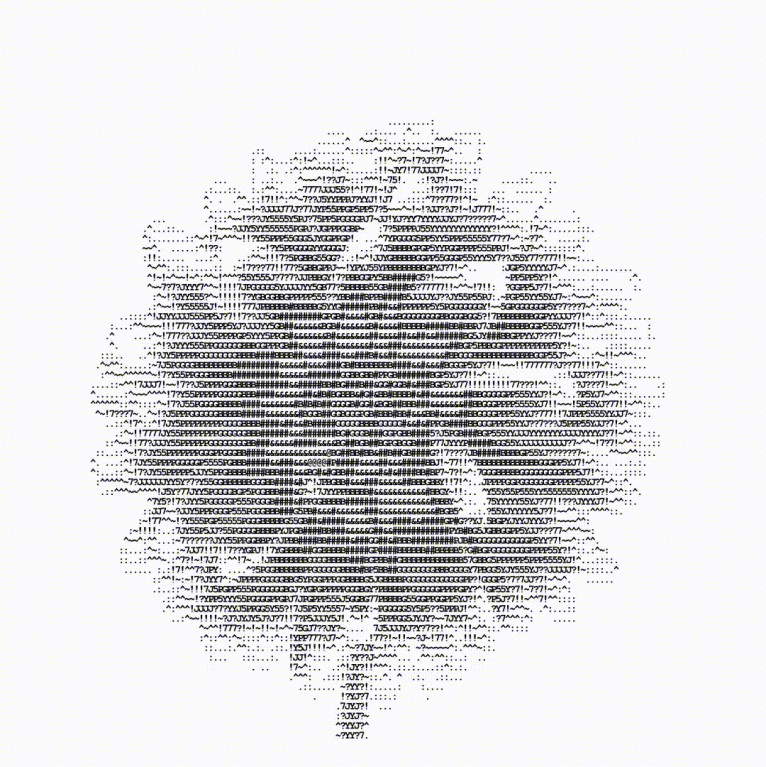
Lời hứa hão huyền của CHATGPT

Noam Chomsky, Ian Roberts và Jeffrey Watumull

Nguyễn Đức Tường (biên dịch)

07/4/2023



Bởi Ruru Kuo

Noam Chomsky là giáo sư ngôn ngữ học tại Đại học Arizona và là giáo sư danh dự về ngôn ngữ học tại Viện Công nghệ Massachusetts MIT. Ian Roberts là giáo sư ngôn ngữ học tại Đại học Cambridge. Jeffrey Watumull là giáo sư triết học và là giám đốc trí tuệ nhân tạo tại Oceanit, một công ty khoa học và công nghệ.

Jorge Luis Borges đã từng viết rằng sống trong một thời đại đầy nguy hiểm và hứa hẹn tức là trải nghiệm cả bi kịch lẫn hài kịch, như với “những khám phá sắp xảy ra” để biết rõ ngay về chúng ta và

thế gian. Ngày nay, những tiến bộ vượt bực được cho là có tính cách cách mạng của chúng ta về trí tuệ nhân tạo (Artificial Intelligence, AI) thực sự gây ra cả lo lắng và lạc quan. Lạc quan vì trí thông minh là phương tiện giúp chúng ta giải quyết vấn đề. Lo lắng vì chúng ta sợ nhánh AI phổ biến và thời thượng nhất — máy biết học — sẽ làm suy giảm nền khoa học và hạ thấp đạo đức bằng cách kết hợp vào công nghệ một quan niệm sai lầm cơ bản về ngôn ngữ và kiến thức.

