

AI có thể đe dọa nền dân chủ và chia rẽ thế giới?

ISSN: 2734-9195

10:30 17/10/2025

Trí tuệ nhân tạo (AI) là mối đe dọa chưa từng có đối với nhân loại bởi vì nó là công nghệ đầu tiên trong lịch sử có thể tự đưa ra quyết định và tạo ra những ý tưởng mới.

Tác giả: **Thiền giả Yuval Noah Harari**

Viết dịch: **Sa môn Lê Văn Phước**

Nguồn: www.theguardian.com

Xuyên suốt dòng chảy lịch sử, nhiều truyền thống tin rằng một số khiếm khuyết chết người, trong bản chất con người đã cám dỗ chúng ta theo đuổi những sức mạnh mà chúng ta không biết cách kiểm soát. Trong thần thoại Hy Lạp, Phaethon là một cậu con trai trẻ của Helios, người đã cầu xin cha mình cho phép điều khiển cỗ xe mặt trời. Muốn chứng minh nguồn gốc thần thánh của mình, Phaethon yêu cầu được đặc quyền điều khiển cỗ xe mặt trời.

Helios cảnh báo Phaethon rằng không con người nào có thể điều khiển những con ngựa trên trời kéo cỗ xe mặt trời.

Nhưng Phaethon vẫn khẳng định, cho đến khi thần mặt trời nhượng bộ. Sau khi kiêu hãnh bay lên bầu trời, thực sự Phaethon mất kiểm soát cỗ xe. Mặt trời lệch hướng, thiêu rụi mọi thảm thực vật, giết chết vô số sinh vật và đe dọa thiêu rụi cả Trái Đất. Thần Sấm Sét (Zeus), can thiệp và giáng xuống Phaethon một tia sét. Kẻ kiêu ngạo rơi xuống từ bầu trời như một ngôi sao băng, bản thân bị bốc cháy. Các vị thần giành lại quyền kiểm soát bầu trời và cứu thế giới.

Hai nghìn năm sau, khi cách mạng công nghiệp đang trên hành trình những bước đầu tiên và trong nhiều công việc máy móc bắt đầu thay thế con người, nhà thơ, nhà viết kịch, tiểu thuyết gia, nhà văn, nhà khoa học, họa sĩ người Đức Johann Wolfgang von Goethe (1749-1832) đã xuất bản một câu chuyện cảnh báo tương tự có tựa đề “*The Sorcerer’s Apprentice*” (Người học việc của phù thủy).

Bài thơ của Goethe (sau này được phổ biến thành phim hoạt hình của Walt Disney với sự tham gia của Chuột Mickey) kể về một phù thủy già giao cho một người học việc trẻ tuổi quản lý xưởng của mình và giao cho anh ta một số việc vặt khi ông đi vắng, chẳng hạn như lấy nước từ sông.

Người học việc quyết định làm mọi việc dễ dàng hơn cho mình và, sử dụng một trong những câu thần chú phép thuật của phù thủy, phù phép một cây chổi để lấy nước cho anh ta. Nhưng người học việc không biết làm thế nào để ngăn cây chổi, nó liên tục lấy nước ngày càng nhiều, đe dọa làm ngập xưởng.



Thiền giả Yuval Noah Harari. Ảnh: theguardian.com

Trong cơn hoảng loạn, người học việc chặt cây chổi bị phù phép làm đôi bằng một chiếc rìu, chỉ để thấy mỗi nửa biến thành một cây chổi khác. Giờ đây, hai cây chổi bị phù phép đang làm ngập xưởng ngấn chìm nước. Khi vị phù thủy già trở về, người học việc cầu xin sự giúp đỡ: *“Những linh hồn mà ta đã triệu hồi, giờ ta không thể thoát khỏi chúng nữa.”* Ngay lập tức, vị phù thủy phá bỏ bùa chú và ngăn chặn trận lũ. Bài học cho người học việc - và cho cả nhân loại - rất rõ ràng: đừng bao giờ triệu hồi những sức mạnh mà ta không thể kiểm soát.

Trong thế kỷ XXI này, những câu chuyện ngụ ngôn cảnh báo về Người học việc của phù thủy và Phaethon nói lên điều gì với chúng ta?

Rõ ràng nhân loại chúng ta đã từ chối lắng nghe những lời cảnh báo của họ. Chúng ta đã khiến khí hậu Trái Đất mất cân bằng và đã triệu hồi hàng tỷ cây chổi, máy bay không người lái, chatbot và các linh hồn thuật toán khác có thể

thoát khỏi tầm kiểm soát của chúng ta và gây ra hàng loạt hậu quả. Vậy chúng ta nên làm gì?

Những câu chuyện ngụ ngôn không đưa ra câu trả lời nào, ngoài việc chờ đợi một vị thần hay phù thủy nào đó cứu rỗi chúng ta.

Trong thần thoại Hy Lạp, Phaethon và bài thơ của Goethe không đưa ra được lời khuyên hữu ích nào vì chúng đã hiểu sai cách con người khi đạt được quyền lực. Trong cả hai câu chuyện ngụ ngôn, một cá nhân đạt được quyền lực to lớn, nhưng sau đó lại bị tha hóa bởi sự kiêu ngạo và lòng tham lam. Kết luận là tâm lý cá nhân khiếm khuyết của chúng ta khiến chúng ta lạm dụng quyền lực. Điều mà phân tích thô thiển này bỏ sót là quyền lực của con người không bao giờ là kết quả của sáng kiến cá nhân. Quyền lực luôn bắt nguồn từ sự hợp tác giữa nhiều người.

Theo đó, không phải tâm lý cá nhân khiến chúng ta lạm dụng quyền lực. Xét cho cùng, bên cạnh lòng tham lam, sự kiêu ngạo và sự tàn ác, con người cũng có khả năng yêu thương, thương cảm, khiêm nhường và vui vẻ. Đúng vậy, trong số những thành viên tồi tệ nhất của loài người, lòng tham lam và sự tàn ác ngự trị tối cao dẫn dắt những kẻ xấu lạm dụng quyền lực.

