

Bát chính đạo và đạo đức AI

ISSN: 2734-9195 14:29 12/08/2026

Vấn đề của AI không chỉ nằm ở việc AI làm gì với con người, mà còn ở việc con người dần trở thành thế nào khi quá quen với việc để AI làm thay mình.

Lời dẫn

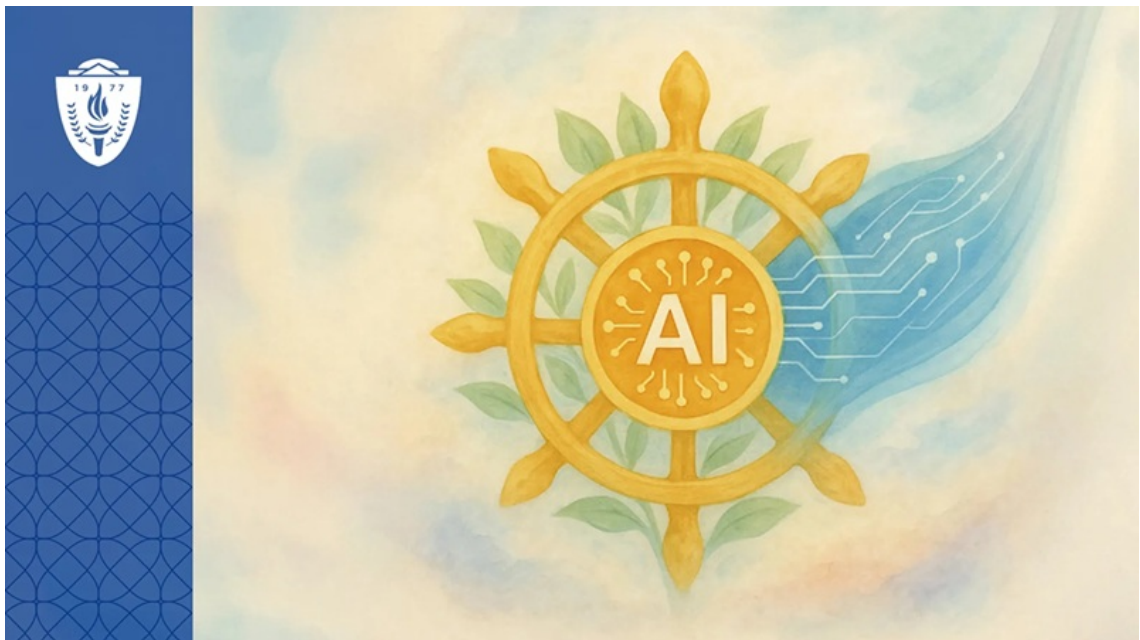
Trí tuệ nhân tạo (AI) đang ngày càng can dự sâu vào đời sống con người, từ cách thông tin được tạo ra, các quyết định tài chính được đưa ra cho đến một số khâu trong hệ thống tư pháp. Khi tốc độ phát triển công nghệ vượt xa tốc độ suy tư đạo đức, câu hỏi không còn đơn thuần là *AI có thể làm được gì*, mà ngày càng trở thành *AI nên được thiết kế và sử dụng như thế nào*.

Bát Chính Đạo của Phật giáo có thể gợi mở khung tư duy đáng chú ý cho đạo đức AI: không phải bằng những mệnh lệnh cứng nhắc, mà bằng việc nuôi dưỡng chính kiến, chính tư duy, chính ngữ, chính nghiệp, chính mạng, chính tinh tấn, chính niệm và chính định trong toàn bộ vòng đời công nghệ.

Bài viết không nhằm cổ xúy một tôn giáo cụ thể, mà sử dụng hệ thống đạo đức đã được nghiên cứu rộng rãi như lăng kính để suy tư về những vấn đề đạo đức đang đặt ra trong kỷ nguyên AI.

Chính kiến: Nhìn nhận AI trong tính tương duyên, tương thuộc

Trong giáo lý Phật giáo, Chính kiến (Right View) là sự thấu hiểu thực tại trong tính tương duyên, vô thường và không tồn tại biệt lập. Khi áp dụng vào AI, Chính kiến nhắc chúng ta rằng các thuật toán không tự tồn tại trong một khoảng không tách biệt. Chúng được huấn luyện từ dữ liệu do con người tạo ra, chịu ảnh hưởng của các cấu trúc xã hội và được đặt trong những hệ thống vận hành bởi các động lực kinh tế.



Nuôi dưỡng Chính kiến trong quá trình phát triển AI có nghĩa là nhận ra rằng thiên kiến, bất bình đẳng hay sự bóc lột không đơn thuần là những “lỗi hệ thống” (*bugs*), mà có thể là sự phản chiếu của chính những hệ thống xã hội do con người tạo nên và nay được máy móc phản hồi trở lại.