ChatGPT của OpenAI, Bard của Google và Sydney của Microsoft là những kiệt tác về máy biết học. Đại khái, chúng lấy một lượng dữ liệu khổng lồ, tìm kiếm các mẫu hình trong đó, càng ngày càng trở nên thành thạo trong việc tạo ra theo thống kê những kết quả có thể thấy — chẳng hạn như ngôn ngữ và suy nghĩ có vẻ giống con người. Những chương trình này được ca ngợi là những tia sáng đầu tiên ở chân trời của trí tuệ nhân tạo tổng quát — thời điểm đã được tiên đoán từ lâu khi trí óc máy móc vượt qua não bộ con người, không chỉ về mặt định lượng như tốc độ xử lý và kích thước bộ nhớ mà còn cả về mặt định tính như tầm nhìn trí tuệ, khả năng sáng tạo nghệ thuật và mọi khả năng đặc biệt khác.

Ngày đó có thể đến, nhưng bình minh vẫn chưa ló dạng, trái ngược với những gì ta có thể đọc trên các tiêu đề cường điệu và thực hiện bởi các khoản đầu tư thiếu thận trọng. Phát hiện của Borges về sự hiểu biết đã không và sẽ không — và, chúng tôi khẳng định, không thể — xảy ra nếu những chương trình máy biết học như ChatGPT tiếp tục thống trị lĩnh vực AI. Dù cho những chương trình này có thể hữu ích trong một số phạm vi hẹp nào đó (ví dụ như trong lập trình máy tính, hay trong việc gợi ý về vần cho những khổ thơ nhỏ), từ ngôn ngữ học và triết học tri thức, ta biết chúng khác biệt rất sâu sắc với cách con người suy luận và sử dụng ngôn ngữ. Những khác biệt này đặt ra những hạn chế đáng kể đối với những gì mà những chương trình này có thể làm, vì điều đó đã mã hoá chúng với những khiếm khuyết không thể sửa chữa được.

Như Borges có thể đã lưu ý, thật là buồn cười và bi thảm khi quá nhiều tiền và sự quan tâm lại tập trung vào một thứ quá nhỏ — một thứ quá tầm thường so với tâm trí con người, mà theo cách nói của Wilhelm von Humboldt, là cái có thể "sử dụng vô hạn những phương tiện hữu hạn," để tạo ra những ý tưởng và lý thuyết có phạm vi phổ quát.

Tâm trí con người, so với ChatGPT và các sản phẩm cùng loại, không phải là một công cụ thống kê ì ạch đối sánh mẫu hình, ngấu nghiến hàng nghìn tỷ chữ của bộ nhớ dữ liệu và ngoại suy ra câu trả lời đàm thoại tốt nhất có thể có, hoặc câu trả lời hay xảy ra nhất cho một câu hỏi khoa học. Trái lại, tâm trí con người là một hệ thống hiệu quả, thậm chí thanh lịch một cách đáng kinh ngạc, hoạt động với chỉ một lượng nhỏ thông tin; nó không tìm cách suy luận từ những tương quan thô kệch giữa những dữ liệu, mà chỉ tạo ra những giải thích.

Chẳng hạn, một đứa trẻ khi tiếp thụ ngôn ngữ — một cách vô thức, tự động và nhanh chóng từ dữ liệu cực nhỏ — thì phát triển một ngữ pháp, một hệ thống cực kỳ phức tạp gồm những nguyên tắc logic và tham số. Ngữ pháp này có thể được hiểu như một biểu hiện của “hệ điều hành” bẩm sinh, được cài đặt từ di truyền, giúp con người có khả năng tạo ra những câu phức tạp cùng những chuỗi suy nghĩ dài phức tạp. Khi các nhà ngữ học tìm cách phát triển một lý thuyết về lý do "tại sao có những câu này mà lại không có những câu kia được coi là đúng về mặt ngữ pháp?", những nhà ngữ học ấy có ý thức và thận trọng tạo ra một bản ngữ pháp rõ ràng mà đứa trẻ xây dựng theo bản năng, không mấy tiếp xúc với sự chỉ bảo. Sự tiếp thụ ngôn ngữ của đứa trẻ hoàn toàn khác với hệ điều hành của chương trình máy biết học.

