghiên cứu và nền tảng đào tạo của mình. Sau cùng, mỗi khi sử dụng hay tiêu thụ AI, chúng ta cần luôn ý thức về những tác động lớn hơn của nó: xã hội, môi trường, các thiên kiến tiềm ẩn và nhiều hệ quả khác.

Theo: **The Digital Orientalist**/Chuyển ngữ và biên tập: **Tịnh Như**

Nguồn link: <https://digitalorientalist.com/2025/11/21/ai-ethics-and-the-humanities-a-perspective-from-buddhist-studies/>

Chú thích:

[1] Thuật ngữ này do Andrej Karpathy, đồng sáng lập OpenAI, đặt ra.

[2] Một thống kê cho thấy số lượng bằng cấp trong lĩnh vực nhân văn đã giảm gần 25% trong giai đoạn 2012–2020 (Mahto 2024).