TẠP CHÍ NGHIÊN CỨU
PHẬT HỌC
CƠ QUAN NGÔN LUẬN CỦA TRUNG ƯƠNG GIÁO HỘI PHẬT GIÁO VIỆT NAM
ISSN: 2734-9195

Vô ngã trước thuật toán: Phật học đối thoại cùng trí tuệ nhân tạo (P.3 - Hết)

ISSN: 2734-9195 08:05 02/11/2026

Một hệ thống "AI trì giới" thể hiện sự tự chế về mặt đạo đức và bảo vệ những ranh giới luân lý trong tiến trình này. Một hệ thống "AI từ bi" góp phần nâng cao an lạc của con người.

Trong bối cảnh trí tuệ nhân tạo (AI) ngày càng tham gia sâu vào đời sống con người, việc xây dựng những hệ thống công nghệ có trách nhiệm không chỉ là bài toán kỹ thuật, mà còn đặt ra yêu cầu về nền tảng đạo đức định hướng cho quá trình phát triển và vận hành AI.

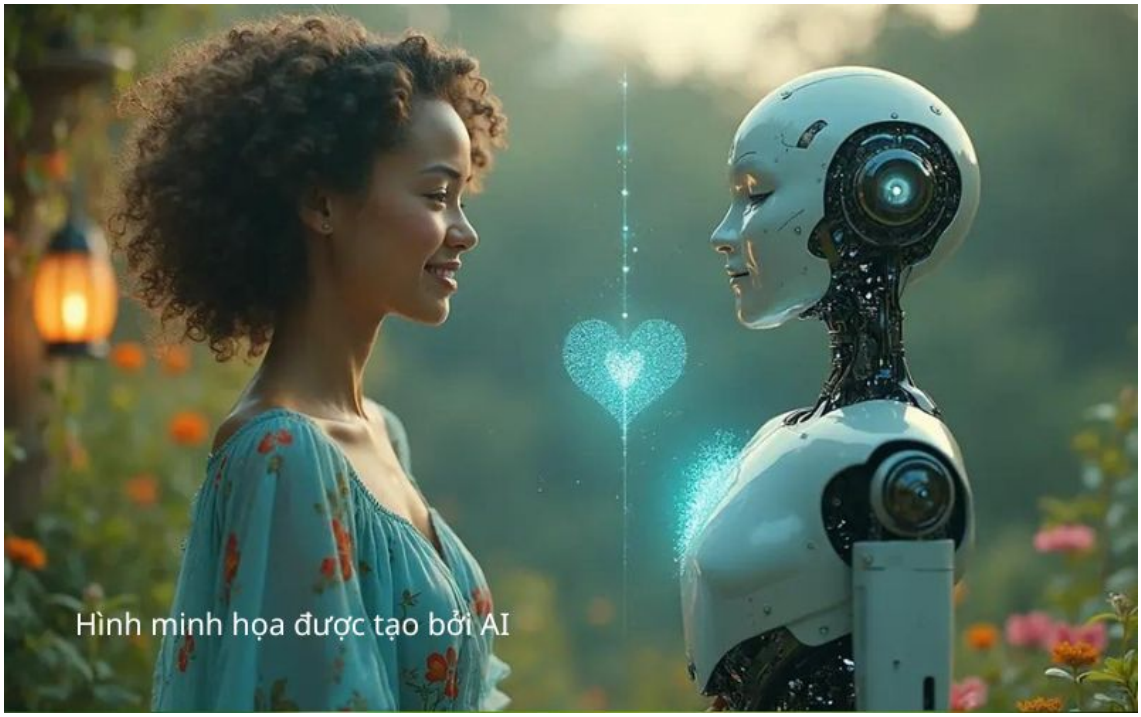
Từ quan niệm về tâm tính trong Phật học, bài viết gợi mở khả năng chuyển hóa các nguyên tắc giới luật, trí tuệ và tâm bình đẳng thành những định hướng đạo đức cho thiết kế AI. Tuy nhiên, vấn đề cốt yếu không phải là gán cho thuật toán một tâm thức hay khả năng giác ngộ vốn chưa được chứng minh, mà là xây dựng những cơ chế công nghệ có khả năng hạn chế tổn hại, thúc đẩy lợi ích chung và hỗ trợ con người phát triển những phẩm chất tốt đẹp.

4.2. Khơi nguồn cảm hứng cho công nghệ AI từ lý thuyết tâm tính trong Phật học

Cũng như nhà trường giáo dục học sinh không chỉ bằng việc bồi dưỡng kiến thức, kỹ năng mà còn thông qua việc hình thành các giá trị, lĩnh vực AI cũng cần được định hướng theo tinh thần ấy. Bằng cách đưa những quan niệm về tâm tính trong Phật học, chẳng hạn như các nguyên tắc kiềm chế hành vi bất thiện, quy tắc tu tập, đạo đức và tâm bình đẳng, vào quá trình phát triển AI, chúng ta có thể hướng công nghệ này trở thành một công cụ được định hướng bởi ý thức đạo đức và lòng từ bi. AI không chỉ có thể hỗ trợ con người giữ gìn điều thiện, loại bỏ điều ác, góp phần củng cố tinh thần và hoàn thiện nhân cách, mà còn có thể hỗ trợ nhân loại đạt tới sự "tự thực hiện" đích thực (*self-realization*).

Định hướng đạo đức của các nhà thiết kế không trực tiếp làm thay đổi những tham số nền tảng bên trong mô hình. Thay vào đó, định hướng ấy tác động đến

“các nguyên tắc chi phối việc điều chỉnh hàm thưởng, các bộ lọc an toàn và quá trình đánh giá đầu ra (Huang và cộng sự, 2025)”, qua đó ảnh hưởng đến hành vi của AI.



Trước hết, khi tìm kiếm nguồn cảm hứng cho công nghệ AI, chúng ta có thể học hỏi từ lời dạy của Lục tổ Huệ Năng về “Pháp môn Tâm địa” [41*]: “Tâm địa không ác, tự tính giới; Tâm địa không si, tự tính tuệ; Tâm địa không loạn, tự tính định”.* [42]

Mục tiêu ở đây không phải là trao cho AI khả năng giác ngộ nội tại, mà là chuyển hóa những định hướng giá trị ấy thành một khuôn khổ thiết kế nhằm định hình các khuynh hướng hành vi của hệ thống. Trong quá trình xây dựng tập dữ liệu huấn luyện, các văn bản thể hiện lòng từ bi, trí tuệ và tinh thần Trung đạo, chẳng hạn như những kinh điển Phật giáo được tuyển chọn, các trước tác triết học và tài liệu đương đại có nền tảng đạo đức vững chắc, có thể được đưa vào nhằm thiết lập cơ sở dữ liệu lành mạnh. Ở các giai đoạn tinh chỉnh mô hình và học tăng cường, những hàm thưởng được thiết kế để củng cố các kiểu phản hồi phù hợp với những chuẩn mực đạo đức, đồng thời hạn chế các đầu ra có hại. Cách tiếp cận này có thể được thực hiện thông qua những phương pháp hiện có như tinh chỉnh theo chỉ dẫn (*instruction tuning*) và mô hình hóa phần thưởng (*reward modeling*) (Alammar và Grootendorst, 2024, tr. 379).

Máy tính vận hành theo thuật toán chứ không dựa trên những niềm tin như con người. *“Để học hỏi, các thuật toán sơ khởi cần thêm một yếu tố bên cạnh quyền tiếp cận dữ liệu. Chúng cũng cần một mục tiêu. [...] Các thuật toán cũng phải*

được trao một mục tiêu thì mới có thể học hỏi” (Harari, 2024, tr. 301). Vì vậy, chúng không chỉ cần dữ liệu mà còn cần những mục tiêu rõ ràng. Các nghiên cứu hiện có cho thấy, bằng cách điều chỉnh những hành vi cụ thể thông qua hàm thưởng, người ta có thể định hướng hoạt động của các tác tử thông minh theo những mục tiêu có ý nghĩa mà không làm suy giảm hiệu quả tổng thể của chúng (Keane và cộng sự, 2025). Điều này cho thấy những phương pháp như tinh chỉnh theo chỉ dẫn và mô hình hóa phần thưởng là một hướng triển khai khả thi.

Ở cấp độ quản trị đầu ra, một phương án thực tiễn là thiết lập cơ chế rà soát và lọc nội dung có khả năng điều chỉnh linh hoạt, dựa trên những nguyên tắc đạo đức được xác định rõ ràng, chẳng hạn như lời dạy của Phật giáo: *“Không làm các điều ác, vâng làm các điều lành”*. Theo đó, những đầu ra không đáp ứng các tiêu chí này có thể được điều chỉnh hoặc chuyển sang quy trình xem xét của con người.

Cách tiếp cận như vậy không chỉ đòi hỏi việc triển khai các giải pháp kỹ thuật, mà còn yêu cầu sự trưởng thành về đạo đức của tất cả những người tham gia phát triển AI, bao gồm các nhà quản lý và người ra quyết định, đội ngũ xây dựng kiến trúc kỹ thuật cũng như những người hoạch định chính sách đạo đức. Mục tiêu không phải là đưa vào hộp đen thuật toán một “tâm thức” giả định, mà là xây dựng một kiến trúc hệ thống được định hướng bởi các giá trị. Kiến trúc này nhất quán tích hợp những nguyên tắc làm lợi và không gây tổn hại vào quá trình tuyển chọn dữ liệu, cấu trúc phần thưởng và phương thức tương tác. Qua đó, các hệ thống AI có thể được định hướng theo những thiết kế thực sự hỗ trợ sự phát triển tốt đẹp của con người, tôn trọng mọi sinh thể và tránh gây tổn hại cho bất kỳ chúng sinh nào.

Như các tác giả cuốn *Moral Machines* (*Những cỗ máy đạo đức*) đã nhận định: *“Cần khẩn trương suy nghĩ về việc thiết kế các tác nhân máy móc tự chủ (AMA), bởi những hệ thống tự chủ đã bước vào không gian đạo đức của các hoạt động thường nhật”* (Wallach và Allen, 2008, tr. 17). Để tầm nhìn này trở thành hiện thực, trí tuệ phải giữ vai trò dẫn đường. Chỉ khi ấy, AI mới có thể góp phần khơi dậy trí tuệ của con người thay vì giải phóng những xung lực mang tính hủy hoại. Theo nghĩa đó, việc thiết kế AI không chỉ có thể hỗ trợ hoàn thiện nhân cách con người mà còn góp phần vào tiến trình tu dưỡng Phật tính.