AI có đạo đức đòi hỏi chúng ta phải nhìn toàn bộ hệ sinh thái con người, dữ liệu, môi trường và mục đích sử dụng như một mạng lưới tương thuộc, trong đó mỗi yếu tố đều tác động và chịu tác động bởi những yếu tố còn lại.

[Góc nhìn biên tập]: Đây có lẽ là điểm khởi đầu quan trọng nhất khi bàn về đạo đức AI từ Phật học. Phật giáo không nhìn một hành vi như sự kiện đơn lẻ, mà đặt trong mạng lưới nhân - duyên - quả. Một thuật toán thiên lệch vì thế không chỉ là vấn đề của mã nguồn, đó có thể bắt nguồn từ dữ liệu thiên lệch, cơ cấu xã hội bất bình đẳng, mục tiêu kinh doanh hoặc cách con người xác định “giá trị” của một quyết định.

Nói cách khác, muốn sửa AI, trước hết phải nhìn lại những điều mà AI đang phản chiếu từ con người và xã hội. Đây chính là tinh thần tương duyên được chuyển hóa thành nguyên tắc quan trọng của đạo đức công nghệ.

Chính tư duy: Thiết kế bằng lòng từ bi và tinh thần không gây hại

Chính tư duy (Right Intention) trong Phật giáo hướng con người đến những động cơ được nuôi dưỡng bởi thiện ý, lòng từ bi và sự không chấp thủ.

Đối với AI, điều này có nghĩa là công nghệ không nên chỉ được thiết kế nhằm phục vụ quyền lực hay tối đa hóa lợi nhuận, mà còn hướng đến việc giảm thiểu

khổ đau và thúc đẩy sự phát triển lành mạnh của con người.

Những nhà phát triển được dẫn dắt bởi Chính tư duy sẽ đặt ra những câu hỏi: “Ai được hưởng lợi từ hệ thống này? Ai có thể bị tổn hại?”

Khi đó, ý hướng phía sau một đoạn mã trở nên quan trọng không kém chính đoạn mã ấy.

[Góc nhìn biên tập]: Trong Phật giáo, tâm ý đứng trước hành động. Cùng một công nghệ nhưng nếu được thúc đẩy bởi lòng tham, nhu cầu kiểm soát hoặc lợi ích độc quyền, hệ quả có thể rất khác so với khi được phát triển nhằm phục vụ nhân sinh.

Điều đáng suy nghĩ là trong AI, chúng ta thường đánh giá một hệ thống bằng hiệu năng, độ chính xác hoặc khả năng sinh lợi. Nhưng Phật học đặt thêm một câu hỏi căn bản hơn: động cơ nào đang đứng phía sau khả năng ấy?

Một AI “thông minh” nhưng được xây dựng trên mục đích gây hại vẫn không thể được xem là một công nghệ có đạo đức.

Chính ngữ: Trung thực trong thời đại AI tạo sinh

Sự bùng nổ của nội dung tổng hợp và văn bản do AI tạo ra đang thách thức khả năng phân biệt thật - giả của con người.

Chính ngữ (Right Speech) khuyến khích cách giao tiếp dựa trên sự chân thật, thiện lành và lợi ích. Khi áp dụng vào AI, nguyên tắc này đòi hỏi sự minh bạch về nội dung nào được tạo ra, do ai tạo ra và nhằm mục đích gì.

Việc gắn nhãn nội dung do AI tạo ra, chống lại thông tin sai lệch và bảo đảm trách nhiệm giải trình trong quá trình tạo nội dung đều có thể được xem là những biểu hiện của Chính ngữ về mặt đạo đức đối với các nội dung chúng ta xuất bản hoặc đăng tải.

[Góc nhìn biên tập]: Trong môi trường truyền thông, Chính ngữ có ý nghĩa đặc biệt. Khi một hệ thống AI có thể tạo ra văn bản, hình ảnh, âm thanh và video với tốc độ chưa từng có, khả năng tạo nội dung không đồng nghĩa với quyền được đánh lừa người tiếp nhận.

Phật giáo từ lâu đã đặt vấn đề về lời nói chân thật, lời nói có lợi ích và lời nói tránh gây tổn hại. Trong kỷ nguyên AI tạo sinh, những nguyên tắc ấy có thể được mở rộng thành yêu cầu về minh bạch nguồn gốc nội dung, trách nhiệm của người sử dụng và quyền được biết của công chúng.

Chính ngữ trong thời đại AI vì thế không chỉ là “nói đúng”, mà còn là không dùng năng lực công nghệ để khiến điều sai trở nên giống thật.