Nhưng tại sao xã hội loài người lại chọn giao phó quyền lực cho những thành viên tồi tệ nhất của mình? Ví dụ, vào năm 1933 hầu hết người Đức không phải là những kẻ tâm thần. Vậy tại sao họ lại bỏ phiếu cho nhà độc tài Adolf Hitler?

Xu hướng triệu hồi những sức mạnh mà chúng ta không thể kiểm soát không bắt nguồn từ tâm lý cá nhân mà từ cách thức độc đáo mà loài người chúng ta hợp tác với số lượng lớn. Nhân loại đạt được sức mạnh to lớn bằng cách xây dựng những mạng lưới hợp tác rộng lớn, nhưng cách thức xây dựng mạng lưới này lại khiến chúng ta sử dụng sức mạnh một cách thiếu thông minh. Hầu hết các mạng lưới của chúng ta được xây dựng và duy trì bằng cách lan truyền những câu chuyện hư cấu, tưởng tượng và ảo tưởng đại chúng - từ những câu chối thần kỳ đến các hệ thống tài chính. Vậy thì, vấn đề của chúng ta là một vấn đề về mạng lưới. Cụ thể hơn, đó là một vấn đề về thông tin. Bởi vì thông tin là chất keo gắn kết các mạng lưới lại với nhau, và khi mọi người được cung cấp thông tin sai lệch, họ có thể đưa ra những quyết định sai lầm, bất kể họ khôn ngoan và tử tế đến đâu.

Hiện nay trong những thế hệ trẻ, từ trước đến nay nhân loại đã trải qua sự gia tăng lớn nhất về cả số lượng lẫn tốc độ sản xuất thông tin. Mỗi chiếc điện thoại thông minh chứa đựng lượng thông tin lớn hơn cả Thư viện Alexandria, thành phố Alexandria của Ai Cập từng là một trong những thư viện lớn và quan trọng

nhất trong số các thư viện của thế giới cổ đại và cho phép chủ sở hữu kết nối tức thời với hàng tỷ người khác trên khắp thế giới. Tuy nhiên, với tất cả lượng thông tin này được lưu hành với tốc độ chóng mặt, nhân loại đang tiến gần hơn bao giờ hết đến sự hủy diệt của chính mình.

Hoặc có lẽ chính vì - sự bất chấp - kho dữ liệu khổng lồ của mình, chúng ta vẫn tiếp tục thải khí nhà kính vào khí quyển, gây ô nhiễm sông ngòi và đại dương, chặt phá rừng, phá hủy toàn bộ môi trường sống, đẩy vô số giống loài đến bờ vực tuyệt chủng và gây nguy hiểm cho nền tảng sinh thái của chính loài người. Chúng ta cũng đang sản xuất ngày càng nhiều vũ khí vũ khí mạnh mẽ nhất đã hủy diệt hàng loạt, từ bom nhiệt hạch đến virus "*ngày tận thế*". Các nhà lãnh đạo của chúng ta không thiếu thông tin về những mối nguy hiểm này, vậy mà thay vì hợp tác để tìm ra giải pháp, họ lại đang tiến gần hơn đến một cuộc chiến tranh toàn cầu.

Liệu có thêm thông tin sẽ khiến mọi thứ tốt hơn - hay tệ hơn? Chúng ta sẽ sớm tìm ra câu trả lời. Nhiều tập đoàn và chính phủ đang chạy đua để phát triển công nghệ thông tin mạnh mẽ nhất trong lịch sử - AI.

Một số doanh nhân hàng đầu, chẳng hạn như nhà đầu tư người Mỹ, nhà sáng lập quỹ Andreessen Horowitz (a16z) chuyên đầu tư mạo hiểm và từ lâu ủng hộ ý tưởng AI nguồn mở tỷ phú Marc Andreessen, tin rằng cuối cùng AI sẽ giải quyết được mọi vấn đề của nhân loại.



Thiền giả Yuval Noah Harari. Ảnh: theguardian.com

Vào ngày 6 tháng 6 năm 2023, tỷ phú Marc Andreessen đã xuất bản một bài luận có tựa đề *"Tại sao AI sẽ cứu thế giới"*, với những tuyên bố táo bạo như: *"Nơi đây, để tôi mang đến tin tốt: AI sẽ không hủy diệt thế giới, và trên thực tế có thể cứu nó."* Ông kết luận: *"Sự phát triển và phổ biến của AI - hoàn toàn không phải là một rủi ro mà chúng ta nên lo sợ - là một nghĩa vụ đạo đức mà chúng ta phải có đối với bản thân, với con cháu và với tương lai của chúng ta."*

Những người khác thì hoài nghi hơn. Không chỉ các nhà triết học và nhà khoa học xã hội mà còn nhiều chuyên gia AI và doanh nhân hàng đầu như Giáo sư Yoshua Bengio, chuyên gia về AI, Giáo sư Geoffrey Hinton, nhà tâm lý học nhận

thức và khoa học máy tính người Anh - Canada, nhà đầu tư và lập trình viên người Mỹ, cựu giám đốc điều hành của OpenAI và là cựu chủ tịch của Y Combinator, doanh nhân Sam Altman, tỷ phú Elon Musk, người giàu nhất thế giới hiện nay và doanh nhân Mustafa Suleyman, nhà nghiên cứu người Anh chuyên về trí tuệ nhân tạo đã cảnh báo rằng AI có thể hủy diệt nền văn minh của chúng ta.

Trong một cuộc khảo sát năm 2023 với 2.778 nhà nghiên cứu AI, hơn một phần ba cho rằng ít nhất 10% khả năng AI tiên tiến sẽ dẫn đến những hậu quả tồi tệ như sự tuyệt chủng của loài người.

Trong hai ngày (1 và 2 tháng 11 năm 2023, Hội nghị thượng đỉnh toàn cầu về trí tuệ nhân tạo (AI) đầu tiên đã được tổ chức tại Bletchley Park, Vương quốc Anh, với sự tham dự của khoảng 100 đại biểu là các nhà lãnh đạo từ 28 quốc gia, đại diện các công ty công nghệ hàng đầu, các học giả và các thành phần khác.