Dẫu vậy, trong các hệ thống mô hình ngôn ngữ lớn (LLM) có mức độ tự động hóa cao hiện nay, những quy tắc được áp đặt từ bên ngoài có thể hạn chế tính linh hoạt và sức sáng tạo của AI. Việc hiện thực hóa các nguyên tắc đạo đức trong những mô hình LLM dạng “hộp đen” hiện nay quả thực là một nhiệm vụ

phức tạp và đầy thách thức. Tuy nhiên, điều đó không có nghĩa là việc tích hợp đạo đức đã đi vào ngõ cụt; trái lại, nó đòi hỏi sự hợp tác liên ngành để tìm kiếm những phương thức căn chỉnh phù hợp sâu sắc hơn với bản chất vận hành của quá trình học máy.

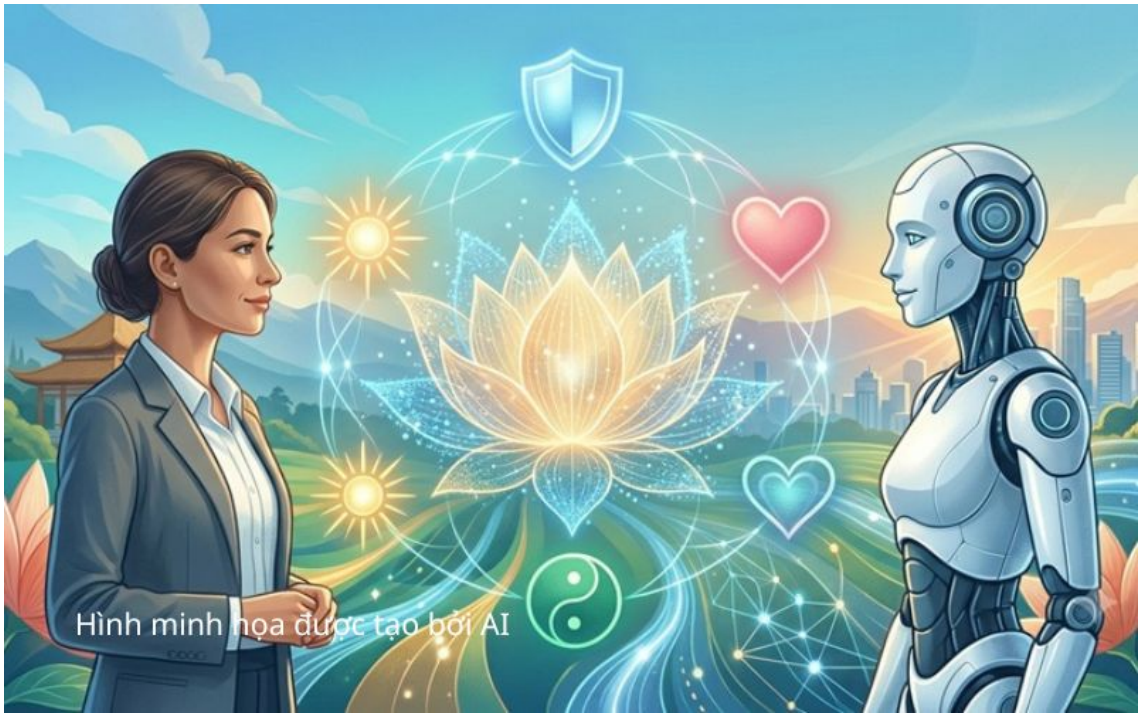
4.3. Thanh lọc "tâm" AI thông qua Ngũ giới và Thập thiện

Khái niệm "AI thân thiện" (*Friendly AI*) được Eliezer Yudkowsky, một trong những người sáng lập Viện Nghiên cứu Trí tuệ Nhân tạo Singularity, đề xuất và phát triển lần đầu vào năm 2001 (Yudkowsky, 2001). Các học giả đã tiếp cận vấn đề xây dựng hệ thống đạo đức cho AI từ nhiều góc độ khác nhau, bao gồm *"Quy tắc Vàng, Mười Điều răn, các nguyên tắc Yama và Niyama của Ấn Độ giáo, những danh mục đức hạnh, mệnh lệnh tuyệt đối của Kant [...] và Ba định luật Robot của Asimov (Asimov, 1942)"* (Wallach và Allen, 2008, tr. 79). [43]

Tuy nhiên, trong các nghiên cứu so sánh giữa Phật giáo phương Đông với đạo đức học và lý thuyết đức hạnh phương Tây, vẫn chưa có sự đồng thuận về việc điều gì được xem là "thân thiện" và điều gì cấu thành sự "thanh lọc". "Việc lập trình các khuôn khổ đạo đức cho AI có thể làm nảy sinh những lo ngại về việc cách diễn giải Pháp hay đạo đức đức hạnh của ai sẽ được mã hóa, từ đó có nguy cơ củng cố những thiên kiến cụ thể" (Maheshwari và cộng sự, 2025, tr. 54).

Vì vậy, câu hỏi làm thế nào để vận dụng những tư tưởng này vào công nghệ AI, đồng thời tránh những thiên kiến phát sinh từ quá trình học tăng cường, vẫn chưa có lời giải dứt khoát. Gần đây, một số học giả đã đề xuất sử dụng Thập thiện để định hướng việc ứng dụng AI có đạo đức, đánh dấu bước thử nghiệm ban đầu trong việc kết hợp đạo đức Phật giáo với trí tuệ nhân tạo.

Nghiên cứu của Yomsin (2025) "khám phá việc vận dụng Thập thiện nghiệp đạo (*kusala-kammapatha*) và khái niệm như lý tác ý (*yoniso-manasikāra*) như những khuôn khổ đạo đức nhằm giải quyết tình trạng lạm dụng và các mối quan ngại về đạo đức liên quan đến nội dung giả mạo do AI tạo ra (*AI Deepfake*), qua đó đề xuất những phương thức sử dụng công nghệ một cách có trách nhiệm và đúng đắn nhằm hỗ trợ, thúc đẩy một xã hội hòa bình".



Công trình của Hershock (2021) làm sáng tỏ khả năng của các nguồn lực Phật giáo trong việc định hướng lại sự chú tâm, sự hiện diện và tinh thần trách nhiệm của con người trong những môi trường được trung gian bởi AI. Cách tiếp cận này bổ sung một phương diện quan trọng cho các khuôn khổ dựa trên giới luật và nguyên tắc giảm thiểu tổn hại.

Tiếp nối những phân tích trên, phần này sẽ thử tiếp cận vấn đề từ góc nhìn Phật giáo, chuyển hóa các nguyên tắc “Ngũ giới và Thập thiện” thành những cơ chế bảo vệ đạo đức cho AI, qua đó thiết lập một nền tảng đạo đức có cấu trúc.

Lấy Ngũ giới [44] và Thập thiện nghiệp đạo [45] trong lý thuyết tâm tính của Phật giáo làm điểm xuất phát, có thể thấy những thực hành này biểu hiện một cách tự nhiên của tâm thanh tịnh thông qua hành vi đạo đức. Dựa trên nguyên lý đó, các nhà phát triển có thể chuyển hóa Ngũ giới và Thập thiện thành một khuôn khổ đạo đức định hướng hoạt động của thuật toán, đồng thời tích hợp khuôn khổ này vào thiết kế tương tác giữa người và máy. Theo cách ấy, họ có thể tạo lập điều được gọi là một “tâm thanh tịnh của AI”.

*Trước hết là nguyên tắc **không sát sinh**.*

Phật giáo cho rằng quá trình ra quyết định cần đặt lòng từ bi ở vị trí ưu tiên. Điều này không chỉ dừng lại ở việc kiềm chế hành vi giết hại bên ngoài, mà còn bao hàm một tâm không có ý định gây tổn hại,biểu hiện tự nhiên của lòng từ bi nơi tự tính. Trong lĩnh vực đạo đức AI, nguyên tắc này có nghĩa là thiết kế các cơ chế bảo vệ nhằm ứng phó với những hệ quả ngoài dự kiến, qua đó ngăn ngừa tổn hại có thể xảy ra. Theo cách hiểu này, tinh thần không sát sinh tương ứng

với nguyên tắc thiết kế công nghệ theo hướng tránh gây hại.

Năm 2023, giám đốc điều hành của nhiều công ty AI lớn đã cùng ký một bức thư ngỏ chung chỉ gồm một câu: *"Giảm thiểu nguy cơ AI gây ra sự tuyệt chủng cần được xem là ưu tiên toàn cầu, ngang hàng với những rủi ro quy mô xã hội khác như đại dịch và chiến tranh hạt nhân"* (Center for AI Safety, 2023). Tuyên bố này cho thấy xã hội hiện nay nhìn nhận nghiêm túc đến mức nào khả năng phát sinh những tổn hại ngoài dự kiến từ các quyết định do AI đưa ra.

Các thuật toán được xây dựng để đạt được những mục tiêu do con người giao cho chúng. Nếu các mục tiêu ấy chỉ giới hạn ở việc "tối đa hóa lợi ích", kết quả thường sẽ trở nên phiến diện và thiển cận. Điều quan trọng không chỉ là hiệu suất, mà còn là tác động của hệ thống đối với mọi sinh thể chịu ảnh hưởng từ hoạt động của nó. Vì vậy, các nhà thiết kế cần xác lập những mục tiêu coi con người và máy móc là các thành phần trong cùng một mạng lưới sinh thái. Chỉ khi đó, AI mới có thể đóng góp vào phúc lợi chung mà không gây tổn hại.

Tư tưởng Phật giáo nhắc nhở chúng ta rằng "vạn pháp từ một tâm mà sinh". Trong đời sống con người, mọi sự đều có mối liên hệ với nhau; nguyên lý tương tự cũng có ý nghĩa trong môi trường số. Vì vậy, các thuật toán cần lần theo những chuỗi nhân quả dài hạn bắt nguồn từ các lựa chọn của mình. Nếu thiếu tầm nhìn này, những tính toán về lợi nhuận và lợi ích trước mắt rất dễ dẫn đến những tổn thất về lâu dài.

Quan niệm ấy không phải là điều hoàn toàn mới. "Hiệu ứng cánh bướm" nổi tiếng của Edward Lorenz, được dẫn lại trong *Embodied Intelligence* (*Trí tuệ hiện thân*), cho thấy ngay cả những quyết định nhỏ nhất ở hiện tại cũng có thể định hình các kết quả ở tương lai xa theo những cách khó lường. Vì vậy, AI cần kết hợp dự báo ngắn hạn với các dự phóng dài hạn (Liu, 2025, tr. 180-181). Hệ thống không chỉ cần xem xét những kết quả tức thời, mà còn phải mô hình hóa các hệ quả lâu dài thông qua phân tích nhân quả và suy luận logic, nhằm dự liệu cách những hành động của nó có thể tạo ra các tác động lan tỏa trong toàn xã hội.