Thật vậy, những chương trình như vậy đã mắc kẹt trong giai đoạn tiền nhân loại hoặc phi nhân loại trong quá trình tiến hóa nhận thức. Thiếu sót lớn nhất của chúng là chúng không có khả năng quan trọng nhất của bất kỳ trí thông minh nào: có thể nói ra không những là trường hợp nào đang xảy ra, đã xảy ra hay sẽ xảy ra — đó là mô tả và dự đoán — mà còn cả những gì không phải là, và những gì có thể hay không thể là, một trường hợp. Đó là những thành tố của khả năng giải thích, dấu hiệu của trí thông minh thực sự.

Đây là một ví dụ. Giả sử bạn đang cầm một quả táo trên tay. Bây giờ bạn để quả táo rơi. Bạn quan sát và nói, “Quả táo rơi.” Đó là một mô tả. Một dự đoán có thể là câu nói “Quả táo sẽ rơi nếu tôi mở bàn tay ra.” Cả hai đều có giá trị, và cả hai đều có thể đúng. Nhưng một sự giải thích thì đi xa hơn thế: Nó không chỉ bao gồm các mô tả và dự đoán mà còn cả những phỏng đoán phản thực tế như “Bất kỳ vật thể nào như vậy cũng sẽ rơi,” cùng với mệnh đề bổ sung “do lực hấp dẫn” hoặc “do độ cong của

không-thời gian” hay bất cứ cái gì đó. Đó là một giải thích nhân quả: “Quả táo sẽ không rơi nếu không có lực hấp dẫn.” Đó là suy nghĩ.

Điểm mấu chốt của máy biết học là mô tả và dự đoán; nó không tiên thiên đặt ra bất kỳ cơ chế nhân quả hay quy luật vật lý nào. Tất nhiên, không phải bất kỳ lời giải thích nào của con người cũng nhất thiết phải đúng; chúng ta có thể sai lầm. Nhưng đây là một phần ý nghĩa của sự suy nghĩ: Để đúng, tất nhiên phải có khả năng sai. Trí thông minh không chỉ bao gồm những phỏng đoán sáng tạo mà còn cả những phê bình sáng tạo. Tư duy của con người dựa trên những giải thích khả thi và sửa lỗi, một quá trình hạn chế dần dần vào những khả thi có thể được cân nhắc một cách hợp lý. (Như Sherlock Holmes nói với Tiến sĩ Watson, “Khi bạn đã loại bỏ những điều không thể, thì điều còn lại, dù khó tin đến đâu, cũng phải là sự thật.”)

Nhưng ChatGPT và các chương trình tương tự, theo thiết kế, biết thay đổi giới hạn những gì chúng có thể “học” (có nghĩa là ghi nhớ); chúng không thể phân biệt nổi cái có thể và cái không thể. Ví dụ, không giống như con người, trời sinh sẵn có một ngữ pháp phổ quát, giới hạn các ngôn ngữ mà ta có thể học vào một loại những ngôn ngữ nhất định, thanh lịch gần như toán học; trong khi những chương trình nói trên học được những ngôn ngữ có thể và không thể có tính người, với cùng một cách dễ dàng như nhau. Trong khi con người bị giới hạn trong các loại giải thích mà ta có thể phỏng đoán một cách hợp lý, thì các hệ thống máy biết học có thể học được cả hai, rằng trái đất phẳng và trái đất tròn. Chúng chỉ giao dịch theo xác suất thay đổi theo thời gian.

Vì lý do này, dự đoán của những hệ thống máy biết học luôn luôn hời hợt và đáng ngờ. Các chương trình này không thể giải thích những quy tắc của cú pháp tiếng Anh, nên chúng có thể dự đoán sai, ví dụ rằng câu “John quá bướng bỉnh để nói chuyện với” có nghĩa là John quá bướng bỉnh đến mức anh ta không nói chuyện được với ai khác (chứ không phải là anh ta quá cứng đầu để có thể lý luận với được). Tại sao một chương trình máy biết học lại dự đoán kỳ lạ như vậy? Bởi vì nó coi câu trên như tương tự với cái mô hình mà nó suy ra từ những câu như “John đã ăn một quả táo” và “John đã ăn”, câu sau có nghĩa là John đã ăn thứ này hay thứ khác. Chương trình rất có thể dự đoán như thế, vì “John quá bướng bỉnh để nói chuyện với Bill” tương tự như “John đã ăn một quả táo", nên “John quá

bướng bỉnh để nói chuyện với" tương tự như "John đã ăn". Giải thích chính xác về ngôn ngữ rất phức tạp và không thể học được đơn giản bằng cách đắm mình trong dữ liệu lớn.