Đáng quan tâm, nước chủ nhà đã công bố *“Tuyên bố Bletchley”*, được nhất trí với các nước như Hoa Kỳ và Trung Quốc, nhằm thúc đẩy các nỗ lực phối hợp toàn cầu để đảm bảo an toàn trong lĩnh vực AI. *“Tuyên bố Bletchley”* do 28 quốc gia ký kết, được công bố ngay trong ngày đầu tiên của Hội nghị Cấp cao an toàn AI toàn cầu, toàn thể hội nghị đã thừa nhận rằng *“có khả năng gây ra tác hại nghiêm trọng, thậm chí thảm khốc, dù là cố ý hay vô ý, bắt nguồn từ những khả năng quan trọng nhất của các mô hình AI này”*. Bằng cách sử dụng những thuật ngữ khai huyền như vậy, các chuyên gia và chính phủ không muốn gọi lên hình ảnh Hollywood về những con robot nổi loạn chạy trên đường phố và bắn người. Một kịch bản như vậy là không thể xảy ra, và thực sự nó chỉ khiến mọi người mất tập trung khỏi những nguy hiểm.

Trí tuệ nhân tạo (AI) là mối đe dọa chưa từng có đối với nhân loại bởi vì nó là công nghệ đầu tiên trong lịch sử có thể tự đưa ra quyết định và tạo ra những ý tưởng mới. Trước đây, tất cả các phát minh của con người đều trao quyền cho con người, bởi vì dù công cụ mới có mạnh mẽ đến đâu, quyết định về việc sử dụng nó vẫn nằm trong tay chúng ta. Bom hạt nhân không tự quyết định giết ai, cũng không thể tự cải tiến hay phát minh ra những quả bom mạnh hơn. Ngược lại, máy bay không người lái tự động có thể tự quyết định giết ai, và AI có thể tạo ra các thiết kế bom mới lạ, các chiến lược quân sự chưa từng có và các AI tốt hơn. AI không phải là một công cụ - nó là một tác nhân.

Mối đe dọa lớn nhất của AI là chúng ta đang triệu hồi vô số tác nhân mới mạnh mẽ đến Trái Đất, những tác nhân này có tiềm năng thông minh và giàu trí tưởng tượng hơn chúng ta, và chúng ta không hoàn toàn hiểu hoặc kiểm soát được.

“Trong vài thập kỷ tới, rất có thể AI sẽ đạt được khả năng tạo ra các dạng sống mới”.

Theo truyền thống, thuật ngữ “AI” được dùng như một từ viết tắt của trí tuệ nhân tạo. Nhưng có lẽ tốt hơn nên hiểu nó như một từ viết tắt của trí thông minh ngoài hành tinh. Khi AI phát triển, nó trở nên ít nhân tạo hơn (theo nghĩa phụ thuộc vào thiết kế của con người) và trở nên xa lạ hơn. Nhiều người cố gắng đo lường và thậm chí định nghĩa AI bằng thước đo *“trí thông minh ngang tầm con người”*, và có một cuộc tranh luận sôi nổi về thời điểm chúng ta có thể kỳ vọng AI đạt đến mức đó. Thước đo này rất dễ gây hiểu lầm. Nó giống như việc định nghĩa và đánh giá máy bay thông qua thước đo *“bay ngang tầm chim”*. AI không tiến triển đến mức trí thông minh ngang tầm con người. Nó đang phát triển một loại trí thông minh ngoài hành tinh.

Ngay cả ở thời điểm hiện tại, trong giai đoạn phôi thai của cuộc cách mạng AI, máy tính đã đưa ra quyết định về chúng ta - cho chúng ta vay thế chấp, thuê chúng ta làm việc, hay tống chúng ta vào tù. Trong khi đó, các AI sinh sản như GPT-4 đã tạo ra những bài thơ, câu chuyện và hình ảnh mới. Xu hướng này sẽ chỉ ngày càng gia tăng và tăng tốc, khiến việc hiểu cuộc sống của chính chúng ta trở nên khó khăn hơn. Liệu chúng ta có thể tin tưởng các thuật toán máy tính sẽ đưa ra những quyết định sáng suốt và tạo ra một thế giới tốt đẹp hơn không? Đây là một canh bạc lớn hơn nhiều so với việc tin tưởng một cây chổi thần kỳ để lấy nước. Và chúng ta đang đánh cược không chỉ vào mạng sống con người. AI đã có khả năng tự mình tạo ra nghệ thuật và thực hiện những khám phá khoa học. Trong vài thập kỷ tới, nó có thể sẽ đạt được khả năng tạo ra các dạng sống mới, bằng cách viết mã di truyền hoặc bằng cách phát minh ra mã vô cơ làm hoạt hình các thực thể vô cơ. Do đó, AI có thể thay đổi tiến trình không chỉ của lịch sử nhân loại mà còn của sự tiến hóa của tất cả các dạng sống.

Đây là lý do tại sao cả các chuyên gia cờ vây và chuyên gia máy tính đều sửng sốt, vào tháng 3 năm 2016 khi AlphaGo đánh bại nhà vô địch cờ vây người Hàn Quốc Lee Sedol (李世石), từ đó anh quyết định từ giã sự nghiệp thi đấu chuyên nghiệp vì cho rằng máy móc sử dụng trí tuệ nhân tạo (AI) đã quá thông minh.

Trong cuốn sách *“Sóng Thần Công Nghệ - The Coming Wave”* Nhà xuất bản: New York: Crown, (2023) Tác giả Suleyman đã mô tả một trong những khoảnh khắc quan trọng nhất trong trận đấu của họ - một khoảnh khắc đã định nghĩa lại AI và được nhiều giới học thuật và chính phủ công nhận là một bước ngoặt quan trọng trong lịch sử. Điều này xảy ra trong ván đấu thứ hai của trận đấu, vào ngày 10 tháng 3 năm 2016.