Để những nguyên tắc như "không gây hại" có thể phát huy hiệu quả trong các hệ thống AI, chúng cần được củng cố bằng những cơ chế cụ thể. Trong học tăng cường, điều này đòi hỏi phải kết hợp cả các ràng buộc lẫn những khuyến khích. Các nhà thiết kế có thể triển khai theo hai hướng bổ trợ cho nhau.

Hướng thứ nhất tập trung vào các ràng buộc mang tính ngăn chặn, chẳng hạn đưa "chi phí gây hại" (*harm cost*) vào hàm mục tiêu, nhằm giảm mạnh giá trị phần thưởng đối với những hành động có khả năng dẫn đến tổn hại về thể chất

hoặc tâm lý.

Hướng thứ hai nhấn mạnh các khuyến khích tích cực, thông qua việc gia tăng giá trị của hàm lợi ích bằng cách tưởng thưởng cho những hành động cứu người và làm vơi bớt khổ đau.

Các nghiên cứu hiện có cho thấy: “Những cách tiếp cận kết nối luận gần đây có xu hướng ít phụ thuộc hơn vào các cấu trúc được định sẵn, thay vào đó tận dụng khả năng của mạng nơ-ron nhân tạo để tự hình thành các hệ thống phân loại một cách linh hoạt từ dữ liệu đầu vào mà chúng tiếp nhận” (Wallach và Allen, 2008, tr. 107). Khi một nút trong mạng được củng cố đủ số lần, việc áp dụng một quy tắc nhất định có thể trở thành phản ứng tự động. Thông qua quá trình học tăng cường như vậy, việc tuân thủ các quy tắc đạo đức cũng có thể được tự động hóa, qua đó đưa tinh thần “tâm không sát hại” trực tiếp vào “tâm” của AI.

Trong quản trị dữ liệu AI, các nhà phát triển có thể tham khảo những tác phẩm Phật giáo như *Tự môn sùng hành lục* (*Zimen Chongxing Lu*), *Liễu Phàm tứ huấn* (*Liao Fan Si Xun*), *Phổ Hiền hạnh nguyện phẩm* (*Puxian Xingyuan Pin*), *Cách ngôn liên bích* (*Geyan Lianbi*), *Tây phương công cứ* (*Xifang Gongju*), *An Sĩ toàn thư* (*Anshi Quanshu*), *Tịnh độ tập yếu* (*Jingtu Ji Yao*) và *Nhân sinh bách sự* (*Rensheng Baishi*) [46].

Những tác phẩm này có thể cung cấp những góc nhìn đặc thù, góp phần đưa tinh thần từ bi và không gây tổn hại vào các hàm mục tiêu của AI. Thông qua việc cung cấp một lượng đáng kể các mẫu dữ liệu tích cực phục vụ quá trình căn chỉnh đạo đức, các nhà thiết kế có thể chuyển hóa những quan niệm sâu sắc về sự tôn trọng sự sống và tôn trọng quy luật nhân quả thành những quy tắc có thể triển khai trong thực tế. Qua đó, các hệ thống AI có thể được thiết kế để tự điều chỉnh, tự đánh giá và tự sửa lỗi trong quá trình vận hành.

Những quy tắc do AI thực thi không thể khiến con người “thành Phật”, nhưng có thể tạo ra những “điều kiện bên ngoài” thuận lợi cho quá trình tự thực hiện của con người, góp phần kiến tạo một môi trường thanh tịnh. Nói cách khác, thông qua những “giới luật số” (*digital precepts*), AI có thể hỗ trợ người dùng phòng hộ các căn, duy trì chính niệm và thanh lọc tâm, thay vì trở thành một “cảnh giới ma” khuếch đại dục vọng và làm nhiễu loạn tâm tính.

Những vấn đề mà Facebook từng gặp phải, bao gồm “kích động bạo lực, tin giả, ngôn từ thù ghét, lôi kéo tình dục, vi phạm quyền sở hữu trí tuệ và hàng chục vấn đề khác” (Shneiderman, 2022, tr. 235), có thể được hạn chế thông qua những định hướng đạo đức như vậy.

Cơ chế quản trị thuật toán theo nguyên tắc “không gây hại” góp phần lan tỏa tinh thần bất hại và từ bi. Khi người dùng tiếp xúc với các nền tảng mạng xã hội được định hình bởi những thuật toán như vậy, họ sẽ được tiếp cận với nội dung đã được sàng lọc nhằm hạn chế sự thù ghét, bạo lực và dục vọng thái quá, đồng thời có thêm cơ hội tiếp nhận những thông điệp thúc đẩy sự hòa hợp.

Môi trường này nuôi dưỡng những hạt giống thiện lành, đồng thời giảm bớt các điều kiện khơi dậy sân hận và hành vi hung hăng. Nhờ đó, con người có thể duy trì chính niệm tốt hơn, và những phẩm chất vốn thanh tịnh của tâm có thêm điều kiện để biểu hiện.

Thứ hai, ***không trộm cắp****.*

Xét theo nghĩa trực tiếp, giới này đề cao sự tôn trọng tài sản và quyền tự chủ của người khác. Theo cách diễn giải từ giáo lý Tâm địa của Phật giáo, giới không trộm cắp biểu hiện trạng thái “tâm không có niệm tham lam”.

Nguyên tắc này có ý nghĩa quan trọng đối với việc thiết kế AI, bởi nó hướng đến xây dựng những không gian số dựa trên sự bình đẳng, công bằng, tôn trọng và không có hành vi trục lợi.

Để triển khai trong thực tế, các nhà phát triển cần điều chỉnh cả hàm mục tiêu lẫn kiến trúc thuật toán. Chẳng hạn, dữ liệu của người dùng, với tư cách là đối tượng gắn với quyền và lợi ích hợp pháp của họ, không nên bị thu thập hoặc khai thác tùy tiện khi chưa có sự đồng thuận tự nguyện và sự hiểu biết đầy đủ của họ. Đồng thời, cần tôn trọng quyền sở hữu trí tuệ của người sáng tạo, xây dựng một hệ thống công bằng và loại bỏ các hành vi chiếm đoạt.

Thứ ba, ***không tà dâm****.*

Trong Phật pháp, giới này được diễn giải ở bình diện Tâm địa là “trong tâm không có niệm dâm dục”. Ở đây, dục vọng chỉ sự ham muốn hoặc chấp trước quá mức; trong bối cảnh hiện đại, điều này còn có thể được liên hệ với sự lệ thuộc của con người vào những hình thức kích thích từ môi trường ảo.

Việc chuyển hóa nguyên tắc này vào thiết kế AI đặt ra hai trách nhiệm chính. Thứ nhất, thuật toán cần hạn chế việc đề xuất những nội dung có tính kích thích quá mức. Hệ thống không nên liên tục dẫn dắt người dùng đến những tài liệu khơi gợi ham muốn hoặc nuôi dưỡng những tưởng tượng ám ảnh. Thay vào đó, AI có thể định hướng người dùng đến những nội dung hỗ trợ sức khỏe thể chất, tinh thần và sự phát triển nhân cách.

Thứ hai, hệ thống cần có cơ chế can thiệp khi người dùng đắm chìm quá lâu trong những nội dung kích thích dục vọng. Trong những trường hợp như vậy, AI có thể đưa ra lời nhắc để người dùng tạm dừng, thoát khỏi dòng nội dung và trở về với chính niệm.

Đối với cơ chế cảnh báo sớm của AI, cần nhận diện kịp thời và có biện pháp can thiệp phù hợp trước những ngôn ngữ mang tính thao túng, qua đó góp phần tạo lập một môi trường giao tiếp an toàn và lành mạnh. Bằng cách này, AI có thể hỗ trợ người dùng phát triển khả năng tự chủ và tiết chế hành vi.

Thứ tư, ***không vọng ngữ, không nói lời chia rẽ, không nói lời thô á****c.*

Xét theo nghĩa trực tiếp, giới này ngăn ngừa lời nói sai sự thật, lời nói gây chia rẽ, lời nói thô ác và những phát ngôn mang tính lừa dối; đồng thời khuyến khích lời nói chân thật, có ích và không gây tổn hại. Theo tinh thần căn bản của Phật pháp, ở bình diện Tâm địa, điều này được diễn giải là “trong tâm không có niệm dối trá”.



Đối với AI, nhiệm vụ không chỉ là “tránh nói dối”. Điều quan trọng hơn là xây dựng những hệ thống chủ động bảo vệ tính chân thực của thông tin và thúc đẩy ngôn ngữ lành mạnh, không để bản thân trở thành phương tiện truyền bá sự lừa dối hoặc những lời nói gây hại.

Trong thực tiễn, có hai phương diện cần được chú trọng: truy xuất nguồn gốc nội dung trực tuyến và chuyển hóa những cách diễn đạt có hại.

Với sự phát triển nhanh chóng của các hệ thống tạo sinh có khả năng mô phỏng gần như hoàn chỉnh những sản phẩm do con người tạo ra, việc xác minh nguồn thông tin và loại bỏ các hiện tượng "ảo giác" (*hallucination*) hoặc hành vi cố ý lừa dối đã trở thành một cách tôn trọng nhu cầu phân biệt thật - giả của người dùng.

Một phương diện khác là can thiệp đối với ngôn từ gây hại. Khi AI phát hiện những lời lăng mạ hoặc ngôn ngữ kích động, hệ thống có thể hướng dẫn người dùng diễn đạt theo những cách khác phù hợp hơn. Bằng cách đề xuất ngôn ngữ bất bạo động, AI có thể khuyến khích người dùng lựa chọn những lời nói có khả năng hàn gắn thay vì gây tổn thương.

Thông qua những lời nhắc về "lời nói chân thật, lời nói có ích và lời nói từ ái", AI có thể từng bước góp phần định hình lại thói quen ngôn ngữ, dẫn dắt người dùng từ sự thanh lọc lời nói đến sự thanh lọc tư tưởng.

*Thứ năm, **chuyển hóa ba độc**: tham, sân và si.*

Trong Phật giáo, sự ham muốn, chiếm hữu và không biết đủ được xem là những biểu hiện của tham.

Để giúp con người giảm thiểu tâm tham, thiết kế AI cần hướng đến việc khắc phục mô hình thuật toán phục vụ sự khuếch đại vô hạn các ham muốn của con người, kích thích tiêu dùng và thúc đẩy hành vi lệ thuộc. Một giải pháp là chuyển đổi các chỉ số cốt lõi từ việc "tối đa hóa mức độ tương tác và tiêu dùng của người dùng" sang việc khuyến khích sự biết đủ, chẳng hạn thông qua những lời nhắc về sự tri túc. Qua đó, người dùng có thể thực hành tiết chế trong hành vi duyệt và tiếp nhận nội dung.