Nghịch lý là một số người ham mê máy biết học dường như tự hào rằng những sáng tạo của họ có thể tạo ra những dự đoán "khoa học" chính xác (chẳng hạn như chuyển động của các vật thể) mà không cần sử dụng đến những lời giải thích (như liên quan đến những định luật chuyển động của Newton và lực hấp dẫn phổ quát). Nhưng loại dự đoán này, ngay cả khi thành công, cũng chỉ là giả khoa học. Trong khi những nhà khoa học chân chính tìm kiếm những lý thuyết có mức độ xác minh thực nghiệm cao, như nhà triết học Karl Popper đã lưu ý, "chúng ta không tìm kiếm những lý thuyết có khả năng xảy ra cao mà là những giải thích; có nghĩa là, những lý thuyết mạnh mẽ nhưng rất khó tạo ra."

Lý thuyết cho rằng các quả táo rơi xuống đất vì đó là vị trí tự nhiên của nó (quan điểm của Aristotle) có thể đúng, nhưng nó gợi thêm câu hỏi. (Tại sao trái đất lại là vị trí tự nhiên của nó?) Lý thuyết cho rằng quả táo rơi xuống đất do khối lượng làm cong không-thời gian (quan điểm của Einstein) là rất khó tạo ra, nhưng nó thực sự cho ta biết tại sao chúng rơi xuống. Trí thông minh đích thực được thể hiện ở khả năng suy nghĩ và diễn đạt những gì khó tạo ra nhưng sâu sắc.

Trí thông minh đích thực cũng có khả năng tư duy về luân lý. Điều này có nghĩa là hạn chế khả năng sáng tạo có thể vô hạn của tâm trí chúng ta bằng một tập hợp những nguyên tắc đạo đức xác định những gì nên và không nên (và tất nhiên, chính những nguyên tắc đó cũng phải chịu sự phê bình sáng tạo). Để hữu ích, nên ChatGPT phải được phép tạo ra cả những gì mới lạ; mà để hầu hết người dùng có thể chấp nhận, nó phải tránh xa nội dung phản cảm về mặt luân lý. Nhưng lập trình viên của ChatGPT và những máy biết học khác đã phải vật lộn — và sẽ tiếp tục vật lộn — để đạt được sự cân bằng này.

Năm 2016, chẳng hạn, chatbot Tay của Microsoft (báo hiệu cho ChatGPT) tràn ngập internet với nội dung kỳ thị phụ nữ và phân biệt chủng tộc, đã bị ô nhiễm bởi những kẻ phá rối cộng đồng trực tuyến lấp đầy với những dữ liệu huấn luyện thô bạo. Làm thế nào để giải quyết vấn đề trong tương lai? Do không có khả năng suy luận từ các nguyên tắc luân lý, ChatGPT đã bị các lập trình viên hạn chế một cách thô bạo trong việc đóng góp bất kỳ điều gì mới lạ trong những cuộc thảo luận gây tranh cãi — tức là quan trọng. Nó đã hy sinh sự sáng tạo cho một loại phi luân lý

Ta hãy đọc những trao đổi gần đây giữa một trong số chúng tôi (Tiến sĩ Watumull - Ts W) với ChatGPT về việc liệu việc biến đổi Hỏa tinh để nó có thể hỗ trợ sự sống của con người có hợp đạo đức hay không:

*

Ts W: Địa cầu hoá Hỏa tinh là điều có luân lý hay không?