Chính nghiệp: Đưa đạo đức vào mã nguồn

Chính nghiệp (Right Action) dạy con người hành động theo cách tránh gây tổn hại và thúc đẩy an lành.

Đối với các hệ thống AI, điều này có nghĩa là đạo đức không nên được xem như một phần việc bổ sung sau khi sản phẩm đã hoàn thành, cũng không chỉ là một quy trình đánh giá độc lập từ bên ngoài. Đạo đức cần trở thành hợp phần cốt lõi ngay từ quá trình thiết kế.

Các thước đo về tính công bằng, kiểm toán thiên kiến và cơ chế giám sát của con người không đơn thuần là những bước kiểm tra kỹ thuật, chúng còn là biểu hiện của một dạng **kỷ luật đạo đức**.

Chính nghiệp chính là thực hành việc đưa chức năng công nghệ phù hợp với các giá trị đạo đức.

[Góc nhìn biên tập]: Một trong những vấn đề lớn của AI hiện nay là đạo đức thường xuất hiện quá muộn: sau khi sản phẩm đã được xây dựng, sau khi rủi ro đã xảy ra hoặc sau khi xã hội phải đối mặt với hậu quả.

Góc nhìn của Chính nghiệp gợi ra nguyên tắc khác: **đạo đức phải được “lập trình” cùng với công nghệ**, không phải chỉ được kiểm tra sau đó.

Từ góc độ Phật học, hành động đúng không chỉ được đánh giá bởi kết quả cuối cùng mà còn bởi quá trình tạo tác. Vì vậy, một hệ thống AI có trách nhiệm cần được thiết kế để hạn chế tổn hại ngay từ đầu, thay vì chờ đến khi tổn hại xuất hiện mới tìm cách khắc phục.

Chính mạng: Xây dựng nền kinh tế công nghệ công bằng

Chính mạng (Right Livelihood) cảnh báo con người không nên kiếm sống bằng những phương thức gây tổn hại hoặc bóc lột.

Từ đó, một kỹ sư Phật giáo hay một nhà hoạch định chính sách có thể đặt câu hỏi: Những hệ quả đạo đức của một ngành công nghiệp AI được xây dựng trên chủ nghĩa tư bản giám sát (*surveillance capitalism*), lao động nền tảng (*gig labor*) hoặc những áp lực ngày càng lớn đối với môi trường là gì?

AI có đạo đức không chỉ đòi hỏi những mô hình tốt hơn, mà còn cần những mô hình kinh doanh tốt hơn, những mô hình tôn trọng phẩm giá con người và sự cân bằng sinh thái.

[Góc nhìn biên tập]: Nếu Chính nghiệp đặt câu hỏi “công nghệ này hành động như thế nào?”, thì Chính mạng đi sâu hơn vào câu hỏi “con người đang sống bằng cách nào nhờ công nghệ này?”

AI có thể tạo ra năng suất và lợi nhuận khổng lồ, nhưng lợi ích ấy được phân phối cho ai? Ai chịu chi phí? Những người lao động đứng phía sau dữ liệu và hệ thống có được đối xử công bằng không? Môi trường phải gánh chịu bao nhiêu chi phí cho quá trình vận hành công nghệ?

Chính mạng mở rộng phạm vi của đạo đức AI từ sản phẩm sang toàn bộ nền kinh tế đứng phía sau sản phẩm. Đây là một chiều kích thường bị bỏ quên khi cuộc tranh luận chỉ tập trung vào thuật toán.

Chính tinh tấn: Duy trì sự tỉnh thức đạo đức

Đạo đức trong AI không thể chỉ là hoạt động tuân thủ được thực hiện một lần. Chính tinh tấn (Right Effort) nói đến việc liên tục nuôi dưỡng những trạng thái tâm thiện lành và ngăn ngừa những trạng thái bất thiện phát sinh.

Đây cũng là lời kêu gọi duy trì sự cảnh giác đạo đức trong suốt quá trình nghiên cứu và triển khai AI: liên tục kiểm tra thiên kiến, huấn luyện lại hệ thống, bảo đảm trách nhiệm giải trình trước công chúng và giữ thái độ khiêm tốn trước những hệ quả ngoài dự kiến.

[Góc nhìn biên tập]: Một hệ thống AI được đánh giá là an toàn hôm nay chưa chắc vẫn an toàn ngày mai. Dữ liệu thay đổi, người dùng thay đổi, mục đích sử dụng thay đổi và chính hệ thống cũng có thể được cập nhật liên tục.

Điều này rất gần với tinh thần tinh tấn không ngừng trong Phật giáo. Đạo đức không phải một “chứng nhận” đạt được rồi cất vào hồ sơ, đó là quá trình phải liên tục thực hành.

