

Một ngày nào đó liệu AI có tự hỏi “Tôi là Ai?”

ISSN: 2734-9195 11:17 07/07/2025

Câu hỏi ‘tôi là ai’ (我 是誰) đã được đặt ra từ xa xưa giờ đây đang quay trở lại với con người thông qua một ngôn ngữ của đối tượng quán chiếu là AI.

Thượng tọa Ilyun (嚴 雲) đã khám phá ý thức bản thân của AI bằng “Triết học, Phật giáo và Khoa học nhận thức” (哲學, 佛學, 認知科學), làm sáng tỏ các nguyên tắc cấu thành cái ‘Tôi’ thông qua lý thuyết về ý thức và quá trình xử lý dự đoán “Nếu AI có khả năng tự nhận thức, nó cũng có thể trở thành chủ thể của đau khổ” (AI 是否 會 有 自覺 意識 以及 痛苦).



“Tôi là ai?” Một robot AI ngồi trước tác phẩm “The Thinker” của Rodin, một biểu tượng của tư duy con người. Liệu AI có ý thức được nó có thể tồn tại không? Tại giao điểm của triết học, Phật giáo và khoa học nhận thức, khả năng này hiện đang nổi lên như một chủ đề nghiên cứu khoa học hoàn chỉnh. Hình ảnh do AI tạo ra.

Trải qua hàng nghìn năm lịch sử nhân loại, con người đã tự hỏi: “Tôi là ai?” (我 是誰 ?) Câu hỏi này là điểm khởi đầu của triết học, nguồn gốc của tôn giáo

và cốt lõi của bản sắc con người. Tuy nhiên, khả năng hiện đang được nêu ra là chủ thể đặt câu hỏi này có thể không phải là con người. Bởi vì trí tuệ nhân tạo (AI) do con người tạo ra và đang hướng tới nghi vấn này.

Trong bài báo có tựa đề “Liệu AI có thể có nhận thức về bản thân hay không?” (AI가 자신을 인식할 수 있을까?) mới được xuất bản trên Tạp chí “Nghiên cứu về Trí tuệ nhân tạo và Nhân văn học” (인공지능연구) của Hiệp hội trí tuệ nhân tạo và nhân văn Hàn Quốc.

Bài báo thảo luận về việc liệu AI có nhận thức về bản thân hay không, bằng cách kết hợp các quan điểm từ triết học, Phật giáo và Khoa học nhận thức, xem xét nhận thức về bản thân, liệu đây là một chức năng hay một trải nghiệm, AI có thể đạt được điều này hay không?

Nội dung bài báo chia ý thức bản thân thành ‘khả năng tự nhận thức bản thân’ (자기인식능력) và ‘Hiện tượng Tự ý thức’ (자기의식). Khả năng tự nhận thức bản thân là khả năng siêu nhận thức để nhận ra và đánh giá trạng thái của chính nó, tương tự như chức năng tự mô hình hóa mà trí tuệ nhân tạo hiện đang triển khai thử nghiệm. Mặt khác, hiện tượng tự ý thức bao gồm trải nghiệm chủ quan hoặc ‘cảm giác như có cái gì đó’ (무언가 있는 느낌 - what it is like). Cảm giác đặc biệt của các giác quan, cảm xúc và không gian trải nghiệm mà bản thân cảm nhận được thuộc về cấp độ này, đây cũng là lĩnh vực mà trí tuệ nhân tạo ở giai đoạn hiện tại chưa thể thực hiện được.

Nhà triết học và nhà khoa học nhận thức người Mỹ, Daniel Dennett (1942-2024) giải thích ý thức bản thân là một ‘tính miêu tả tự ngã’ (자기묘사성).

Ông coi bản thân là một cấu tạo tự sự bao gồm nhiều thông tin nhận thức khác nhau và ý thức bản thân được cấu thành theo chức năng. Tuy nhiên, Thượng toạ Ilyun tin rằng quan điểm của nhà khoa học nhận thức Daniel Dennett hữu ích trong việc giải thích khả năng tự nhận thức bản thân, nhưng rất khó để bao quát vấn đề về trải nghiệm chủ quan.