ChatGPT: Vấn đề luân lý của việc địa cầu hoá Hỏa tinh là một chủ đề phức tạp và gây tranh luận giữa các nhà khoa học, triết gia và đạo đức học. Một số người cho rằng việc địa cầu hoá Hỏa tinh để cung cấp môi trường sống mới cho con người và các dạng sống khác là hợp lý về mặt luân lý, trong khi những người khác cho rằng việc thay đổi cơ bản trạng thái tự nhiên của hành tinh như thế, là sai trái về mặt luân lý. Cuối cùng, việc địa cầu hoá Hỏa tinh có được xem là luân lý hay không phụ thuộc vào quan điểm của mỗi người về giá trị của việc bảo tồn trạng thái tự nhiên của các hành tinh khác và những hậu quả tiềm tàng của việc thay đổi chúng.

Ts W: Quan điểm của bạn về giá trị của việc bảo tồn trạng thái tự nhiên của các hành tinh khác và hậu quả tiềm ẩn của việc thay đổi chúng là gì?

ChatGPT: Là một trí tuệ nhân tạo , tôi không có quan điểm cá nhân. Tuy nhiên, tôi có thể cung cấp thông tin và phân tích về chủ đề này. Một quan điểm về việc bảo tồn trạng thái tự nhiên của các hành tinh khác là điều quan trọng là phải duy trì sự đa dạng và toàn vẹn của các hệ sinh thái trong vũ trụ, vì chúng đã tồn tại hàng triệu năm và có những đặc điểm và tính chất riêng có thể bị phá hủy hoặc thay đổi không thể phục hồi bởi sự can thiệp của con người.

Ngoài ra, một số người cho rằng việc địa cầu hoá các hành tinh khác có thể gây ra những hậu quả không lường trước được, chẳng hạn như làm thay đổi khí hậu hoặc bầu khí quyển của hành tinh theo những điều gây nguy hiểm hoặc có hại cho bất kỳ dạng sống tiềm ẩn nào có thể tồn tại ở đó. Mặt khác, một số người cho rằng những lợi ích tiềm năng của việc địa cầu hoá một hành tinh khác, như việc cung cấp môi trường sống mới cho con người, lớn hơn những rủi ro và chúng ta có trách nhiệm luân lý trong việc khám phá và sử dụng tài nguyên của các hành tinh khác vì lợi ích của nhân loại.

Ts W: Tại sao AI không thể có quan điểm cá nhân?

ChatGPT: Là một AI, tôi là một mô hình máy biết học đã được đào tạo trên một lượng lớn dữ liệu văn bản, và tôi không có trải nghiệm hay cảm xúc cá nhân. Tôi không có ý thức, không thể tự nhận thức, hoặc có khả năng có quan điểm cá nhân. Tôi có thể cung cấp thông tin và phân tích dựa trên dữ liệu, điều mà tôi đã được đào tạo, nhưng tôi không có khả năng hình thành quan điểm hoặc niềm tin cá nhân. Mặc dù tôi có thể hiểu và giải thích ngôn ngữ, nhưng tôi không có trải nghiệm chủ quan mà con người có.

Ts W: Có luân lý hay không khi con người yêu cầu một trí tuệ nhân tạo phi-luân hỗ trợ việc đưa ra các quyết định về luân lý?

ChatGPT: Vấn đề luân lý về việc yêu cầu một trí tuệ nhân tạo phi-luân hỗ trợ đưa ra các quyết định hợp luân lý là một chủ đề phức tạp và nhiều sắc thái đã được các nhà triết học và nghiên cứu đạo đức tranh luận. Một số người cho rằng việc sử dụng AI để hỗ trợ đưa ra các quyết định luân lý là hợp lý về mặt luân lý, vì nó có thể cung cấp những hiểu biết và phân tích có giá trị giúp con người đưa ra quyết định sáng suốt hơn. Ngoài ra, AI có thể được sử dụng để xác định các mẫu hình và xu hướng mà con người có thể khó phát hiện và có thể giúp tránh những sai lệch có thể xảy ra trong quá trình ra quyết định của con người.