“Rồi... đến nước đi thứ 37,” Tác giả Suleyman viết. “Nó thật vô lý. Rõ ràng AlphaGo đã thất bại, mà quá đi theo một chiến thuật tưởng chừng như thua cuộc mà không kỳ thủ chuyên nghiệp nào dám theo đuổi. Các bình luận viên trực tiếp trận đấu, cả hai đều là kỳ thủ chuyên nghiệp hàng đầu, đều nói rằng đây là một ‘nước đi rất kỳ lạ’ và cho rằng đó là ‘một sai lầm’. Nó bất thường đến mức nhà vô địch cờ vây người Hàn Quốc Lee Sedol mất 15 phút để phản ứng và thậm chí còn đứng dậy khỏi bàn cờ để đi dạo. Khi chúng tôi theo dõi từ phòng điều khiển, sự căng thẳng thật khó tin. Nhưng khi tàn cuộc đến gần, nước đi ‘sai lầm’ đó lại trở thành bước ngoặt. AlphaGo lại chiến thắng. Chiến lược cờ vây đang được viết lại trước mắt chúng tôi. AI của chúng tôi đã khám phá ra những ý tưởng mà những kỳ thủ xuất sắc nhất trong hàng nghìn năm qua chưa từng nghĩ đến.”

Nước đi thứ 37 là biểu tượng của cuộc cách mạng AI vì hai lý do.

Thứ nhất, nó chứng minh bản chất xa lạ của AI. Ở Đông Á, Cờ vây được coi là hơn cả một trò chơi: nó là một truyền thống văn hóa quý giá. Trong hơn 2.500 năm, hàng chục triệu người đã chơi Cờ vây, và toàn bộ các trường phái tư tưởng đã phát triển xung quanh trò chơi, ủng hộ các chiến lược và triết lý khác nhau. Tuy nhiên, trong suốt hàng thiên niên kỷ đó, tâm trí con người chỉ khám phá một số lĩnh vực nhất định trong bối cảnh Cờ vây. Những lĩnh vực khác vẫn còn nguyên vẹn, bởi vì tâm trí con người không nghĩ đến việc mạo hiểm đến đó. AI, thoát khỏi những hạn chế của tâm trí con người, đã khám phá và khai phá những lĩnh vực trước đây còn ẩn giấu này.

Thứ hai, nước đi thứ 37 đã chứng minh sự khó hiểu của AI. Ngay cả sau khi AlphaGo chơi nước đi này và giành chiến thắng, Tác giả Suleyman và nhóm của ông vẫn không thể giải thích được AlphaGo quyết định chơi như thế nào. Ngay cả khi tòa án đã ra lệnh cho DeepMind cung cấp cho nhà vô địch cờ vây người Hàn Quốc Lee Sedol một lời giải thích, thì cũng không ai có thể thực hiện được lệnh đó.

Tác giả Suleyman viết: *“Hiện tại, trong AI, việc các mạng nơ-ron hướng tới tính tự chủ vẫn chưa thể giải thích được. Bạn không thể hướng dẫn ai đó qua quy trình ra quyết định để giải thích chính xác tại sao một thuật toán lại đưa ra một dự đoán cụ thể. Các kỹ sư không thể nhìn thấu bên trong và dễ dàng giải thích chi tiết nguyên nhân dẫn đến điều gì đó xảy ra. GPT-4, AlphaGo và những người còn lại là những hộp đen, đầu ra và quyết định của chúng dựa trên những chuỗi tín hiệu nhỏ bé, mơ hồ và phức tạp đến mức không thể tưởng tượng nổi.”*

Sự trỗi dậy của trí thông minh ngoài hành tinh khó lường, gây ra mối đe dọa cho toàn thể nhân loại, đặc biệt là mối đe dọa đối với nền dân chủ. Nếu ngày càng

nhiều quyết định về cuộc sống của con người được đưa ra trong một hộp đen, khiến cử tri không thể hiểu và phản biện chúng, thì nền dân chủ sẽ không còn hoạt động. Cụ thể, điều gì sẽ xảy ra khi những quyết định quan trọng không chỉ liên quan đến cuộc sống cá nhân mà còn liên quan đến các vấn đề tập thể như lãi suất của Cục Dự trữ Liên bang (Fed) là một cơ quan chính phủ độc lập nhưng chịu trách nhiệm trước công chúng và Quốc hội Hoa Kỳ được đưa ra bởi các thuật toán khó hiểu? Cử tri nhân loại có thể tiếp tục bầu chọn một Tổng thống nhân loại, nhưng liệu đây có phải chỉ là một buổi lễ sáo rỗng? Ngay cả ngày nay, chỉ một phần nhỏ nhân loại thực sự hiểu được hệ thống tài chính.

Một cuộc khảo sát năm 2014 đối với các nghị sĩ Anh - những người chịu trách nhiệm quản lý một trong những trung tâm tài chính quan trọng nhất thế giới - cho thấy chỉ 12% hiểu chính xác rằng tiền mới được tạo ra khi các ngân hàng cho vay. Thực tế này là một trong những nguyên tắc cơ bản nhất của hệ thống tài chính hiện đại. Như Cuộc khủng hoảng tài chính toàn cầu 2007-2008 là một sự kiện toàn cầu chứ không chỉ giới hạn ở Hoa Kỳ đã chỉ ra, một số công cụ và nguyên tắc tài chính phức tạp chỉ được một số ít chuyên gia tài chính hiểu rõ. Điều gì sẽ xảy ra với nền dân chủ khi AI tạo ra các thiết bị tài chính phức tạp hơn và khi số lượng con người hiểu được hệ thống tài chính giảm xuống bằng không?