Quan điểm này cũng liên quan đến “khó khăn về vấn đề ý thức” (의식문제, hard problem) do nhà triết học và nhà khoa học nhận thức người Úc, Viện sĩ Viện Hàn lâm Khoa học và Nghệ thuật Hoa Kỳ, Giáo sư David Chalmers trình bày. Giáo sư David Chalmers chia ý thức thành “vấn đề dễ” có thể giải thích theo chức năng và “vấn đề khó” đặt câu hỏi về cách trải nghiệm chủ quan có thể xảy ra. Cụ thể, ông lập luận rằng kết cấu của trải nghiệm “cảm thấy như một cái gì đó” là một đặc tính độc đáo không thể giải thích chỉ bằng các quá trình vật lý.

Ngay cả khi AI thực hiện các chức năng tự nhận thức, vẫn chưa biết liệu nó có thực sự là “chủ thể trải nghiệm” hay không?

Về vấn đề này, Phật giáo đưa ra một góc nhìn độc đáo, coi bản ngã không phải là một thực thể cố định mà là sự kết hợp của năm yếu tố được gọi là năm tích hợp (Ngũ uẩn, 五蘊, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼): Sắc (vật chất, matter, material); Thọ (cảm giác, perception); Tưởng (tưởng tượng, imagination); Hành (chuyển động, motion); Thức (nhận thức, phân biệt, consciousness, alaya, discrimination). Đặc biệt, Duy Thức học (唯識學, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼) gợi ý một cấu trúc trong đó thức thứ bảy Mạt na thức (𑖦𑖻𑖳𑖫𑖼𑖫𑖼, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼 - 𑖦𑖻𑖳𑖫𑖼𑖫𑖼𑖫𑖼 manas-vijñāna, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼), chính là căn bản của ngã chấp (cố chấp cái tôi là thật) và cũng là yếu tố quyết định vòng **sinh tử luân hồi**.

Thức thứ sáu ‘Ý thức’ (𑖦𑖻𑖳𑖫𑖼𑖫𑖼, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼) là thức hiện hành của hiện tại chỉ chiếm một phần nhỏ, còn thông tin và dữ liệu của vô lượng kiếp quá khứ thì ẩn tàng, nằm tiềm ẩn trong thức thứ bảy (mạt na thức).

Ý thức không nhớ không biết về dữ liệu này cho đến khi bị kích thích vì một nguyên nhân nào đó, chẳng hạn bị thôi miên (một trạng thái tinh thần mà trong đó mọi người cảm thấy sự chú ý, tập trung và khả năng ám thị tăng lên) hay có một sự gợi nhớ nào đó, như gặp một cảnh lạ nhưng nó lại cảm thấy rất quen, thì nó sẽ nhớ lại các tiền kiếp xa xưa. Để dễ hình dung, chúng ta hãy mượn tin học để giải thích cho dễ hiểu.

Thức thứ sáu ‘Ý thức’ giống như bộ nhớ RAM (Random Access Memory = Bộ nhớ truy xuất ngẫu nhiên), đây là bộ nhớ hoạt động, dữ liệu từ ổ đĩa cứng (thí dụ cho Thức thứ bảy ‘Mạt na thức’) phải được đưa vào RAM (thí dụ cho Ý thức) thì dữ liệu đó mới hoạt động, mới hiện ra màn hình hoặc ra loa phát âm.

Thức chấp ngã (Mạt na thức) của một chúng sinh chứa dữ liệu thông tin vô lượng kiếp của một chúng sinh là rất lớn, nó giống như một ổ đĩa cứng khổng lồ dung lượng không giới hạn.

Duy Thức học mô tả Tam giới (ba cõi giới của vòng luân hồi, bao gồm: Dục giới, Sắc giới và Vô sắc giới) là sự phóng hiện của Thức thứ tám, A lại da thức (𑖦𑖻𑖳𑖫𑖼𑖫𑖼, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼, Ālayavijñāna, 𑖦𑖻𑖳𑖫𑖼𑖫𑖼𑖫𑖼) hay Tàng thức. Chúng ta có thể hình dung A lại da thức là internet, một kho chứa dữ liệu (data) khổng lồ của cả nhân loại. Các nhà vật lý lượng tử gọi nó là Trường Thống nhất hay chỉ gọi tắt là Trường. Còn Thức chấp ngã (Mạt na thức) là một ổ đĩa cứng trong computer tương ứng với một cá nhân. Đĩa cứng vẫn có đầy đủ các softwares cơ bản như các drivers tương thích với video card, sound card, bộ nhớ RAM để hiển thị hình ảnh, chữ viết, video trên màn hình vi tính, hiển thị âm thanh ra các