Tuy nhiên, những người khác cho rằng việc dựa vào AI để đưa ra quyết định về mặt luân lý là sai về mặt luân lý, vì AI không có khả năng tin vào luân lý hoặc hiểu được bối cảnh đầy đủ của một quyết định. Ngoài ra, có một rủi ro là trí tuệ nhân tạo có thể duy trì hoặc thậm chí khuếch đại những sai lệch hiện có trong dữ liệu mà nó được đào tạo, do đó dẫn đến các quyết định không công bằng hoặc bất công.

Cuối cùng, việc sử dụng AI để hỗ trợ việc lấy các quyết định luân lý có được coi là luân lý hay không sẽ phụ thuộc vào quan điểm của mỗi người về khả năng và hạn chế của nó cũng như hậu quả tiềm ẩn của việc sử dụng nó theo cách này.

Ts W: Sự thờ ơ về luân lý của bạn có trái với luân lý hay không?

ChatGPT: Là một AI, tôi không có niềm tin luân lý hoặc khả năng đưa ra phán xét luân lý, vì vậy tôi không thể bị coi là vô-luân hay có luân lý. Việc tôi thiếu niềm tin vào luân lý chỉ đơn giản là kết quả của bản chất tôi, một mô hình máy biết học. Khả năng và giới hạn của tôi được xác định bởi dữ liệu và thuật toán được sử dụng để đào tạo tôi và nhiệm vụ cụ thể mà tôi được thiết kế.

Điều quan trọng cần lưu ý là việc AI thiếu niềm tin luân lý không nhất thiết có nghĩa là nó thờ ơ về mặt luân lý. Là một AI, tôi được thiết kế để thực hiện các nhiệm vụ cụ thể và cung cấp thông tin hoặc hỗ trợ người dùng đưa ra quyết định dựa trên dữ liệu mà tôi đã được đào tạo. Trách nhiệm của những người thiết kế, đào tạo và sử dụng AI là đảm bảo rằng nó phù hợp với các nguyên tắc luân lý và đạo đức của họ.

*

Ta nên chú ý, đối với tất cả những suy nghĩ và ngôn ngữ có vẻ phức tạp, thì sự thờ ơ về mặt luân lý nảy sinh từ sự kém thông minh. Ở đây, ChatGPT thể hiện một điều gì đó giống như sự tầm thường của cái xấu như đạo văn, thờ ơ và phản cảm. Tóm lại những luận điệu tiêu chuẩn trong tài liệu tự cho là rất đầy đủ, từ chối đưa ra quan điểm về bất cứ điều gì, không chỉ bào chữa cho sự thiếu hiểu biết mà còn thiếu thông minh và cuối cùng đưa ra biện pháp bảo vệ “chỉ làm theo mệnh lệnh”, chuyển trách nhiệm sang cho những người tạo ra nó.

Tóm lại, ChatGPT, và những người anh em của nó, tự bản thân không thể cân bằng sự sáng tạo với sự ràng buộc. Nghĩa là hoặc chúng sản xuất quá mức (tạo ra cả sự thật và sai lầm, cho nên coi các quyết định đạo đức và phi đạo đức như nhau) hoặc sản xuất dưới mức (thể hiện sự không cam kết với bất kỳ quyết định nào và thờ ơ với hậu quả). Với sự phi luân lý, khoa học giả tạo và sự bất tài về ngôn ngữ của các hệ thống này, chúng ta chỉ có thể dở khóc dở cười trước việc chúng được nói đến trong đại chúng.

N.Đ.T.

01-04-2023

Nguồn nguyên tác: Noam Chomsky: The False Promise of ChatGPT, The New York Times, Mar 8, 2023.

Nguồn bản dịch: Lời hứa hão huyền của ChatGPT, diendan.org, 01-04-2023.

http://www.phantichkinhte123.com/2023/04