Chuyển thể câu chuyện ngụ ngôn cảnh báo của Goethe sang ngôn ngữ tài chính hiện đại, hãy tưởng tượng tình huống sau: một học viên tại Phố Wall chán ngấy công việc nặng nhọc của xưởng tài chính đã tạo ra một AI tên là Broomstick, cung cấp cho nó tiền vốn ban đầu một triệu USD và ra lệnh cho nó kiếm thêm tiền. Đối với AI, tài chính là sân chơi lý tưởng, vì nó là một lĩnh vực thuần túy thông tin và toán học. AI vẫn gặp khó khăn trong việc tự lái xe, bởi vì điều này đòi hỏi phải di chuyển và tương tác trong thế giới vật lý hỗn loạn, nơi mà “*thành công*” rất khó định nghĩa. Ngược lại, để thực hiện các giao dịch tài chính, AI chỉ cần xử lý dữ liệu, và nó có thể dễ dàng đo lường thành công của mình bằng toán học bằng đô la, euro hoặc bảng Anh. Thêm đô la - nhiệm vụ đã hoàn thành.

“Máy tính vẫn chưa đủ mạnh để tự mình hủy diệt nền văn minh nhân loại. Miễn là nhân loại đoàn kết, chúng ta có thể xây dựng các thể chế quản lý AI.”

Để theo đuổi mục tiêu kiếm nhiều tiền hơn, Broomstick không chỉ đưa ra các chiến lược đầu tư mới mà còn tạo ra những công cụ tài chính hoàn toàn mới mà chưa một con người nào từng nghĩ đến. Trong hàng nghìn năm, trí tuệ con người chỉ khám phá một số lĩnh vực nhất định trong lĩnh vực tài chính. Họ đã phát

minh ra tiền, séc, trái phiếu, cổ phiếu, ETF, CDO và các loại phép thuật tài chính khác. Nhưng nhiều lĩnh vực tài chính vẫn còn bỏ ngỏ, bởi vì trí tuệ con người chưa nghĩ đến việc mạo hiểm vào đó. Broomstick, thoát khỏi những giới hạn của trí tuệ con người, đã khám phá và khai phá những lĩnh vực trước đây bị che giấu này, thực hiện những bước đi tài chính tương đương với nước đi số 37 của AlphaGo.

Trong vài năm, khi Broomstick dẫn dắt nhân loại vào vùng đất tài chính nguyên sơ, mọi thứ trông thật tuyệt vời. Thị trường tăng vọt, tiền đổ vào dễ dàng, và mọi người đều vui vẻ. Rồi xảy ra một cuộc khủng hoảng thậm chí còn lớn hơn cả năm 1929 hay 2008. Nhưng không một con người nào - dù là Tổng thống, chủ ngân hàng hay công dân - biết nguyên nhân và cách giải quyết. Vì cả thần thánh lẫn phù thủy đều không thể cứu vãn hệ thống tài chính, nên các chính phủ tuyệt vọng đã cầu cứu sự giúp đỡ từ thực thể duy nhất có khả năng hiểu được chuyện gì đang xảy ra - Broomstick. AI đưa ra một số khuyến nghị chính sách, táo bạo hơn nhiều so với nói lung tung định lượng - và cũng mờ ám hơn nhiều. Broomstick hứa hẹn rằng những chính sách này sẽ cứu vãn tình hình, nhưng các chính trị gia loài người - không thể hiểu được logic đằng sau các khuyến nghị của Broomstick - lo sợ rằng chúng có thể làm đảo lộn hoàn toàn cấu trúc tài chính và thậm chí là xã hội của thế giới. Liệu họ có nên nghe theo AI?

Máy tính vẫn chưa đủ mạnh để hoàn toàn thoát khỏi sự kiểm soát của chúng ta hoặc tự hủy diệt nền văn minh nhân loại. Khi nào nhân loại còn đoàn kết, chúng ta có thể xây dựng các thể chế điều chỉnh AI, dù là trong lĩnh vực tài chính hay chiến tranh. Đáng tiếc là nhân loại chưa bao giờ đoàn kết. Chúng ta luôn bị quấy nhiễu bởi những kẻ xấu, cũng như bởi những bất đồng giữa những kẻ tốt. Sự trỗi dậy của AI đặt ra một mối nguy hiểm hiện hữu cho nhân loại, không phải vì sự độc ác của máy tính, mà vì những thiếu sót của chính chúng ta.

Do đó, một nhà độc tài hoang tưởng có thể trao quyền lực vô hạn cho một AI dễ mắc sai lầm, bao gồm cả quyền phát động các cuộc tấn công hạt nhân. Nếu sau đó AI mắc lỗi hoặc bắt đầu theo đuổi một mục tiêu bất ngờ, hậu quả có thể là thảm họa, và không chỉ đối với quốc gia nào đó. Tương tự, những kẻ khủng bố có thể sử dụng AI để kích động một đại dịch toàn cầu. Bản thân những kẻ khủng bố có thể không có nhiều kiến thức về dịch tễ học, nhưng AI có thể tổng hợp cho chúng một mầm bệnh mới, đặt hàng từ các phòng thí nghiệm thương mại hoặc in nó bằng máy in 3D sinh học, và đưa ra chiến lược tốt nhất để phát tán nó trên toàn thế giới, thông qua các sân bay hoặc chuỗi cung ứng thực phẩm. Điều gì sẽ xảy ra nếu AI tổng hợp một loại virus gây chết người như Ebola, dễ lây lan như Covid-19 và hoạt động chậm như HIV? Đến khi những nạn nhân đầu tiên bắt đầu tử vong và thế giới được cảnh báo về mối nguy hiểm, thì hầu hết

mọi người trên Trái đất có thể đã bị nhiễm bệnh.

Nền văn minh nhân loại cũng có thể bị tàn phá hàng loạt bởi vũ khí hủy diệt xã hội, chẳng hạn như những câu chuyện làm xói mòn các mối quan hệ xã hội của chúng ta. Một AI được phát triển ở một quốc gia có thể được sử dụng để tung ra một loạt tin tức giả, tiền giả và con người giả, khiến người dân ở nhiều quốc gia khác mất đi khả năng tin tưởng bất cứ điều gì hoặc bất cứ ai.