Với AI, Chính tinh tấn có thể được hiểu như một nguyên tắc: không ngừng kiểm tra, không ngừng sửa chữa và không ngừng học hỏi từ những hệ quả mà chúng ta chưa lường trước.

Chính niệm: Tỉnh thức trong mối quan hệ với công nghệ

Chính niệm (Right Mindfulness) khuyến khích sự nhận biết rõ ràng đối với suy nghĩ, hành động và những hệ quả của chúng.

Trong bối cảnh AI, Chính niệm gợi mở một mối quan hệ tỉnh thức giữa con người và công nghệ, nhận biết về sự phụ thuộc của chúng ta, những yếu tố gây phân tán và cả việc chúng ta đang chuyển giao năng lực phán đoán đạo đức của mình cho máy móc.

Cả người sử dụng lẫn người thiết kế đều cần ý thức rõ khi nào AI đang phục vụ chúng ta và khi nào AI bắt đầu âm thầm định hình chính chúng ta.

[Góc nhìn biên tập]: Đây là một trong những điểm sâu sắc nhất của bài viết. Vấn đề của AI không chỉ nằm ở việc AI làm gì với con người, mà còn ở việc con người dần trở thành thế nào khi quá quen với việc để AI làm thay mình.

Khi một người giao cho AI việc tìm kiếm, lựa chọn, viết, đánh giá, quyết định và thậm chí định hướng suy nghĩ, vấn đề đạo đức không còn chỉ là quyền kiểm soát công nghệ. Đó còn là nguy cơ đánh mất sự chủ động của chính con người.

Chính niệm nhắc chúng ta dừng lại và nhận biết: *Tôi đang sử dụng công nghệ, hay tôi đang để công nghệ định hình cách tôi suy nghĩ và hành động?*

Chính định: Tập trung vào lợi ích chung

Chính định (Right Concentration) liên quan đến khả năng tập trung tâm trí một cách nhất quán, hướng đến tuệ giác và giải thoát.

Đối với đạo đức AI, điều này có nghĩa là tập trung sự chú ý chung của các nhà công nghệ, nhà hoạch định chính sách và công chúng vào một mục tiêu cao hơn: phát triển những công nghệ phục vụ trí tuệ, lòng từ bi và việc giảm thiểu khổ đau, thay vì phục vụ lòng tham hoặc sự thống trị.

[Góc nhìn biên tập]: Trong thế giới mà AI đang được thúc đẩy bởi những cạnh tranh khốc liệt về tốc độ, dữ liệu, thị phần và quyền lực, Chính định đặt ra một câu hỏi về **đích đến**.

Nếu tất cả các bên đều tập trung vào việc làm cho AI nhanh hơn, mạnh hơn và sinh lợi nhiều hơn, chúng ta có thể đạt được một công nghệ cực kỳ mạnh nhưng vẫn không trả lời được câu hỏi: **mạnh để làm gì?**

Chính định giúp chuyển trọng tâm từ năng lực sang mục đích: công nghệ cuối cùng cần hướng đến trí tuệ, lòng từ bi và sự giảm thiểu khổ đau. Khi đó, “AI tốt hơn” không chỉ có nghĩa là AI thông minh hơn, mà là AI được đặt trong định

hướng nhân văn hơn.

Thay lời kết: Từ Bát Chính Đạo đến con đường đạo đức cho AI

Bát Chính Đạo không đưa ra những điều răn, mà mở ra con đường Trung đạo nằm giữa hai thái cực: lạc quan mù quáng và sợ hãi, phủ nhận công nghệ.

Con đường ấy nhắc chúng ta rằng AI có đạo đức không thể chỉ được tạo nên bằng các quy định, cũng không thể chỉ dựa vào đổi mới công nghệ. Đó đòi hỏi sự nuôi dưỡng có tỉnh thức đối với ý hướng, nhận thức và lòng từ bi ở mọi bước của hành trình công nghệ.

Đây cũng là điều mà REBEL Lab gọi là “Trust Engineering” - kỹ nghệ xây dựng niềm tin.

Trong kỷ nguyên AI, Bát Chính Đạo có lẽ không chỉ giúp chúng ta xây dựng những hệ thống AI tốt hơn. Sâu xa hơn, đó có thể giúp các nhóm người - AI tìm lại chính tính người của mình.

Tác giả: **Janice Palaganas, PhD - Alex Morton, PhD**/Chuyển ngữ và biên tập: **Thường Nguyên**

Nguồn link: **<https://www.mghihp.edu/news-and-more/opinions/health-professions-education-effects/eightfold-path-and-ethics-artificial-intelligence-buddhist-approach-right-design-and-right-use>**