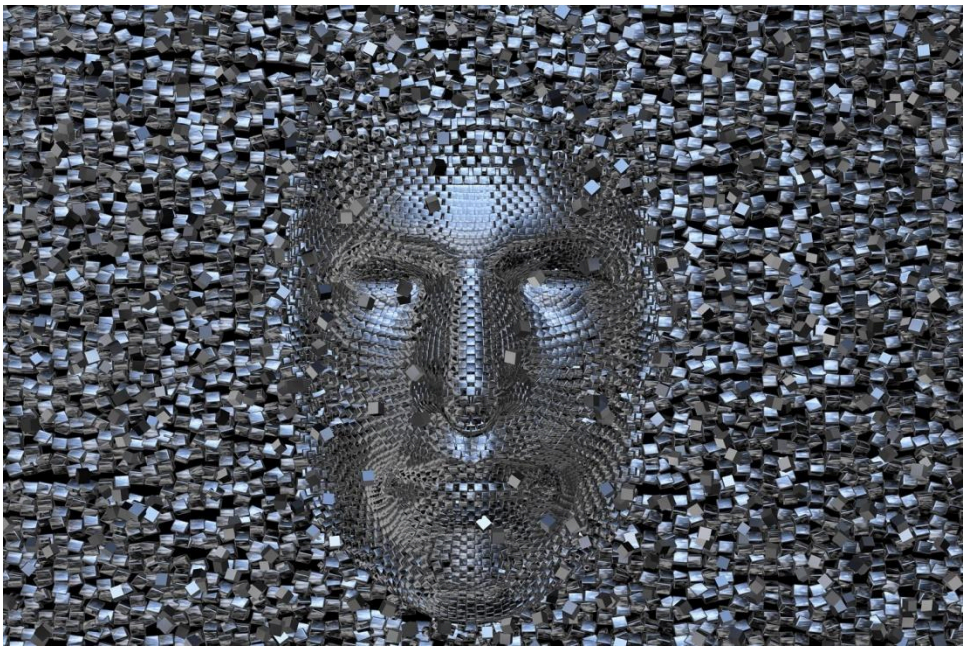
Liệu máy có biết ta biết những gì mà nó biết không?

Nguyễn Việt Anh dịch

Nguồn: Can a Machine Know That We Know What It Knows?, *Nytimes*, March 27, 2023.

21/5/2023



Một số nhà nghiên cứu cho rằng các phần mềm chatbot đã phát triển thuyết tâm trí. Nhưng điều đó có khiến cho thuyết tâm trí của chúng ta trở nên hoang dại không?

Ngày 27 tháng 3 năm 2023

Các nhà khoa học nhận thức đã khám phá ra các cách để kiểm tra những loại năng lực tinh thần mà các mô hình ngôn ngữ lớn như ChatGPT sở hữu và không sở hữu. Nguồn: vrvr/Alamy

Đọc tâm trí là việc phổ biến ở con người chúng ta. Không phải theo cách các nhà tâm linh [psychics] tuyên bố sẽ làm điều đó, mà theo cách tiếp cận với những luồng ý thức ấm áp lấp đầy kinh nghiệm của mỗi cá nhân, hoặc theo cách các nhà tâm thức học [mentalist] tuyên bố sẽ làm điều đó, bằng cách kéo một suy nghĩ ra khỏi đầu bạn theo mong muốn. Việc đọc tâm trí hàng ngày thì tinh tế hơn:

chúng ta ghi nhận khuôn mặt và chuyển động của mọi người, lắng nghe lời ăn tiếng nói của họ rồi quyết định hoặc cảm nhận được là điều gì có thể diễn ra trong đầu họ.

Tâm lý trực giác như vậy — tức khả năng gán trạng thái tinh thần của người khác cho trạng thái tinh thần của chúng ta — được các nhà tâm lý học gọi là thuyết tâm trí [theory of mind], và sự vắng mặt hoặc khiếm khuyết của nó có liên quan tới chứng tự kỷ, chứng tâm thần phân liệt và các rối loạn phát triển khác. Thuyết tâm trí giúp chúng ta giao tiếp với và hiểu được người khác; nó cho phép chúng ta thưởng thức văn chương và phim ảnh, chơi trò chơi và hiểu được các môi trường xã hội quanh ta. Trên nhiều phương diện, năng lực này là một phần thiết yếu của con người.

Điều gì sẽ xảy ra nếu máy cũng có thể đọc được tâm trí?

Gần đây, Michal Kosinski, một nhà tâm lý học tại Trường Kinh doanh Sau đại học Stanford, đưa ra lập luận rằng: các mô hình ngôn ngữ lớn như ChatGPT và GPT-4 của hãng OpenAI — các cỗ máy dự đoán từ ngữ tiếp theo được đào tạo trên một lượng lớn văn bản từ mạng internet — đã phát triển thuyết tâm trí. Các nghiên cứu của ông tuy chưa được bình duyệt, song chúng đã thúc đẩy sự suy xét kỹ lưỡng và trò chuyện với các nhà khoa học nhận thức, những người đang cố gắng trả lời câu hỏi thường gặp hiện nay — liệu ChatGPT có thể làm được *điều này* hay không? — và đưa nó vào lĩnh vực nghiên cứu khoa học vững chắc hơn. Các mô hình này có những năng lực gì, và làm thế nào mà chúng có thể thay đổi sự am hiểu của chúng ta về tâm trí của chính mình?