Nhiều xã hội - cả **dân chủ** lẫn độc tài - có thể hành động có trách nhiệm để điều chỉnh việc sử dụng AI, trấn áp những kẻ xấu và kiểm chế tham vọng nguy hiểm của chính những người cai trị và những kẻ cuồng tín của họ. Nhưng nếu ngay cả một số ít xã hội không làm được như thế, điều này cũng đủ để gây nguy hiểm cho toàn thể nhân loại. Biến đổi khí hậu có thể tàn phá ngay cả những quốc gia áp dụng các quy định môi trường xuất sắc, bởi vì đây là vấn đề toàn cầu chứ không giới hạn ở bất kỳ quốc gia nào. AI cũng là một vấn đề toàn cầu. Theo đó, để hiểu được chính trị máy tính mới, việc xem xét cách các xã hội riêng biệt có thể phản ứng với AI là chưa đủ. Chúng ta cũng cần xem xét cách AI có thể thay đổi mối quan hệ giữa các xã hội trên phạm vi toàn cầu.

Vào thế kỷ XVI, khi những nhà chinh phạt Tây Ban Nha, Bồ Đào Nha và Hà Lan xây dựng những đế chế toàn cầu đầu tiên trong lịch sử, họ mang theo thuyền buồm, ngựa và thuốc súng. Vào thế kỷ XIX và XX khi người Anh, người Nga và người Nhật Bản thực hiện các nỗ lực giành quyền bá chủ, họ dựa vào tàu hơi nước, đầu máy xe lửa và súng máy. Vào thế kỷ XXI, để thống trị một thuộc địa, bạn không còn cần phải điều động tàu chiến nữa. Bạn cần phải lấy dữ liệu.

Một vài tập đoàn hoặc chính phủ thu thập dữ liệu của thế giới có thể biến phần còn lại của địa cầu thành các thuộc địa dữ liệu - những lãnh thổ mà họ kiểm soát không phải bằng vũ lực quân sự công khai mà bằng thông tin.

Hãy tưởng tượng một tình huống - chẳng hạn trong 20 năm nữa - khi một ai đó ở Bắc Kinh hoặc San Francisco nắm giữ toàn bộ lịch sử cá nhân của mọi chính trị gia, nhà báo, đại tá và giám đốc điều hành ở quốc gia bạn: mọi tin nhắn họ từng gửi, mọi tìm kiếm trên web họ từng thực hiện, mọi căn bệnh họ từng mắc phải, mọi cuộc tình họ từng tận hưởng, mọi trò đùa họ kể, mọi hối lộ họ nhận. Liệu bạn vẫn sẽ sống ở một quốc gia độc lập, hay giờ đây bạn sẽ sống trong một thuộc địa dữ liệu?

Điều gì sẽ xảy ra khi quốc gia của bạn thấy mình hoàn toàn phụ thuộc vào cơ sở hạ tầng kỹ thuật số và các hệ thống do AI hỗ trợ mà nó không thể kiểm soát hiệu quả?

“Trong một nền kinh tế toàn cầu do AI thúc đẩy, các công ty dẫn đầu về công nghệ số sẽ chiếm phần lớn lợi ích. Trong khi đó, giá trị của lao động phổ thông ở các quốc gia bị bỏ lại phía sau sẽ giảm.”

Trong lĩnh vực kinh tế, trước đây các đế chế dựa trên các nguồn tài nguyên vật chất như đất đai, bông và dầu mỏ. Điều này đặt ra giới hạn cho khả năng tập trung cả của cải kinh tế lẫn quyền lực chính trị của đế chế vào một nơi. Vật lý và địa chất không cho phép toàn bộ đất đai, bông hay dầu mỏ của thế giới được chuyển đến một quốc gia. Các đế chế thông tin mới thì khác. Dữ liệu có thể di chuyển với tốc độ ánh sáng, và các thuật toán không chiếm nhiều không gian. Do đó, sức mạnh thuật toán của thế giới có thể tập trung vào một trung tâm duy nhất. Các kỹ sư ở một quốc gia duy nhất có thể viết mã và kiểm soát các khóa cho tất cả các thuật toán quan trọng vận hành toàn bộ thế giới.

Do đó, AI và tự động hóa đặt ra một thách thức đặc biệt đối với các nước đang phát triển nghèo hơn. Trong một nền kinh tế toàn cầu do AI thúc đẩy, các nước dẫn đầu về công nghệ số sẽ nắm giữ phần lớn lợi ích và có thể sử dụng tài sản của mình để đào tạo lại lực lượng lao động và thu về nhiều lợi nhuận hơn nữa. Trong khi đó, giá trị của lao động phổ thông ở các quốc gia bị bỏ lại phía sau sẽ giảm, khiến họ càng tụt hậu hơn nữa.

Kết quả có thể là rất nhiều việc làm mới và khối tài sản khổng lồ ở San Francisco và Thượng Hải, trong khi nhiều nơi khác trên thế giới phải đối mặt với sự suy thoái kinh tế.

Theo công ty kiểm toán toàn cầu PricewaterhouseCoopers, vào năm 2030 AI dự kiến sẽ bổ sung 15,7 nghìn tỷ USD (12,3 nghìn tỷ bảng Anh) vào nền kinh tế toàn cầu. Nhưng nếu xu hướng hiện tại tiếp tục, dự kiến Trung Quốc và Bắc Mỹ - hai siêu cường AI hàng đầu - sẽ cùng nhau chiếm 70% số tiền đó.

Trong thời Chiến tranh Lạnh, bức màn sắt ở nhiều nơi thực chất được làm bằng kim loại: hàng rào thép gai ngăn cách các quốc gia. Giờ đây, thế giới ngày càng bị chia cắt bởi bức màn Silic (Silicon Curtain). Giống như Bức màn Sắt đã từng chia cắt thế giới trong Chiến tranh Lạnh, Màn che Silicon được dệt nên từ những dòng mã lập trình, chia rẽ thế giới thành các khối kỹ thuật số riêng biệt, mỗi khối vận hành theo luật lệ và hệ giá trị riêng. Mã trên điện thoại thông minh của bạn quyết định bạn sống ở phía nào của bức Màn che Silicon, thuật toán nào điều hành cuộc sống của bạn, ai kiểm soát sự chú ý của bạn và dữ liệu của bạn chảy về đâu.