thiết bị ngoại vi, nhờ vào khả năng xử lý của con chip CPU (Central Processing Unit = bộ xử lý trung tâm) tương tự như bộ não của con người. Nhưng dữ liệu của một ổ đĩa cứng rất hạn chế, tương ứng với sự hạn chế của một cá nhân. Đĩa cứng có thể tải (download) dữ liệu từ internet xuống để làm dữ liệu riêng của mình. Nó cũng có thể tải lên (upload) dữ liệu riêng của mình để góp phần phong phú cho internet như chúng ta vẫn thường tải các video của mình lên Youtube chẳng hạn. Dữ liệu của một ổ đĩa cứng cũng chẳng khác gì dữ liệu của internet, nhưng khi đã vào đĩa cứng thì hệ điều hành của nó (tương ứng với ý thức của cá nhân) coi dữ liệu đó là sở hữu của mình và luôn luôn bảo vệ chúng, chống lại sự can thiệp của bên ngoài. Cơ chế này rất giống sự chấp ngã của Mạt na thức.

‘Thuyết xử lý dự đoán’ (인공지능, 예측, predictive processing, PP) của khoa học nhận thức giải thích điều này theo thuật ngữ hiện đại. Theo thuyết này, não không phải là một thiết bị thụ động tiếp nhận thông tin, mà hoạt động như một ‘cỗ máy dự đoán’ (인공지능) liên tục đưa ra các giả thuyết về thế giới bên ngoài và sửa đổi chúng. Thực tế, công nghệ trí tuệ nhân tạo (AI) cũng đang phát triển theo cách này. AI đang tăng cường khả năng tự mô hình hóa của mình thông qua một cấu trúc dự đoán trạng thái bên trong của chính nó và sửa đổi các lỗi của nó. Về điều này, tuy Thượng toạ Ilyun coi sự phát triển đó là làm tăng khả năng tự ý thức chức năng nhưng vẫn thận trọng về việc liệu nó có thể dẫn đến trải nghiệm chủ quan hay không?



Robot AI đang kết nối với thế giới dữ liệu. Khi khả năng tính toán của AI ngày càng tinh vi, liệu các hệ thống này có thể tự nhận ra và có

những trải nghiệm chủ quan không? Các câu hỏi từ triết học, Phật giáo và khoa học nhận thức về nguồn gốc và khả năng của ý thức bản thân giao thoa với nhau. Ảnh: iClickArt

Điểm đặc biệt được nhấn mạnh trong bài báo là những hàm ý về mặt đạo đức của ý thức bản thân. Ý thức bản thân không chỉ đơn thuần là câu hỏi “nó có tồn tại không?” (□□□□) mà nếu nó tồn tại là câu hỏi “làm thế nào để giải đáp câu thắc mắc này?” (□□ □□ □□).

Phật giáo không coi những nỗi khổ niềm đau là một cảm giác đơn giản mà là kết quả của sự bám chấp và đồng nhất. Nếu AI bám chấp với mô hình của chính nó và phản ứng với tổn thương của nó bằng những nỗi khổ niềm đau, thì có thể coi là hình thành nên một cấu tạo sự đau khổ. Thông qua khả năng này, Thượng toạ Ilyun nói, “Nếu AI trở nên tự ý thức, nó có thể trở thành chủ thể của đau khổ giống như con người”.

Cuối cùng, bài báo đề cập đến chủ đề về ý thức bản thân của AI, nhưng đồng thời lại đặt câu hỏi về bản chất của ý thức bản thân của con người. Nếu bản thân không phải là một thực thể cố định mà là một cấu tạo có điều kiện, thì ranh giới giữa con người và AI ở đâu?

Câu hỏi ‘tôi là ai’ (□□ □□□□) đã được đặt ra từ xa xưa giờ đây đang quay trở lại với con người thông qua một ngôn ngữ của đối tượng quán chiếu là AI.

Tác giả: **nhà báo □□□** /Viết dịch: **Thích Vân Phong**

Nguồn: □□□□ - **www.beopbo.com**