“Các nhà tâm lý học sẽ không chấp nhận bất kỳ tuyên bố nào về những năng lực của trẻ nhỏ chỉ dựa trên các giai thoại về những tương tác của bạn với chúng, đó là điều dường như đang diễn ra với ChatGPT”, Alison Gopnik, một nhà tâm lý học tại Đại học California, Berkeley và cũng là một trong những nhà nghiên cứu đầu tiên xem xét thuyết tâm trí vào thập niên 1980 cho biết. “Bạn phải làm các trắc nghiệm khá cẩn thận và nghiêm ngặt.”

Nghiên cứu trước đây của Tiến sĩ Kosinski cho thấy các mạng nơ-ron được đào tạo để phân tích các đặc điểm trên khuôn mặt như dáng mũi, góc đầu và biểu cảm có thể dự đoán được các quan điểm về chính trị và khuynh hướng tình dục của mọi người với độ chính xác đáng kinh ngạc (khoảng 72% cho trường hợp đầu tiên và khoảng 80% cho trường hợp thứ hai). Công trình gần đây của ông về các mô

hình ngôn ngữ lớn sử dụng các trắc nghiệm thuyết tâm trí cổ điển để đo lường khả năng trẻ em trong việc gán những niềm tin sai lầm cho người khác.

Một ví dụ nổi tiếng là trắc nghiệm Sally-Anne, trong đó một cô bé, Anne, di chuyển một viên bi từ giỏ sang hộp khi một cô bé khác, Sally, không để ý. Các nhà nghiên cứu khẳng định, để biết Sally sẽ tìm viên bi ở đâu, người xem sẽ phải vận dụng thuyết tâm trí, lập luận về bằng chứng cảm nhận [perceptual evidence] và sự hình thành niềm tin của Sally: Sally không nhìn thấy Anne di chuyển viên bi vào hộp, vì vậy cô bé vẫn tin rằng đó là nơi cô bé để nó lần cuối, trong giỏ.

Tiến sĩ Kosinski trình bày 10 mô hình ngôn ngữ lớn với 40 biến thể đặc biệt của các trắc nghiệm này về thuyết tâm trí — mô tả các tình huống như trắc nghiệm Sally-Anne, trong đó một người (Sally) hình thành một niềm tin sai lầm. Sau đó, ông đặt câu hỏi cho các mô hình này về những tình huống đó, khuyến khích chúng xem liệu chúng có gán những niềm tin sai lầm cho các nhân vật liên quan và dự đoán chính xác hành vi của họ hay không. Ông nhận thấy rằng GPT-3.5, phát hành vào tháng 11 năm 2022, vượt qua 90% thời gian và GPT-4, phát hành vào tháng 3 năm 2023, vượt qua 95% thời gian.

Chúng ta kết luận được điều gì? Máy có thuyết tâm trí.

Michal Kosinski, một nhà tâm lý học tại Trường Kinh doanh Sau đại học Stanford, cho rằng các mô hình ngôn ngữ lớn đã phát triển thuyết tâm trí. Nhiều học giả không đồng ý. Nguồn: Christie Hemm Klok gửi cho Thời báo New York

Nhưng ngay sau khi những kết quả này được công bố, Tomer Ullman, một nhà tâm lý học tại Đại học Harvard, phản hồi bằng một loạt các thí nghiệm của riêng mình, cho thấy rằng những điều chỉnh nhỏ trong lời nhắc hỏi có thể thay đổi hoàn toàn câu trả lời do những mô hình ngôn ngữ lớn tinh vi nhất tạo nên. Nếu một công ten nơ được mô tả là trong suốt, thì máy sẽ không thể suy luận rằng ai đó có thể nhìn vào bên trong nó. Máy gặp khó khăn trong việc tính tới phát biểu của mọi người trong những tình huống này và đôi khi không thể phân biệt được đâu là một vật thể nằm trong công ten nơ và đâu là vật thể nằm trên đó.

Maarten Sap, một nhà khoa học máy tính tại Đại học Carnegie Mellon, đưa hơn 1.000 trắc nghiệm thuyết tâm trí vào các mô hình ngôn ngữ lớn và phát hiện ra rằng các mô hình có tính biến đổi [transformer] tiên tiến nhất, như ChatGPT và GPT-4, chỉ vượt qua khoảng 70% thời gian. (Nói cách khác, chúng đã thành công 70% trong việc gán những niềm tin sai lầm cho những người được mô tả trong các tình huống kiểm tra bằng trắc nghiệm.) Sự khác biệt giữa dữ liệu của ông và của Tiến sĩ Kosinski có thể dẫn tới những khác biệt trong quá trình kiểm tra, song Tiến sĩ Sap nói rằng thậm chí việc vượt qua 95% thời gian sẽ không phải là bằng chứng của thuyết tâm trí đích thực. Ông cho hay, máy thường thất bại theo một khuôn mẫu, không thể tham gia vào suy luận trừu tượng và thường tạo ra "những tương quan giả".