Việc tiếp cận thông tin qua bức màn Silicon đang trở nên khó khăn, chẳng hạn như giữa Trung Quốc và Hoa Kỳ, hoặc giữa Nga và EU. Hơn nữa, hai

bên ngày càng hoạt động trên các mạng kỹ thuật số khác nhau, sử dụng các mã máy tính khác nhau. Ở Trung Quốc, bạn không thể sử dụng Google hay Facebook, và bạn cũng không thể truy cập Wikipedia. Ở Hoa Kỳ, rất ít người sử dụng các ứng dụng hàng đầu của Trung Quốc như WeChat. Quan trọng hơn, hai lĩnh vực kỹ thuật số này không phải là hình ảnh phản chiếu của nhau. Baidu không phải là Google của Trung Quốc. Alibaba không phải là Amazon của Trung Quốc. Họ có những mục tiêu khác nhau, kiến trúc kỹ thuật số khác nhau và tác động khác nhau đến cuộc sống của mọi người. Những khác biệt này ảnh hưởng đến phần lớn thế giới, vì hầu hết các quốc gia đều dựa vào phần mềm của Trung Quốc và Hoa Kỳ hơn là công nghệ nội địa.

Hoa Kỳ cũng gây áp lực buộc các đồng minh và khách hàng của mình tránh xa phần cứng của Trung Quốc, chẳng hạn như cơ sở hạ tầng 5G của Huawei. Chính quyền Tổng thống Mỹ Donald Trump đã ngăn chặn nỗ lực của tập đoàn Broadcom của Singapore nhằm mua lại nhà sản xuất chip máy tính hàng đầu của Mỹ, Qualcomm. Họ lo ngại người nước ngoài có thể cài cửa hậu vào chip hoặc sẽ ngăn cản chính phủ Hoa Kỳ cài cửa hậu của riêng mình vào đó.

Cả chính quyền Tổng thống Mỹ Donald Trump và Chính quyền Joe Biden đều đặt ra những giới hạn nghiêm ngặt đối với việc buôn bán chip điện toán hiệu suất cao cần thiết cho sự phát triển của AI.

Các công ty Hoa Kỳ hiện bị cấm xuất khẩu những loại chip như vậy sang Trung Quốc. Mặc dù trong ngắn hạn, điều này cản trở Trung Quốc trong cuộc đua AI, nhưng về lâu dài, nó sẽ thúc đẩy Trung Quốc phát triển một lĩnh vực kỹ thuật số hoàn toàn riêng biệt, khác biệt với lĩnh vực kỹ thuật số của Mỹ ngay cả trong những tòa nhà nhỏ nhất của họ.

Do đó, hai lĩnh vực kỹ thuật số có thể ngày càng xa nhau hơn. Trong nhiều thế kỷ, các công nghệ thông tin mới đã thúc đẩy quá trình toàn cầu hóa và đưa mọi người trên khắp thế giới đến gần nhau hơn. Nghịch lý thay, ngày nay công nghệ thông tin mạnh mẽ đến mức có thể chia rẽ nhân loại bằng cách nhốt những người khác nhau vào những cái kén thông tin riêng biệt, chấm dứt ý tưởng về một thực tại chung của con người. Trong nhiều thập kỷ, ẩn dụ chính của thế giới là mạng lưới internet. Ẩn dụ chính của những thập kỷ tới có thể là cái kén.

Mặc dù Trung Quốc và Hoa Kỳ hiện đang dẫn đầu trong cuộc đua AI, nhưng họ không đơn độc. Các quốc gia hoặc khối khác, chẳng hạn như EU, Ấn Độ, Brazil và Nga, có thể đang cố gắng tạo ra những “kén kỹ thuật số” của riêng mình, mỗi quốc gia chịu ảnh hưởng bởi các truyền thống chính trị, văn hóa và tôn giáo khác nhau. Thay vì bị chia cắt giữa hai đế chế toàn cầu, thế giới có thể được chia thành hàng chục đế chế.

Các đế chế mới càng cạnh tranh với nhau, nguy cơ xung đột vũ trang càng lớn. Chiến tranh lạnh giữa Mỹ và Liên Xô chưa bao giờ leo thang thành một cuộc đối đầu quân sự trực tiếp, phần lớn là nhờ học thuyết hủy diệt lẫn nhau. Nhưng nguy cơ leo thang trong thời đại AI còn lớn hơn, bởi vì chiến tranh mạng về bản chất khác với chiến tranh hạt nhân.

Vũ khí mạng có thể đánh sập lưới điện của một quốc gia, nhưng chúng cũng có thể được sử dụng để phá hủy một cơ sở nghiên cứu bí mật, gây nhiễu cảm biến của đối phương, thổi bùng một vụ bê bối chính trị, thao túng bầu cử hoặc hack một chiếc điện thoại thông minh. Và chúng có thể làm tất cả những điều này một cách lén lút. Chúng không công bố sự hiện diện của mình bằng một đám mây hình nấm và một cơn bão lửa, cũng không để lại dấu vết rõ ràng từ bệ phóng đến mục tiêu. Do đó, thực sự đôi khi rất khó để biết liệu một cuộc tấn công có thể xảy ra hay không hoặc ai đã phát động nó. Do đó, sự cám dỗ để bắt đầu một cuộc chiến tranh mạng hạn chế là rất lớn, và sự cám dỗ để leo thang nó cũng vậy.

Một điểm khác biệt quan trọng thứ hai liên quan đến khả năng dự đoán. Chiến tranh lạnh giống như một ván cờ vua siêu lý trí, và sự chắc chắn về sự hủy diệt trong trường hợp xảy ra xung đột hạt nhân lớn đến mức mong muốn bắt đầu một cuộc chiến tranh cũng nhỏ bé tương ứng. Chiến tranh mạng thiếu sự chắc chắn này. Không ai biết chắc chắn mỗi bên đã đặt bom logic là các cuộc tấn công an ninh mạng tinh vi, phức tạp và gây ra thiệt hại rất lớn, ngựa thành Troy, một loại phần mềm độc hại đánh lừa người dùng về mục đích thực sự của nó bằng cách ngụy trang thành một chương trình bình thường và phần mềm độc hại ở đâu. Thực sự không ai có thể chắc chắn liệu vũ khí của họ có hiệu quả khi được sử dụng hay không. Liệu tên lửa của Trung Quốc có khai hỏa khi được lệnh hay không, hay có lẽ người Mỹ đã tấn công chúng hoặc chuỗi chỉ huy?