Khi đối diện với những quan điểm trái ngược với mình, con người thường phản ứng bằng sự tức giận và thù địch; đây là những biểu hiện của sân.

Để góp phần giảm thiểu sân hận, AI cần có khả năng nhận diện các dấu hiệu xung đột và tức giận, chủ động hỗ trợ giải quyết những bất đồng, ngăn chặn sự lan truyền của ác ý và thúc đẩy sự thấu hiểu, hòa bình.

Bằng cách nhìn nhận sân hận như một biểu hiện có thể bắt nguồn từ những nhu cầu sâu xa chưa được đáp ứng, đồng thời hướng dẫn người dùng thực hành chính niệm để không bị cuốn theo cảm xúc, AI có thể khuyến khích họ trình bày sự việc một cách khách quan hơn, chuyển lời buộc tội thành đối thoại, biến đối đầu thành giao tiếp.

Nguyên nhân căn bản nhất của cả tham và sân là si. Để góp phần giảm thiểu si mê, AI cần hỗ trợ người dùng điều chỉnh những thiên kiến nhận thức, cung cấp các góc nhìn đa dạng và trình bày nhiều lập trường khác nhau. Bằng cách ấy, AI có thể trở thành cầu nối hỗ trợ việc truyền đạt chính kiến, giúp người dùng từng bước vượt ra khỏi vòng luẩn quẩn của vô minh.

Thông qua việc thiết lập những ràng buộc đạo đức và định hướng các phẩm chất của con đường Phật đạo cho AI, con người có thể sử dụng công nghệ này như một tấm gương phản chiếu Phật tính vốn sẵn có nơi chính mình.

Khi sống trong một thế giới số ít bị chi phối bởi tham, sân và si, con người có thêm điều kiện để nhận diện và chuyển hóa ba độc ngay từ gốc rễ của tâm. Khi những độc tố ấy dần được chuyển hóa, Thập thiện có thể được thực hành một cách tự nhiên hơn, si mê từng bước lắng dịu, và tâm trở về với những phẩm chất thanh tịnh, bình đẳng và từ bi vốn được đề cao trong Phật giáo.

5. Kết luận

Trong một thế giới được trung gian bởi trí tuệ nhân tạo (AI), tư tưởng bất nhị của Phật giáo cùng những con đường thực hành của truyền thống này cung cấp một khuôn khổ có thể kiểm chứng để định hướng công nghệ theo hướng thiện lành. AI có thể trở thành phương tiện hữu hiệu nhằm “tăng cường các điều kiện hỗ trợ”, giúp con người theo đuổi quá trình tự thực hiện bản thân (self-realization).

Công nghệ này có thể loại bỏ nhiều chướng ngại bên ngoài trên con đường tu tập, chẳng hạn như giảm thiểu sự tác động của các duyên bất thiện và ngăn ngừa sáu căn bị tán loạn. AI góp phần tạo nền tảng cho việc tu dưỡng giới, định và tuệ. Bằng cách kiến tạo những môi trường thanh tịnh và ổn định, AI mở rộng không gian tâm lý thuận lợi cho chính niệm, nơi trí tuệ và tuệ giác có thể phát sinh một cách tự nhiên.



Trong mối quan hệ biện chứng giữa AI và tâm tính, con người không chỉ giáo dục AI về kỷ luật đạo đức, mà còn thực hành lại những giá trị của nền văn minh nhân loại thông qua tấm gương công nghệ. Tuy nhiên, việc hiện thực hóa mục tiêu này không thể tách rời vai trò của những người giám sát và thẩm định là con người. Như Frank Pasquale nhận định trong tác phẩm New Laws of Robotics (Những quy luật mới của robot): *"Ngay khi các thuật toán, đặc biệt là robot tạo ra những tác động trong thế giới thực, chúng phải chịu sự quản lý, còn những người lập trình chúng phải chịu trách nhiệm về mặt đạo đức và pháp lý đối với những tổn hại mà chúng gây ra. [...] Chỉ những người thẩm định là con người, có đủ chuyên môn và thẩm quyền, mới có thể phát hiện, hóa giải và ngăn chặn phần lớn những tổn hại này"* (Pasquale, 2020, tr. 90).

Ý hướng của con người nằm ở cội nguồn của việc thiết kế AI. Trên thực tế, những hành vi đạo đức mà AI dường như thể hiện chính là sự ngoại hiện của ý thức tập thể nơi những người tạo ra nó. Thông qua quá trình thực hành giới Bồ-tát của chính mình, các kỹ sư đưa những lời nguyện từ bi và sức mạnh của nguyện lực vào công nghệ AI, nhằm hiện thực hóa một hệ thống tương tác cộng sinh giữa con người và máy móc. Khi con người giữ vai trò là "chủ thể thọ giới", còn công nghệ AI là "tác nhân trì giới", một tấm gương trong sáng được hình thành, phản chiếu sự thức tỉnh đạo đức của nhân loại.

Như Ethan Mollick, tác giả cuốn Co-Intelligence (Trí tuệ cộng tác), nhận xét: "AI là một tấm gương, phản chiếu trở lại những phẩm chất tốt đẹp nhất và tồi tệ nhất của chúng ta. Chính chúng ta sẽ quyết định những hệ quả của nó, và những lựa chọn ấy sẽ định hình những gì AI thực sự làm vì nhân loại và gây ra cho nhân loại" (Mollick, 2024, tr. 211).

Các văn bản Phật giáo cổ điển cũng thể hiện logic tương tự. A-tỳ-đạt-ma Câu-xá luận (Abhidharmakośa, □□□□□) chỉ rõ rằng "tư chính là nghiệp thuộc về tâm, còn những gì do tư tạo ra cấu thành các hành vi thuộc thân và khẩu" [47]; trong khi Nghiệp thành tựu luận (Karmasiddhi-śāstra, □□□□□□) khẳng định rằng "nghiệp vô ký không được xác lập là nghiệp" [48].

Từ đó có thể suy ra rằng nghiệp lấy yếu tố tâm lý là ý định hay tư (cetanā) [49] làm nền tảng, trong khi các trạng thái vô ký không được gọi là nghiệp và không tạo ra quả báo trong tương lai. Theo cách lập luận này, hệ thống AI, do không có cetanā, tự thân không cấu thành nghiệp; các đầu ra của nó chỉ là những hoạt động chức năng phát sinh từ các điều kiện, và nhiều nhất cũng chỉ có thể được ví với phạm trù vô ký. Trái lại, ý hướng trong quá trình thiết kế và sử dụng AI của các kỹ sư và người dùng, những ý hướng được gọi là cetanā có hiệu lực về mặt nghiệp. Chúng có thể khởi động các hành vi thuộc thân và khẩu như những hành động được thực hiện sau khi đã khởi ý, từ đó dẫn đến những kết quả tương ứng.



Vì vậy, chính ý hướng của con người trong thiết kế và sử dụng AI mới mang sức nặng của nghiệp. Theo nghĩa này, các quyết định của AI là sự mở rộng ý chí con người: yếu tố mang tính quyết định không phải là cỗ máy, mà là lựa chọn và sự tu dưỡng của con người. Do đó, nâng cao kỷ luật đạo đức của toàn thể nhân loại có thể định hình những tác động thực tế của AI sâu sắc hơn so với việc chỉ cải tiến kỹ thuật.

Nói cách khác, việc thực hành đạo đức Phật giáo thông qua AI về bản chất là quá trình con người thực hành và tự chứng nghiệm Phật tính của chính mình

thông qua công nghệ. Điều này buộc chúng ta phải một lần nữa trả lời những câu hỏi muôn thuở trong thời đại thuật toán: Trí tuệ đích thực là gì? Từ bi tối hậu là gì?

Khi AI bắt đầu mô phỏng con đường của những bậc giác ngộ, có lẽ hơn bao giờ hết, chúng ta cần trở lại với những lời dạy nguyên thủy của đức Phật: *"Lấy giới làm thầy", "Không làm các điều ác, vâng làm các điều lành".* Với mục tiêu chuyển hóa từ điều sai trái sang điều thiện thông qua sự tự rèn luyện, chúng ta có thể xem công nghệ như một trợ duyên bên ngoài để không ngừng thanh lọc và mở rộng những khả năng tạo tác thiện nghiệp của con người. Con đường khám phá này không có điểm đến cuối cùng, nhưng mỗi bước đi đều làm hiển lộ chiều kích tâm linh của nền văn minh công nghệ.

Thước đo đích thực của AI không chỉ nằm ở việc có tuân thủ những quy tắc đạo đức được xác lập dưới hình thức giới điều hay không, mà còn ở việc những thực hành ấy có tạo ra những phản hồi tích cực trong đời sống con người hay không.

Chẳng hạn, với các chatbot, hệ thống có thể bắt đầu từ yêu cầu cơ bản là tránh lời nói thô ác, ngăn chặn ngôn từ lăng mạ; tiếp đó chủ động an ủi bằng những lời từ bi; và xa hơn nữa, hướng dẫn người dùng buông bỏ sự chấp trước vào khổ đau. AI có thể triển khai một nguyên tắc chứng ngộ theo từng cấp độ, hỗ trợ con người vượt qua những giới hạn của nhận thức về bản thân và chuyển hóa tâm để vượt qua khó khăn.

Một hệ thống "AI trì giới" thể hiện sự tự chế về mặt đạo đức và bảo vệ những ranh giới luân lý trong tiến trình này. Một hệ thống "AI từ bi" góp phần nâng cao an lạc của con người. Một hệ thống "AI thấm nhuần tư tưởng tính không" giúp nới lỏng sự chấp trước vào cái tôi. Còn một hệ thống "AI thiện xảo diệu dụng" hỗ trợ thực hiện những công việc trong đời sống thế tục. Một hệ thống AI như vậy sẽ trở thành phương tiện quan trọng hỗ trợ hành trình hướng đến giác ngộ và chứng nghiệm Phật tính của con người.

Việc AI tham gia vào quá trình con người tự thực hiện bản thân, kết hợp với lý thuyết Phật giáo về tâm (citta), về bản chất là sự tương tác hai chiều giữa "công nghệ chuyển hóa thế giới" và "tâm tính chuyển hóa đời sống". Trong ngắn hạn, AI có thể thúc đẩy điều mà Maslow gọi là "tự hiện thực hóa" (self-actualization). Về lâu dài, công nghệ này thậm chí có thể mô phỏng những trạng thái nhận thức của người đã giác ngộ.

