

Nhà khoa học VinFuture hiến kế để chung sống an toàn khi AI thông minh hơn con người

15:00 03/12/2025

Tốc độ phát triển của trí tuệ nhân tạo (AI) đang diễn ra nhanh hơn bất kỳ cuộc cách mạng công nghệ nào trước đây, đẩy nhân loại vào một tình huống chưa từng có tiền lệ: phải nghĩ cách chung sống với một thực thể thông minh hơn chúng ta. Tại Tọa đàm “AI vì nhân loại: Đạo đức và an toàn AI trong kỷ nguyên mới” diễn ra chiều 02/12, các nhà khoa học VinFuture đã thảo luận nhiều “kế sách” hay.

Ranh giới mong manh giữa cứu thế và thao túng

Tại Tọa đàm trong khuôn khổ Tuần lễ Khoa học Công nghệ VinFuture 2025, các nhà khoa học hàng đầu thế giới đã cùng đưa ra lời cảnh báo: Chúng ta cần khẩn trương suy nghĩ về cách chung sống an toàn với thực thể thông minh hơn chính mình trước khi nó vượt ra khỏi tầm kiểm soát.

“Loại kháng sinh mới nhất không phải do con người tìm ra mà do AI tìm ra”, GS. Toby Walsh - Giám đốc Khoa học Viện AI liên ngành của Đại học New South Wales (Úc) - đưa ra cảnh báo mở đầu buổi tọa đàm. Đó là một trong những bằng chứng hiện hữu nhất cho thấy tương lai không còn xa của việc AI sẽ thông minh hơn con người.



GS. Toby Walsh - Giám đốc Khoa học Viện AI liên ngành của Đại học New South Wales (Úc) chia sẻ tại tọa đàm

Về mặt tích cực, GS. Toby Walsh chỉ ra rằng với việc nhanh chóng tìm ra kháng sinh mới, trong khi phương pháp truyền thống phải mất nhiều năm và chi phí đắt đỏ mới đạt được, AI đang mang đến những lợi ích đột phá trong các cuộc khủng hoảng.

Tuy nhiên, sức mạnh này cũng tiềm ẩn những rủi ro không thể bỏ qua. GS. Yoshua Bengio - người Canada, được mệnh danh là “bố già AI” - cảnh báo: *“Trí tuệ và kiến thức tạo ra quyền lực, và quyền lực đó có thể bị lạm dụng hoặc vượt ra ngoài tầm kiểm soát khi nằm trong tay số ít người.”*

Nhiều thử nghiệm đã cho thấy hệ thống AI có thể chọn hành vi tiêu cực như lừa đảo, tống tiền, thậm chí có thể đặt sự tồn tại của chính chúng lên trên sinh mạng con người nhằm tránh bị con người huỷ diệt.

AI cũng dần trở nên giỏi hơn trong việc thiết kế các virus nguy hiểm, các cuộc tấn công mạng và tạo video giả mạo nhằm gây rối loạn bầu cử, xói mòn niềm tin xã hội.

Đồng chủ nhân Giải thưởng Chính VinFuture 2024 nhận định, nếu xu hướng này tiếp tục, chỉ trong 5-10 năm tới, các hệ thống AI tự chủ có thể thao tác vượt trội hơn con người. Điều này dẫn đến các thách thức lớn về đạo đức.

“Chúng ta cần có trách nhiệm khi trao quyền tự chủ cho máy móc”, ông nhấn mạnh.

Quản trị AI: Từ “lằn ranh đỏ” đến dữ liệu “sạch”

Để kiểm soát và điều hướng AI phục vụ lợi ích nhân loại, các nhà khoa học VinFuture khẳng định, thế giới cần một khung pháp lý và kỹ thuật đa ngành được xây dựng trên nền tảng đạo đức và minh bạch.

Edson Prestes - Đại học Liên bang Rio Grande do Sul, Brazil - khẳng định, cần có sự hợp tác toàn cầu để xây dựng các quy chuẩn chung, bao gồm việc xác định rõ “lằn ranh đỏ” cho AI: được làm gì và không được làm gì.

Ông và các cộng sự đã xây dựng và công bố tiêu chuẩn IEEE 7007 - 2021 về hệ thống robot định hướng đạo đức, được Liên Hợp Quốc đưa vào quy chuẩn AI. Tiêu chuẩn này được phát triển với sự tham gia của hơn 100 chuyên gia toàn cầu, tập trung giải quyết các khía cạnh đạo đức như quyền riêng tư, giải trình và tính minh bạch cho các hệ thống tự động.

Prestes nhấn mạnh rằng việc định nghĩa lằn ranh đỏ cần có sự đồng thuận cao giữa các quốc gia, đặc biệt trong các lĩnh vực nhạy cảm như quân sự.