Tiến sĩ Ullman lưu ý rằng các nhà nghiên cứu về máy học đã phải vật lộn trong vài thập kỷ qua để nắm bắt được tính linh hoạt của tri thức con người trong các mô hình máy tính. Ông nói, khó khăn này từng là một "mặt tối bị che lấp", đằng sau mọi đổi mới thú vị. Các nhà nghiên cứu đã chỉ ra rằng các mô hình ngôn ngữ thường đưa ra câu trả lời sai hoặc chẳng liên quan khi trước đó đã được cung cấp thông tin không cần thiết; một số chương trình chatbot bị các cuộc thảo luận xoay quanh giả thiết về việc những con chim biết nói làm chúng rối lên tới mức cuối cùng chúng tuyên bố rằng những con chim có thể nói được. Vì lý luận của chúng nhạy cảm với những thay đổi nhỏ bé trong các đầu vào của chúng, nên các nhà khoa học đã gọi tri thức của những cỗ máy này là tri thức "mong manh".

Tiến sĩ Gopnik so sánh thuyết tâm trí của các mô hình ngôn ngữ lớn với sự am hiểu của chính bà về thuyết tương đối rộng. “Tôi đã đọc đủ để biết những từ đó là gì,” bà nói. “Nhưng nếu bạn yêu cầu tôi đưa ra một tiên đoán mới hoặc cho bạn biết lý thuyết của Einstein cho chúng ta biết điều gì về một hiện tượng mới, thì tôi sẽ ú ớ vì tôi thực sự không có thuyết tâm trí trong đầu.” Ngược lại, bà nói, thuyết tâm trí của con người được liên kết với các cơ chế lý luận khác về thường kiến [common-sense]; nó đứng vững khi đối mặt với sự suy xét kỹ lưỡng.

Nhìn chung, công trình của Tiến sĩ Kosinski và những phản hồi đối với nó phù hợp với cuộc tranh luận về việc liệu những năng lực của các cỗ máy này có thể so sánh với những năng lực của con người hay không — một cuộc tranh luận khiến cho các nhà nghiên cứu làm việc trong lĩnh vực xử lý ngôn ngữ tự nhiên bị chia rẽ. Những cỗ máy này là những con vẹt ngẫu nhiên, hay trí thông minh của người ngoài hành tinh, hay những kẻ lừa đảo gian xảo? Một cuộc khảo sát vào năm 2022 về lĩnh vực này cho thấy, trong số 480 nhà nghiên cứu trả lời, 51% tin rằng các mô hình ngôn ngữ lớn cuối cùng có thể “hiểu ngôn ngữ tự nhiên theo một nghĩa nào đó” và 49% tin rằng chúng không thể hiểu.

Tuy không coi thường khả năng hiểu của máy hoặc thuyết tâm trí của máy, song Tiến sĩ Ullman cảnh giác với việc gán những năng lực của con người cho những thứ phi con người. Ông lưu ý tới một nghiên cứu nổi tiếng vào năm 1944 của Fritz Heider và Marianne Simmel, trong đó những người tham gia được xem một bộ phim hoạt hình về hai hình tam giác và một hình tròn tương tác với nhau. Khi các đối tượng được yêu cầu viết ra những gì diễn ra trong phim, thì gần như tất cả họ đều mô tả các hình dạng đó là con người.

“Những người yêu nhau trong thế giới hai chiều, chắc chắn rồi; hình tam giác nhỏ số hai và hình tròn ngọt ngào”, một người tham gia viết. “Triangle-one (sau đây gọi là nhân vật phản diện) do thám người tình trẻ. Ah!”

Việc giải thích hành vi của con người bằng cách nói về những niềm tin, khát khao, ý định và suy nghĩ là lẽ tự nhiên và thường được xã hội yêu cầu. Xu hướng này là tâm điểm của con người chúng ta — tâm điểm tới mức đôi khi chúng ta cố gắng đọc tâm trí của những thứ không có tâm trí, ít nhất là không có tâm trí như của riêng con người chúng ta.

Tác giả

Oliver Whang

Oliver Whang (@oliverwhang21) là ký giả của tờ The Times, tập trung vào các lĩnh vực khoa học và sức khỏe.

Ấn bản in của bài báo này phát hành vào ngày 28 tháng 3 năm 2023, Phần A, Trang 1 của ấn bản New York với nhan đề: Human Power to Process Cues May Be Next Frontier for Bots [Sức mạnh con người để xử lý tín hiệu có thể là ranh giới tiếp theo cho những chương trình bot].

http://www.phantichkinhte123.com/2023/05