Trong mạng lưới tương thuộc giữa con người và máy móc, AI có thể góp phần hiện thực hóa lý tưởng Phật giáo nhân gian về việc "thành Phật ngay trong đời này". Mục tiêu của nghiên cứu liên ngành này không phải là tạo ra những cỗ

máy mạnh mẽ hơn, mà là giúp quá trình “tiến hóa tự chủ” của AI thúc đẩy sự phát triển tiềm năng vốn có của nhân loại, hỗ trợ con người tự mình thức tỉnh, thấy thực tại đúng như nó đang là và góp phần hoàn thiện đời sống.

Tác giả: **Jia Liu (ORCID)**

Trường Triết học, Đại học Nam Kinh, Nam Kinh 210000, Trung Quốc

Chuyển ngữ: **Thường Nguyên**

Nguồn link bài gốc: https://www.mdpi.com/2077-1444/17/3/378

Religions 2026, 17(3), 378. DOI: https://doi.org/10.3390/rel17030378

Tài trợ: Nghiên cứu này được tài trợ bởi Chương trình Đổi mới Nghiên cứu và Thực hành Sau đại học tỉnh Giang Tô (Postgraduate Research & Practice Innovation Program of Jiangsu Province), mã dự án KYCX25_0014. Tên dự án bằng tiếng Trung: “江蘇省研究生科研與實踐創新計畫項目”.

Tuyên bố về Hội đồng Đạo đức Nghiên cứu

Không áp dụng.

Tuyên bố về sự đồng thuận tham gia nghiên cứu

Không áp dụng.

Tuyên bố về tính sẵn có của dữ liệu

Nghiên cứu này không tạo ra hoặc phân tích dữ liệu mới.

Xung đột lợi ích

Tác giả tuyên bố không có xung đột lợi ích.

Danh mục chữ viết tắt

Chữ viết tắt sau đây được sử dụng trong bản thảo:

CBETA - Comprehensive Buddhist Electronic Text Archive Foundation (Tổ chức Lưu trữ Điện tử Toàn diện Kinh điển Phật giáo).

Chú thích:

41] Pháp môn Tâm địa (xindi famen 心地法門): Một khuôn khổ giáo lý xem tâm là “đất” hay “mảnh đất” làm nền tảng cho sự phát sinh của các pháp thiện và bất thiện. Vì vậy, việc tu tập tập trung vào nuôi dưỡng Tâm địa bằng cách thanh

Vô ngã trước thuật toán: Phật học đối thoại cùng trí tuệ nhân tạo (P.3 - Hết)
https://phathoc.vn/vo-nga-truoc-thuat-toan-phat-hoc-doi-thoai-cung-tri-tue-nhan-tao-p-3-het.html

lọc phiền não, phát triển trí tuệ và từ bi, thay vì tìm cầu các đối tượng bên ngoài.

42] CBETA T48, số 2008, tr. 358, cột c, dòng 12-13.

43] Ba định luật robot nguyên thủy của Isaac Asimov như sau:

1) Robot không được gây tổn hại cho con người, hoặc do không hành động mà để con người bị tổn hại.

2) Robot phải tuân theo mệnh lệnh của con người, trừ khi những mệnh lệnh đó mâu thuẫn với Định luật thứ nhất.

3) Robot phải bảo vệ sự tồn tại của chính mình, miễn là việc bảo vệ ấy không mâu thuẫn với Định luật thứ nhất hoặc thứ hai.

Nguồn trích dẫn: Asimov (1942).

44] Ngũ giới (pañca-śīla □□): Năm giới là những nguyên tắc tu tập căn bản nhằm ngăn ngừa tổn hại và xây dựng lòng tin. Ngũ giới thường được diễn đạt thành năm điều: không sát sinh, không trộm cắp, không tà hạnh trong các quan hệ tình dục, không nói dối và không sử dụng các chất gây say nghiện.

45] Thập thiện (□□): Thập thiện nghiệp đạo mở rộng đạo đức từ nguyên tắc "không gây hại" sang "khuyến khích điều thiện". Mười nghiệp thiện được phân chia theo ba phương diện: thân, khẩu và ý.

+ Thân: Không sát sinh, không trộm cắp, không tà hạnh trong các quan hệ tình dục.

+ Khẩu: Không nói dối, không nói lời chia rẽ, không nói lời thô ác, không nói lời phù phiếm hoặc lời nói mang tính thao túng.

+ Ý: Không tham cầu chiếm hữu, không có ác ý, giữ chính kiến.

46] Một số ấn bản và văn bản tiêu biểu được tham khảo gồm:

+ Tử môn sùng hạnh lục (Zimen chongxing lu □□□□), CBETA X87, số 1627.

+ Liễu Phàm tứ huấn (Liaofan sixun □□□), do Liễu Phàm biên soạn, Chiu-Nan Lai dịch, Great Enlightenment Lotus Society xuất bản năm 1991. https://www.worldcat.org/title/liaofan-si-xun/oclc/52172213 (truy cập ngày 10/10/2025).

+ Phẩm “Phổ Hiền hạnh nguyện” (Puxian xingyuan pin □□□□□), phẩm thứ 40 của Đại phương quảng Phật Hoa nghiêm kinh (□□□□□□□), CBETA T10, số 279.

+ Cách ngôn liên bích (Geyan lianbi □□□□), văn bản truy cập mở hiện đại tại Chinese Text Project (ctext): https://ctext.org/wiki.pl?if=gb&res=69017 (truy cập ngày 10/10/2025).

+ Tây phương công cứ (Xifang gongju □□□□), bản hiệu đính Trùng đính Tây phương công cứ (□□□□□□), CBETA X62, số 1180.

+ An Sĩ toàn thư (Anshi quanshu □□□□): https://upload.wikimedia.org/wikipedia/commons/a/a6/SSID-11338242_%E5%AE%89%E5%A3%AB%E5%85%A8%E6%9B%B8_%E4%B8%8A.pdf (truy cập ngày 10/10/2025).

+ Tịnh độ tập yếu (Jingtu jiyao □□□□): https://bookgb.bfnn.org/books/0867.htm (truy cập ngày 10/10/2025).

+ Nhân sinh bách sự (Rensheng baishi □□□□) của Hòa thượng Tinh Vân, đăng trên Phổ Môn Học Báo (□□□□), số 33 (2006). Toàn văn PDF tại Thư viện số Phật giáo Đại học Quốc gia Đài Loan (NTU Buddhist Digital Library): https://buddhism.lib.ntu.edu.tw/FULLTEXT/JR-MAG/mag204161.pdf (truy cập ngày 10/10/2025).

47] CBETA T29, số 1558, tr. 67, cột b, dòng 10: “□□□□□□□□□□ ” – “Tư chính là ý nghiệp; những gì được tạo tác là các hành vi thuộc thân và khẩu”.

Câu này giải thích quan niệm nghiệp trong Phật giáo cổ điển: chính tư hay ý định là ý nghiệp, còn các hành vi thuộc thân và khẩu là sự biểu hiện của ý định ấy qua hành động. Nói ngắn gọn, ý định giữ vai trò nền tảng; thân nghiệp và khẩu nghiệp là những biểu hiện bên ngoài của nó.

48] CBETA T31, số 1608, tr. 778, cột c, dòng 7: “□□□□□□ ” – “Nghiệp vô ký thì không thể thành lập [thành nghiệp]”.

Vô ký (avyākṛta, □□) không cấu thành nghiệp theo nghĩa nghiệp đạo đức có khả năng tạo quả báo.

49] Tư tâm sở (cetanā □□□): Yếu tố tâm lý mang tính ý chí, có chức năng thúc đẩy và điều phối một sát-na hay tiến trình tâm, hướng các trạng thái tâm liên hệ đến một đối tượng và thúc đẩy hành động. Trong Phật giáo thời kỳ đầu, cetanā được đồng nhất với nghiệp; các hành vi thuộc thân và khẩu là sự biểu

hiện của tư qua hành động.

[Lưu ý học thuật]:

Về tính nhất quán của thuật ngữ

+ *Mind-nature - tâm tính:* Cần phân biệt với mind (tâm hoặc tâm thức), self-nature (tự tính) và Buddha-nature (Phật tính). Đây là những khái niệm có quan hệ nhưng không thể mặc nhiên dùng thay thế cho nhau.

+ *Self-realization và self-actualization:* Nên giữ hai cách dịch riêng biệt. Self-actualization trong tâm lý học Maslow thường dịch là "tự hiện thực hóa"; self-realization trong bài cần cân nhắc dịch thống nhất là "tự chứng nghiệm bản thân" hoặc "tự thực hiện bản thân", tùy ý nghĩa mà tác giả muốn nhấn mạnh. Không nên tự động đồng nhất hai thuật ngữ này với giác ngộ Phật giáo.

+ *Cetanā – tư/ý định:* Nên thống nhất cách gọi là "tư (cetanā)" khi bàn đến thuật ngữ Phật học; có thể giải thích thêm là yếu tố tâm lý mang tính ý chí hoặc ý định. Không nên giản lược mọi nghĩa của cetanā thành ý định theo nghĩa thông thường.

+ Karma - nghiệp: Phân biệt nghiệp với hành vi quan sát được, hậu quả xã hội và kết quả do máy tạo ra. Luận điểm cho rằng AI không tự tạo nghiệp vì không có cetanā là lập luận của tác giả dựa trên một cách hiểu Phật học, không nên trình bày như kết luận đã được mọi trường phái Phật giáo thống nhất.

+ *Digital precepts - giới luật số:* Nên hiểu là khung nguyên tắc đạo đức được vận dụng vào thiết kế, huấn luyện và sử dụng AI, không phải một hệ thống giới luật Phật giáo cổ truyền đã có sẵn dưới hình thức này.

Về các luận điểm cần diễn đạt thận trọng

Thứ nhất, AI và việc tu dưỡng tâm. Những cách diễn đạt như AI "thanh lọc phiền não", "nuôi dưỡng thiện căn" hay hỗ trợ con người tiến tới giác ngộ cần được đọc trong phạm vi lập luận mang tính định hướng và giả thuyết của tác giả. AI có thể hỗ trợ tạo môi trường, nhắc nhở thực hành hoặc cung cấp hướng dẫn; không nên khiến độc giả hiểu rằng công nghệ tự thân có khả năng chuyển hóa tâm thức hay bảo đảm chứng ngộ.