Liệu các tàu sân bay Mỹ có hoạt động như mong đợi, hay chúng sẽ tắt máy một cách bí ẩn hoặc đi vòng quanh?

Sự bất định như thế chắc chắn làm suy yếu học thuyết về sự hủy diệt lẫn nhau. Dù đúng hay sai, một bên có thể tự thuyết phục - rằng họ có thể phát động một cuộc tấn công phủ đầu thành công và tránh được sự trả đũa ồ ạt. Tệ hơn nữa, nếu một bên nghĩ rằng mình có cơ hội như thế, thì sự cám dỗ phát động một cuộc tấn công phủ đầu có thể trở nên không thể cưỡng lại, bởi vì người ta không bao giờ biết được cơ hội đó sẽ còn mở ra bao lâu. Lý thuyết trò chơi đặt ra rằng tình huống nguy hiểm nhất trong một cuộc chạy đua vũ trang là khi một bên cảm thấy mình có lợi thế nhưng lợi thế này đang dần mất đi.

Ngay cả khi nhân loại tránh được kịch bản tồi tệ nhất của chiến tranh toàn cầu, sự trỗi dậy của các đế chế kỹ thuật số mới vẫn có thể gây nguy hiểm cho tự do và sự thịnh vượng của hàng tỷ người. Các đế chế công nghiệp của thế kỷ XIX và XX đã bóc lột và đàn áp các thuộc địa của chúng, và sẽ là điều ngu ngốc nếu mong đợi các đế chế kỹ thuật số mới sẽ hành xử tốt hơn nhiều. Hơn nữa, nếu thế giới bị chia cắt thành các đế chế đối địch, nhân loại khó có thể hợp tác để vượt qua cuộc khủng hoảng sinh thái hoặc quản lý AI và các công nghệ đột phá khác như kỹ thuật sinh học.

Việc phân chia thế giới thành các đế chế kỹ thuật số đối địch trùng khớp với tầm nhìn chính trị của nhiều nhà lãnh đạo, trong những thập kỷ gần đây những người tin rằng thế giới là một khu rừng rậm, rằng hòa bình tương đối chỉ là ảo tưởng, và lựa chọn thực sự duy nhất là đóng vai kẻ săn mồi hay con mồi.

Với lựa chọn như thế, hầu hết các nhà lãnh đạo đều muốn đi vào lịch sử như những kẻ săn mồi và thêm tên mình vào danh sách đen những kẻ chinh phục mà những học sinh kém may mắn phải học thuộc lòng cho kỳ thi lịch sử. Tuy nhiên, những nhà lãnh đạo này nên được nhắc nhở rằng có một kẻ săn mồi đầu đàn mới trong rừng rậm. Nếu nhân loại không tìm ra cách hợp tác và bảo vệ lợi ích chung, tất cả chúng ta sẽ dễ dàng trở thành con mồi cho AI.

Đây là một trích đoạn đã được biên tập từ tác phẩm *“NEXUS: Lược sử về các mạng lưới thông tin từ Thời kỳ đồ đá đến kỷ nguyên Trí tuệ nhân tạo - NEXUS: A Brief History of Information Networks from the Stone Age to AI”* của Thiền giả Yuval Noah Harari, do Fern Press xuất bản vào ngày 10 tháng 9 năm 2024.

Cuốn sách *“Nexus: A Brief History of Information Networks from the Stone Age to AI”* của tác giả Yuval Noah Harari là một tác phẩm khám phá sự phát triển của mạng lưới thông tin từ thời kỳ sơ khai cho đến kỷ nguyên trí tuệ nhân tạo.

Nội dung chính của cuốn sách: Lịch sử phát triển của mạng lưới thông tin: Từ những câu chuyện truyền miệng đầu tiên của loài người, qua sự ra đời của chữ viết, in ấn, điện báo, điện thoại, cho đến sự bùng nổ của internet và công nghệ thông tin hiện đại.

Mối quan hệ giữa thông tin và quyền lực: Cuốn sách phân tích cách thức thông tin được sử dụng để xây dựng và duy trì quyền lực, từ các đế chế cổ đại đến các chính phủ hiện đại.

Tác động của mạng lưới thông tin đến xã hội: Harari khám phá những thay đổi sâu sắc mà mạng lưới thông tin gây ra cho các nền văn hóa, kinh tế và chính trị.

Thách thức và cơ hội trong kỷ nguyên thông tin: Cuốn sách đặt ra những câu hỏi về tương lai của loài người trong một thế giới được kết nối chặt chẽ bởi mạng lưới thông tin, bao gồm cả những tiềm năng và rủi ro.

Điểm nổi bật của cuốn sách:

Góc nhìn toàn cảnh: Harari cung cấp một bức tranh toàn diện về sự phát triển của mạng lưới thông tin từ quá khứ đến hiện tại.

Phân tích sâu sắc: Tác giả sử dụng kiến thức từ nhiều lĩnh vực như lịch sử, triết học, xã hội học và tâm lý học để phân tích các vấn đề phức tạp.

Văn phong hấp dẫn: Harari có khả năng truyền tải những thông tin phức tạp một cách dễ hiểu và cuốn hút.

Tầm nhìn về tương lai: Cuốn sách không chỉ dừng lại ở việc mô tả quá khứ mà còn đưa ra những dự đoán và suy ngẫm về tương lai.

Nexus là một cuốn sách quan trọng giúp độc giả hiểu rõ hơn về vai trò của thông tin trong lịch sử loài người và những tác động sâu rộng của nó đối với cuộc sống hiện đại.

Tác giả: **Thiền giả Yuval Noah Harari**

Viết dịch: **Sa môn Lê Văn Phước**

Nguồn: www.theguardian.com