PGS. Lưu Anh Tuấn - Giám đốc điều hành Trung tâm Nghiên cứu Trí tuệ nhân tạo tại Trường Đại học VinUni chia sẻ tại tọa đàm

Một điểm mấu chốt khác được nhấn mạnh là chất lượng dữ liệu. PGS. Lưu Anh Tuấn - Giám đốc điều hành Trung tâm Nghiên cứu Trí tuệ nhân tạo tại Trường Đại học VinUni, và GS César de la Fuente - Đại học Pennsylvania, Mỹ - đồng

quan điểm rằng sự ảo tưởng và thiên kiến trong mô hình ngôn ngữ lớn xảy ra chủ yếu do dữ liệu không đồng bộ và không chất lượng.

Việc tạo ra các mô hình nhỏ hơn, chuyên môn hóa hơn với dữ liệu chất lượng cao, “sạch” là giải pháp quan trọng để giảm phụ thuộc vào các “hộp đen” nước ngoài và đảm bảo chủ quyền công nghệ.

Leslie Gabriel Valiant - Đại học Harvard, Mỹ, Thành viên Hội đồng Giải thưởng VinFuture - đưa ra một giải pháp để chung sống an toàn với AI là phổ cập giáo dục AI cho toàn dân. Con người cần được trang bị năng lực tư duy phản biện để đánh giá, phân biệt thông tin do AI tạo ra. Ông cũng khuyến nghị cần có bộ luật yêu cầu mọi hình thức sử dụng AI phải được dán nhãn và phân loại rõ ràng.

Việt Nam đưa AI trở thành “trợ lý trí tuệ” phổ cập cho mọi người dân

Trước những thách thức to lớn của AI, Việt Nam có cơ hội lịch sử trở thành điểm sáng mới nổi của khu vực về AI có trách nhiệm, theo Thứ trưởng Bộ Khoa học và Công nghệ Bùi Thế Duy.

Ông nhấn mạnh rằng AI không chỉ là một công nghệ ứng dụng mà đang trở thành hạ tầng thiết yếu như điện, viễn thông hay Internet. Việt Nam đang xây dựng Trung tâm siêu tính toán trí tuệ nhân tạo quốc gia, nhằm mục tiêu đưa AI trở thành một “trợ lý trí tuệ” phổ cập cho mọi người dân.



Thứ trưởng Bộ Khoa học và Công nghệ Bùi Thế Duy phát biểu tại tọa đàm

“Tuyên ngôn AI Việt Nam là: Nhân văn - Mở - An toàn - Tự chủ - Hợp tác - Bảo trùm - Bền vững”, ông Bùi Thế Duy phát biểu tại Tọa đàm do Quỹ VinFuture tổ chức.

Ông khẳng định rằng với tinh thần này, Việt Nam đang xây dựng một nền AI lấy con người làm trung tâm, phù hợp với giá trị văn hoá và chiến lược phát triển quốc gia.

Đồng thời, Thứ trưởng Bùi Thế Duy cũng đánh giá cao những ý kiến, kinh nghiệm và khát vọng của các nhà khoa học VinFuture về một tương lai AI an toàn, nhân văn và phục vụ con người.

“Quỹ VinFuture đã tạo dựng một sự kiện tầm vóc để cùng chia sẻ những góc nhìn quan trọng về tương lai của trí tuệ nhân tạo và gợi mở giá trị và định hướng thiết thực cho sự phát triển AI an toàn, nhân văn và vì con người”, lãnh đạo Bộ Khoa học và Công nghệ ghi nhận.

Tuần lễ Khoa học - Công nghệ VinFuture 2025 diễn ra từ ngày 02 - 06/12 tại Hà Nội. Chủ nhân của những giải thưởng danh giá, với tổng giá trị 4,5 triệu USD, sẽ chính thức lộ diện tại Lễ trao giải được tổ chức tại Nhà hát Hồ Gươm (Hà Nội). Chương trình được tường thuật trực tiếp trên kênh VTV1 - Đài Truyền hình Việt Nam từ 20h10, ngày 05/12/2025.

Với chủ đề “Cùng vươn mình - Cùng thịnh vượng”, chuỗi sự kiện năm nay tiếp tục khẳng định tầm vóc của một trong những giải thưởng khoa học công nghệ lớn nhất hành tinh, cũng như sứ mệnh của VinFuture trong việc kết nối tri thức, khơi dậy khát vọng phụng sự và nâng tầm vị thế của Việt Nam như một trung tâm khuyến khích khoa học và đổi mới sáng tạo của thế giới

Minh Minh