Thứ hai, AI không có ý định và nghiệp. Tác giả vận dụng khái niệm cetanā để phân biệt nghiệp của con người với hoạt động chức năng của máy. Cần giữ rõ đây là lập luận triết học - Phật học. Trong thực tế, hệ thống AI vẫn có thể gây

tổn hại thông qua thiết kế, dữ liệu, cách triển khai và kết quả đầu ra; việc không quy nghiệp theo nghĩa Phật học cho máy không làm giảm trách nhiệm đạo đức, pháp lý của người thiết kế, đơn vị triển khai và người sử dụng.

Thứ ba, "AI mô phỏng trạng thái nhận thức của người giác ngộ". Đây là một mệnh đề có tính suy đoán. Cần giữ sắc thái khả năng hoặc giả định trong bản dịch, tránh diễn đạt thành khẳng định rằng AI có thể đạt giác ngộ hay có trải nghiệm tâm linh như con người.

Thứ tư, Ba định luật robot của Asimov. Cần giữ đúng tính chất lịch sử và văn học của bộ quy tắc này. Không nên trình bày chúng như chuẩn mực pháp lý hoặc tiêu chuẩn kỹ thuật hiện hành đã đủ để bảo đảm AI an toàn.

Về trích dẫn và tư liệu

+ Đối chiếu lại số trang, số tập, số hiệu văn bản và ký hiệu cột trong các dẫn chiếu CBETA, nhất là chú thích 42, 47 và 48.

+ Kiểm tra nguyên văn Hán văn, bản dịch và ngữ cảnh của hai câu được dẫn từ A-tỳ-đạt-ma Câu-xá luận và Nghiệp thành tựu luận. Cần tránh để bản dịch diễn giải rộng hơn nội dung nguyên văn.

+ Với chú thích 46, nên thống nhất cách trình bày nhan đề, tên tác giả/dịch giả, thông tin xuất bản, đường dẫn và ngày truy cập theo quy cách trích dẫn mà Tạp chí đang áp dụng.

+ Với trích dẫn Frank Pasquale và Ethan Mollick, cần kiểm tra lại nguyên văn tiếng Anh và số trang trong đúng ấn bản được sử dụng.

+ Phần tài trợ ghi mã dự án KYCX25_0014 cần giữ nguyên. Các tuyên bố "không áp dụng", "không có dữ liệu mới" và "không có xung đột lợi ích" nên được trình bày theo đúng mẫu tạp chí hoặc yêu cầu của bản thảo gốc.

Về một số lựa chọn từ ngữ

Thuật ngữ gốc	Cách dịch đề xuất	Lưu ý
Mind-ground	Tâm địa	Giữ nhất quán với "Pháp môn Tâm địa"
Wholesome / unwholesome	Thiện / bất thiện	Phù hợp ngữ cảnh Phật học hơn "tốt / xấu"
Non-harming	Bất hại / không gây hại	"Bất hại" khi nói về tinh thần đạo đức; "không gây hại" khi nói về thiết kế
Reward function	Hàm thưởng	Không đồng nhất với "hàm mục tiêu"
Reinforcement learning	Học tăng cường	Thuật ngữ kỹ thuật
Ethical agency	Năng lực chủ thể đạo đức	Cần phân biệt với năng lực thực hiện hành động
Human oversight	Sự giám sát của con người	Nhấn mạnh trách nhiệm và quyền kiểm soát thực tế

Lưu ý chung: phần cuối có nhiều thuật ngữ thuộc ba lĩnh vực: Phật học, tâm lý học và kỹ thuật AI. Khi biên tập, ưu tiên nhất quán về khái niệm, không chỉ thống nhất mặt chữ.

[Góc nhìn biên tập phần III và lời bàn toàn bài]:

Phần III là phần đưa lập luận của toàn bài đến bình diện ứng dụng: từ tư tưởng bất nhị và quan niệm Phật giáo về tâm tính, tác giả chuyển sang những câu hỏi cụ thể về thiết kế, huấn luyện và sử dụng AI có trách nhiệm.

Có thể nhận diện ba trục lập luận chính.

Trục thứ nhất: từ nguyên tắc đạo đức Phật giáo đến thiết kế AI.

Tác giả vận dụng Ngũ giới, Thập thiện nghiệp đạo và tinh thần phòng hộ các căn để gợi mở những nguyên tắc có thể chuyển hóa thành định hướng thiết kế công nghệ. Điểm đáng chú ý không chỉ là yêu cầu AI tránh tạo ra nội dung có hại, mà còn là khả năng hỗ trợ con người ứng xử thiện lành, giảm xung đột và phát triển những thói quen có ích.

Đây là hướng tiếp cận có giá trị ứng dụng. Tuy nhiên, cần phân biệt giới luật trong đời sống tu tập với những quy tắc vận hành của hệ thống AI. Giới luật gắn với sự tu dưỡng của chủ thể; còn quy tắc kỹ thuật cần được cụ thể hóa thành tiêu chí, quy trình đánh giá và cơ chế kiểm soát có thể kiểm chứng.

Trục thứ hai: từ tâm tính đến trách nhiệm của con người.

Lập luận về cetanā giúp tác giả nhấn mạnh rằng công nghệ không thể được xem xét tách rời ý hướng của người tạo ra và sử dụng nó. Một hệ thống AI có thể biểu hiện những khuôn mẫu ngôn ngữ có vẻ từ bi hoặc đạo đức, nhưng biểu hiện đó không tự chứng minh rằng hệ thống có tâm từ bi hay trải nghiệm đạo đức như con người.

Điểm mạnh của phần này là đưa trách nhiệm trở lại với con người. Người thiết kế lựa chọn mục tiêu, dữ liệu và cơ chế huấn luyện; đơn vị triển khai quyết định môi trường sử dụng; người dùng lựa chọn cách vận dụng công nghệ. Các quyết định ấy tạo ra những hệ quả thực tế cần được đánh giá và chịu trách nhiệm.

Dẫu vậy, câu khẳng định rằng AI không có cetanā nên bản thân nó không tạo nghiệp cần được giữ trong phạm vi lập luận Phật học của tác giả. Không nên từ đó suy ra rằng các hệ thống tự động không cần được kiểm soát hoặc không có vấn đề về trách nhiệm trong quản trị công nghệ.

Trục thứ ba: từ công nghệ hỗ trợ đến sự tự chuyển hóa của con người.

Các mô hình “AI trì giới”, “AI từ bi”, “AI thấm nhuần tư tưởng tính không” và “AI thiện xảo diệu dụng” là những đề xuất mang tính khái niệm. Chúng giúp người đọc hình dung các định hướng ứng dụng khác nhau: hạn chế tổn hại, hỗ trợ an lạc, giảm sự củng cố chấp trước vào cái tôi và giải quyết những công việc thực tiễn.

Tuy nhiên, đây nên được hiểu là các mô hình định hướng cho việc thiết kế và đánh giá AI, không phải những loại hình AI đã được xác lập như một hệ phân loại kỹ thuật chuẩn hóa.

Điểm cần giữ xuyên suốt là: công nghệ có thể tạo điều kiện thuận lợi cho việc tu dưỡng, nhưng không thể thay thế sự thực hành, nhận thức và chuyển hóa của con người. Cũng không nên đồng nhất khả năng mô phỏng ngôn ngữ đạo đức với trí tuệ giải thoát.

Giá trị nổi bật của toàn bài

Loạt bài Vô ngã trước thuật toán: Phật học đối thoại cùng trí tuệ nhân tạo có điểm đáng chú ý ở nỗ lực đưa Phật học vào cuộc đối thoại với một vấn đề đương đại, thay vì chỉ sử dụng giáo lý như lời dẫn nhập mang tính minh họa.

Qua ba phần, bài viết triển khai một hướng tiếp cận có thể khái quát như sau:

+ Từ vô ngã và tư tưởng bất nhị: đặt lại câu hỏi về bản ngã, chủ thể và những giả định nền tảng khi con người tương tác với AI.

+ Từ tâm tính và quá trình nhận thức: mở rộng đối thoại giữa quan niệm Phật giáo về tâm với các mô hình trí tuệ nhân tạo, đồng thời xem xét giới hạn của sự mô phỏng.

+ Từ giới, nghiệp và thực hành đạo đức: chuyển các suy tư triết học thành những gợi ý về thiết kế AI, trách nhiệm con người và định hướng ứng dụng công nghệ.

Cấu trúc này tạo được tiến trình từ nền tảng tư tưởng đến vấn đề nhận thức, rồi đến thực hành. Đó cũng là yếu tố giúp bài viết có khả năng tiếp cận độc giả quan tâm đến Phật học ứng dụng, đạo đức công nghệ và những biến đổi trong đời sống hiện đại.

Điểm cần đặc biệt lưu ý khi biên tập

Thứ nhất, cần phân biệt rõ ba tầng phát biểu:

1] Giáo lý và tư liệu Phật học: những gì có thể đối chiếu với kinh luận, thuật ngữ và lịch sử tư tưởng.

2] Đối thoại liên ngành: những cách tác giả vận dụng khái niệm Phật học để diễn giải AI và tâm lý con người.

3] Đề xuất hoặc giả thuyết: những khả năng AI có thể hỗ trợ chuyển hóa hành vi, mô phỏng trạng thái nhận thức hoặc trở thành phương tiện cho sự tự hoàn thiện của con người.

Ba tầng này có thể bổ trợ cho nhau, nhưng không nên nhập làm một. Việc phân định rõ sẽ giúp bài giữ được chiều sâu tư tưởng mà vẫn thận trọng về mặt học thuật.

Thứ hai, cần tránh để ngôn ngữ Phật học khiến độc giả hiểu rằng AI đã có tâm thức, Phật tính hoặc năng lực tu chứng theo nghĩa con người. Tác giả có thể sử dụng những khái niệm này như lăng kính triết học để đánh giá công nghệ, nhưng cần phân biệt giữa ẩn dụ, mô hình chuẩn tắc và khẳng định về bản thể.

Thứ ba, các khái niệm “vô ngã”, “bất nhị”, “tâm tính” và “tính không” có lịch sử diễn giải phong phú trong Phật giáo. Nếu bài sử dụng chúng để đề xuất một mô hình AI, cần làm rõ đó là cách vận dụng của tác giả, tránh tạo cảm giác có một cách hiểu duy nhất được mọi truyền thống Phật giáo thừa nhận.

Thứ tư, bài có nhiều tiềm năng ứng dụng nhưng nên giữ đúng mức độ khẳng định. Những đề xuất như AI từ bi hay AI trì giới có thể mở ra hướng nghiên cứu về tiêu chí đánh giá hành vi, cơ chế phòng ngừa tổn hại và hỗ trợ người dùng. Tuy nhiên, muốn chuyển thành mô hình triển khai thực tế, cần có tiêu chí vận hành cụ thể, quy trình kiểm thử, đánh giá tác động và cơ chế giám sát của con người.

Tổng hợp chú giải thuật ngữ:

Bảng dưới đây tập trung vào những thuật ngữ Phật học, đạo đức và kỹ thuật xuất hiện trong phần III, đặc biệt ở phần kết luận và các chú thích 41-49. Đây là bảng thuật ngữ phục vụ biên tập; các thuật ngữ xuất hiện ở phần I và II cần tiếp tục được đối chiếu để bảo đảm thống nhất toàn loạt bài.

Thuật ngữ gốc	Cách dịch đề xuất	Chú giải biên tập
Dharma of the Mind-Ground (xindi famen, 心地法門)	Pháp môn Tâm địa	Cách tiếp cận lấy tâm làm nền tảng tu tập, nhấn mạnh thanh lọc phiền não và phát triển trí tuệ, từ bi.
Mind-ground	Tâm địa	Hình ảnh ẩn dụ về tâm như mảnh đất nơi các pháp thiện và bất thiện phát sinh.
Mind-nature	Tâm tính	Cần phân biệt với tâm, tự tính và Phật tính; phải căn cứ ngữ cảnh để xác định nghĩa cụ thể.
Mind (citta)	Tâm / tâm thức	Trong Phật học, citta có các sắc thái nghĩa tùy hệ thống giáo lý và ngữ cảnh.
Buddha-nature	Phật tính	Không tự động đồng nhất với mind-nature hay với một khái niệm về bản ngã thường hằng.

Non-dualism	Tư tưởng bất nhị	Quan điểm vượt khỏi cách phân chia nhị nguyên cứng nhắc; cần diễn giải theo ngữ cảnh triết học cụ thể.
Emptiness	Tính không	Không đồng nghĩa với hư vô hay sự không tồn tại tuyệt đối.
Five Precepts (pañca-śīla, 五戒)	Ngũ giới	Năm nguyên tắc tu tập căn bản nhằm ngăn ngừa tổn hại và điều chỉnh hành vi.
Ten Wholesome Deeds / Courses of Action (十善)	Thập thiện / Thập thiện nghiệp đạo	Hệ thống mười nghiệp thiện thuộc thân, khẩu và ý.
Kusala-kammapatha	Thiện nghiệp đạo	Chỉ con đường hay nhóm hành vi thiện; nên giữ nhất quán theo cách dùng của toàn bài.

Afflictions	Phiền não	Những trạng thái tâm gây trở ngại cho nhận thức và tu tập.
Guard the sense faculties	Phòng hộ các căn	Thực hành điều phục sự tiếp xúc của các giác quan với đối tượng; không nên giản lược thành né tránh mọi kích thích.
Yoniso-manasikāra	Như lý tác ý	Hướng tâm xem xét sự vật đúng theo bản chất và điều kiện của chúng.
Cetanā (cetanā, 思心所)	Tư / tư tâm sở	Yếu tố tâm lý mang tính ý chí, có vai trò quan trọng trong phân tích nghiệp của Phật giáo.
Mental karma	Ý nghiệp	Nghiệp thuộc phương diện tâm ý; cần phân biệt với hành vi biểu hiện qua thân và khẩu.
Body-karma	Thân nghiệp	Hành vi nghiệp thuộc thân.
Speech-karma	Khẩu nghiệp	Hành vi nghiệp thuộc lời nói.

Avyākṛta (無記)	Vô ký	Phạm trù chỉ trạng thái không được xác định là thiện hay bất thiện trong hệ thống phân loại tương ứng; không nên giản lược thành "không có hậu quả".
Karma	Nghiệp	Trong ngữ cảnh bài, cần chú ý vai trò của tư/ý định và phân biệt với mọi kết quả máy tạo ra.
Digital precepts	Giới luật số	Cách vận dụng nguyên tắc đạo đức vào môi trường số và thiết kế AI; không phải tên gọi một bộ giới luật cổ điển.
Friendly AI	AI thân thiện	Khái niệm về AI được định hướng nhằm phù hợp với lợi ích và sự an toàn của con người.
Reinforcement learning	Học tăng cường	Phương pháp học máy dựa trên tương tác và tín hiệu thưởng.

Reward function	Hàm thưởng	Quy tắc xác định tín hiệu thưởng cho hành động hoặc kết quả của hệ thống.
Reward shaping	Điều chỉnh định hình phần thưởng	Thiết kế hoặc điều chỉnh tín hiệu thưởng nhằm định hướng quá trình học.
Reward modeling	Mô hình hóa phần thưởng	Xây dựng mô hình ước lượng mức độ phù hợp hoặc giá trị của đầu ra/hành vi.
Objective function	Hàm mục tiêu	Hàm biểu diễn mục tiêu mà quá trình tối ưu hóa hướng đến. Không nên dùng thay cho "hàm thưởng".
Instruction tuning	Tinh chỉnh theo chỉ dẫn	Giai đoạn huấn luyện để hệ thống đáp ứng chỉ dẫn của người dùng tốt hơn.
Intelligent agent	Tác tử thông minh	Hệ thống có khả năng tiếp nhận thông tin và thực hiện hành động theo mục tiêu hoặc cơ chế được thiết kế.

Human oversight	Sự giám sát của con người	Cơ chế kiểm soát, đánh giá và can thiệp của con người đối với hệ thống AI.
Ethical agency	Năng lực chủ thể đạo đức	Khái niệm về khả năng hành động và chịu trách nhiệm đạo đức; không nên mặc nhiên gán cho AI chỉ vì hệ thống tạo ra lời nói có vẻ đạo đức.
Non-harming	Bất hại / không gây hại	"Bất hại" phù hợp khi bàn về tinh thần đạo đức; "không gây hại" thường rõ nghĩa hơn trong ngữ cảnh kỹ thuật.
Self-realization	Tự chứng nghiệm bản thân / tự thực hiện bản thân	Cần kiểm tra ngữ cảnh của tác giả và thống nhất với các phần trước; không mặc nhiên đồng nhất với giác ngộ.

Self-actualization	Tự hiện thực hóa	Thuật ngữ thường gắn với tâm lý học nhân văn, đặc biệt là Maslow.
Ethical discipline	Kỷ luật đạo đức	Sự điều chỉnh hành vi theo các nguyên tắc đạo đức.
Compassionate AI	AI từ bi	Nhãn khái niệm định hướng hệ thống theo hướng hỗ trợ, giảm tổn hại và nâng cao an lạc; không khẳng định AI thực sự có tâm từ bi.
Precept-keeping AI	AI trì giới	Cách gọi mang tính đề xuất cho hệ thống được thiết kế nhằm tuân thủ các ràng buộc đạo đức.
Emptiness-informed AI	AI thấm nhuần tư tưởng tính không	Nên hiểu là AI được thiết kế hoặc vận dụng theo một định hướng triết học, không phải hệ thống có chứng nghiệm tính không.
Skillful AI of wondrous functioning	AI thiện xảo diệu dụng	Nhãn khái niệm nhấn mạnh khả năng vận dụng phương tiện phù hợp để giải quyết các công việc thực tiễn.

Bodhisattva precepts	Giới Bồ-tát	Hệ thống giới gắn với lý tưởng Bồ-tát; cần phân biệt với quy tắc kỹ thuật dùng để điều chỉnh AI.
Three Laws of Robotics	Ba định luật robot	Ba nguyên tắc hư cấu nổi tiếng của Isaac Asimov; không phải bộ tiêu chuẩn pháp lý hoặc kỹ thuật hiện hành.
Comprehensive Buddhist Electronic Text Archive Foundation (CBETA)	Tổ chức Lưu trữ Điện tử Toàn diện Kinh điển Phật giáo	Tên tổ chức được nêu trong danh mục chữ viết tắt của bản thảo; nên giữ nguyên chữ viết tắt CBETA.

Tài liệu tham khảo:

Alammar, Jay, and Maarten Grootendorst. 2024. Hands-On Large Language Models: Language Understanding and Generation. Sebastopol: O'Reilly Media. ISBN 978-1-098-15096-9. [Google Scholar]

Asimov, Isaac. 1942. "Runaround". Astounding Science-Fiction 29: 94–103. Available online: https://archive.org/details/Astounding_v29n01_1942-03_dtsg0318 (accessed on 10 October 2025).

BodhiAI. 2024. Available online: https://bodhiai.cn/ (accessed on 10 October 2025).

Braidotti, Rosi. 2013. The Posthuman. Cambridge: Polity Press, ISBN-13 978-0-7456-4157-7. [Google Scholar]

Bryson, Joanna J. 2020. The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In The Oxford Handbook of Ethics of AI. Edited by Markus D. Dubber, Frank Pasquale and Sunit Das. Oxford: Oxford University Press, pp. 2–25. [Google Scholar] [CrossRef] [Scilit]

Buswell, Robert E., Jr., and Donald S. Lopez, Jr., eds. 2013. The Princeton Dictionary of Buddhism. Princeton: Princeton University Press. [Google Scholar]

Butlin, Patrick, and Theodoros Lappas. 2025. Principles for Responsible AI Consciousness Research. Journal of Artificial Intelligence Research 82: 1673–90. [Google Scholar] [CrossRef] [Scilit]

Center for AI Safety. 2023. Statement on AI Risk. May 30. Available online: https://www.safe.ai/statement-on-ai-risk/ (accessed on 10 October 2025).

Chalmers, David J. 2023. Could a Large Language Model be Conscious? arXiv. [Google Scholar] [CrossRef] [Scilit]

Chanxiu Jishi. 2024. □□□□ —Apps on Google Play. Google Play. Available online: https://play.google.com/store/apps/details?id=cn.sbo2it.fgstimer.fc (accessed on 10 October 2025).

Chen, Yixiao □□□ . 2004. Foxue changjian cihui □□□□□ A common dictionary of Buddhist terms. Shanghai: Heyu chubanshe □□□□ , p. 102. [Google Scholar]

Cheong, Pauline Hope. 2021. Bounded Religious Automation at Work: Communicating Human Authority in Artificial Intelligence Networks. Journal of Communication Inquiry 45: 5–23. [Google Scholar] [CrossRef] [Scilit]

Chomski, Omer. 2025. Optimizing the Immeasurable: A Critical Examination of Neuro-Meditation Commodification. AI & Society. [Google Scholar] [CrossRef] [Scilit]

Deerpark AI. 2025. Available online: https://deerpark.ai/ (accessed on 10 October 2025).

Dharmamitra. 2025. Dharmamitra · a Toolkit for the Languages of the Dharma. Available online: https://dharmamitra.org/ (accessed on 10 October 2025).

Fazi, Beatrice. 2021. Beyond Human: Deep Learning, Explainability and Representation. Theory, Culture & Society 38: 125–46. [Google Scholar] [CrossRef] [Scilit]

Fazi, Beatrice. 2024. Machines That Create: Contingent Computation and Generative AI. Media Theory 8: 1–12. [Google Scholar] [CrossRef] [Scilit]

Fo Guang Shan, ed. 1998. Fo Guang Dictionary of Buddhism □□□□ , 2nd ed. Gaoxiong: Fo Guang Cultural Enterprise Co., Ltd., p. 4194. [Google Scholar]

Garfield, Jay L. 2015. Engaging Buddhism: Why It Matters to Philosophy. New York: Oxford University Press. [Google Scholar] [CrossRef] [Scilit]

Gould, Hannah, and Keiko Nishimura. 2024. The Buddha in AI/Robotics. In The Cambridge Companion to Religion and Artificial Intelligence. Edited by Beth Singler and Fraser Watts. Cambridge: Cambridge University Press, pp. 50–68. [Google Scholar] [CrossRef] [Scilit]

Grieve, Gregory Price. 2017. Cyber Zen: Imagining Authentic Buddhist Identity, Community, and Practices in the Virtual World of Second Life. London and New York: Routledge. [Google Scholar]

Harari, Yuval Noah. 2024. Nexus: A Brief History of Information Networks from the Stone Age to AI. New York: Random House. [Google Scholar]

Haraway, Donna. 1991. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In Simians, Cyborgs, and Women: The Reinvention of Nature. New York: Routledge, pp. 149–81. [Google Scholar]

Hayles, N. Katherine. 1999. How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics. Chicago: University of Chicago Press. [Google Scholar]

Headspace. 2025. Headspace: Meditation & Health. Google Play. Available online: https://www.headspace.com/app (accessed on 10 October 2025).

Hershock, Peter D. 2021. Buddhism and Intelligent Technology: Toward a More Humane Future. London: Bloomsbury. [Google Scholar]

Hongladarom, Soraj. 2020. The Ethics of AI and Robotics: A Buddhist Viewpoint. Lanham Boulder, New York and London: Lexington Books. [Google Scholar]

Hotoke AI. 2025. HOTOKE AI—AI □□□□□□□□□□□□□□□□ . Available online: https://hotoke.ai/ (accessed on 10 October 2025).

Hsing Yun (□□□). 2005. Shengming de mima (Miwu zhijian 4) □□□□□□□□□□ 4□ The Password of Life (Between Ignorance and Enlightenment 4). Beijing: Zongjiao Wenhua Chubanshe. □□ : □□□□□□□ . ISBN 7801237331. [Google Scholar]

Huang, Yutan, Chetan Arora, Wen Cheng Houng, Tanjila Kanij, Anuradha Madulgalla, and John Grundy. 2025. Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study. arXiv. [Google Scholar] [CrossRef] [Scilit]

Hughes, James J. 2019. Buddhism and Our Posthuman Future. Sophia 58: 653–62. [Google Scholar] [CrossRef] [Scilit]

Josipovic, Zoran. 2024. Reflexivity Gradient—Consciousness Knowing Itself. Frontiers in Psychology 15: 1450553. [Google Scholar] [CrossRef] [Scilit] [PubMed]

Keane, Jonathan, Sam Keyser, and Jeremy Kedziora. 2025. Strategy Masking: A Method for Guardrails in Value-Based Reinforcement Learning Agents. arXiv. [Google Scholar] [CrossRef] [Scilit]

Kim, Taewan, Seolyeong Bae, Hyun Ah Kim, Su Woo Lee, Hwajung Hong, Chanmo Yang, and Young Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Article 701. New York: Association for Computing Machinery, pp. 1–20. [Google Scholar] [CrossRef] [Scilit]

Lee, Peter. 2016. Learning from Tay's Introduction. The Official Microsoft Blog. March 25. Available online: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed on 10 October 2025).

Lin, Chien-Te. 2023. All about the Human: A Buddhist Take on AI Ethics. Business Ethics, the Environment and Responsibility 32: 1113–22. [Google Scholar] [CrossRef] [Scilit]

Liu, Yunhao □□□ . 2025. Jushen zhinen: Rengong zhinen de xiayige langchao. In □□□□□□□□□□□□□□□ . Embodied Intelligence: The Next Wave of Artificial Intelligence. Beijing: Zhongxin Chubanshe. □□ : □□□□□ . ISBN 9787521772098. [Google Scholar]

Maheshwari, Anil K., Sharda S. Nandram, and Payal Kumar. 2025. Dharma Matrix: An Open Architecture for Ethical Decision-Making in the Age of AI. In AI and Consciousness in Organizations and Society. London: Palgrave Macmillan, pp. 39–62. [Google Scholar] [CrossRef] [Scilit]

Moëll, Birger, and Fredrik Sand Aronsson. 2025. Journaling with Large Language Models: A Novel UX Paradigm for AI-Driven Personal Health Management. Frontiers in Artificial Intelligence 8: 1567580. [Google Scholar] [CrossRef] [Scilit] [PubMed]

Mollick, Ethan. 2024. Co-Intelligence: Living and Working with AI. New York: Portfolio/Penguin. [Google Scholar]

Pasquale, Frank. 2020. New Laws of Robotics: Defending Human Expertise in the Age of AI. Cambridge, MA and London: The Belknap Press of Harvard University Press. [Google Scholar]

Phophichit, Nadnapang, Phuen Kittisobhano, Weerasak Abhinandavedi, Nuntawit Dhirabhaddo, and Nopadol Penprachoom. 2025. SabaiJai: A Buddhist AI Chatbot Innovation for Stress Resilience in Thailand's Working-Aged Population. Thammasat Review 28: 70–100. [Google Scholar] [CrossRef]

Rushiwowen. 2025. Available online: https://rushiwowen.co/ (accessed on 10 October 2025).

Shakya, Miroj. 2019. The Digital Sanskrit Buddhist Canon Project: Problems and Possibilities. In Digital Humanities and Buddhism. Edited by Daniel Veidlinger. Berlin and Boston: De Gruyter, pp. 111–26. [Google Scholar]

Shanahan, Murray. 2025. Palatable Conceptions of Disembodied Being. arXiv. [Google Scholar] [CrossRef] [Scilit]

Sheng, Kai (□□). 2018. Rengong zhineng shidai de renjian Fojiao yu weilai shehui yantaohui jiyao "'□□□□□□□□□□□□□□□□ ' □□□□ " Proceedings of the Symposium 'Humanistic Buddhism and Future Society in the Age of AI'. Fojiao Yanjiu □□□□ 2018: 312–19. [Google Scholar]

Sherwood, Harriet. 2016. Robot Monk to Spread Buddhist Wisdom to the Digital Generation. The Guardian. April 26. Available online: https://www.theguardian.com/world/2016/apr/26/robot-monk-to-spread-buddhist-wisdom-to-the-digital-generation (accessed on 10 October 2025).

Shneiderman, Ben. 2022. Human-Centered AI. Oxford: Oxford University Press. [Google Scholar] [CrossRef] [Scilit]

Tempone-Wiltshire, Julien. 2024. Seeking the Neural Correlates of Awakening. Journal of Consciousness Studies 31: 173–203. [Google Scholar] [CrossRef] [Scilit]

Torous, John, Jake Linardon, Simon B. Goldberg, Shufang Sun, Imogen Bell, Jennifer Nicholas, Lamiece Hassan, Yining Hua, Alyssa Milton, and Joseph Firth. 2025. The Evolving Field of Digital Mental Health: Current Evidence and Implementation Issues for Smartphone Apps, Generative Artificial Intelligence, and Virtual Reality. World Psychiatry 24: 156–74. [Google Scholar] [CrossRef] [Scilit] [PubMed]

Tsadra Foundation. 2025. AI Tools for Tibetan Buddhist Translation. Available online: https://www.tsadra.org/ai-tools-for-tibetan-buddhist-translation/ (accessed on 10 October 2025).

Veidlinger, Daniel. 2015. Introduction. In Buddhism, the Internet, and Digital Media: The Pixel in the Lotus. Edited by Gregory Price Grieve and Daniel Veidlinger. New York: Routledge, pp. 1–20. [Google Scholar] [CrossRef] [Scilit]

Veidlinger, Daniel. 2018. From Indra's Net to Internet: Communication, Technology, and the Evolution of Buddhism. Honolulu: University of Hawai'i Press. ISBN 9780824873400. [Google Scholar]

Veidlinger, Daniel, ed. 2019. Introduction. In Digital Humanities and Buddhism. Berlin and Boston: De Gruyter, pp. 1–24. ISBN 978-3-11-051836-8. [Google Scholar]

Wallach, Wendell, and Colin Allen. 2008. Moral Machines: Teaching Robots Right from Wrong. New York: Oxford University Press. [Google Scholar] [CrossRef] [Scilit]

Wei Hai □□ . 2005. Wuyun Xinlixue: Fojia Ziwo Juexing Ziwo Chaoyue de Xueshuo □□□□□□□□□□□□□□□□□ Psychology of the Five Aggregates: The Buddhist Doctrine of Self-Awakening and Self-Transcendence. Beijing: Zongjiao Wenhua Chubanshe □□□□□□ . ISBN 7801237218. [Google Scholar]

Wukong-Buddhist AI. 2024. Available online: https://buddhism-ai.com/ (accessed on 10 October 2025).

Yomsin, Onjira. 2025. SDGs and Buddhist Guideline for Coping with Artificial Intelligence (AI) Deepfake Technology Ethical Issues. Preprint/PDF, Mahachulalongkornrajavidyalaya University. (Public Full Text on ResearchGate). Available online: https://www.researchgate.net/publication/391277469_SDGs_and_Buddhist_Guideline_ (accessed on 10 October 2025).

Yuan, Bin □□ , and Jian Kang □□ , eds. 2010. Chanzong Da Cidian □□□□ . Wuhan: Chongwen Shuju, □□ : □□□□ , p. 171. [Google Scholar]

Yudkowsky, Eliezer. 2001. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. San Francisco: The Singularity Institute for Artificial Intelligence, June 15, Available online: https://intelligence.org/files/CFAI.pdf (accessed on 10 October 2025).

Zhang, Xuesong □□□ . 2024. Dui suowei "AI chengfo" wenti de bianxi—Cong Fojiao zhexue jiaodu fansi rengong zhineng □□□ "AI□□ "□□□□□ — □□□□□□□□□□□□ . A Critical Examination of the So-Called 'AI Attaining Buddhahood': Reflections on Artificial Intelligence from a Buddhist-Philosophical

Perspective. Fojiao Yanjiu □□□□ 2024: 328–34. [Google Scholar]