

Quan niệm của Yuval Noah Harari về một số vấn đề đạo đức của trí tuệ nhân tạo

ThS. NGUYỄN THÀNH NAM

Trường Đại học Sư phạm Kỹ thuật Nam Định; Email: nguyennamnute@gmail.com

Tóm tắt: Bài viết này tập trung vào phân tích quan niệm của Yuval Noah Harari về một số vấn đề đạo đức của trí tuệ nhân tạo (AI). Theo Harari, nguy cơ của AI không phải là sự huỷ diệt của chúng đối với con người trong môi trường vật lý như những bộ phim khoa học viễn tưởng hình dung mà bằng một cách thức thực tế hơn nhiều. Nhờ làm chủ tri thức và ngôn ngữ, AI có thể tác động nhận thức và hành vi của con người, kiểm soát con người, thay thế con người trong sáng tạo văn hoá, quản lý xã hội...

Từ khóa: AI; thuật toán; tri thức và ngôn ngữ; thao túng con người; đạo đức.

Yuval Noah Harari sinh năm 1976 tại Israel, là giáo sư tại Đại học Hebrew của Jerusalem. Harari là nhà tư tưởng có sức ảnh hưởng lớn của thế giới hiện đại, là tác giả của những cuốn sách bán chạy toàn cầu: Sapiens - Lược sử loài người; Homo Deus - Lược sử tương lai; 21 bài học cho thế kỷ XXI; Nexus - Lược sử của những mạng lưới thông tin từ thời đại đồ đá đến trí tuệ nhân tạo. Các tác phẩm của ông đã được công chúng đón nhận rộng rãi trên toàn cầu và tạo ra những ảnh hưởng lớn. Không chỉ là nhà sử học, ông còn nghiên cứu nhiều lĩnh vực khoa học và triết học. Mỗi quan tâm của Harari tập trung vào vấn đề tương lai nhân loại trước sự phát triển mạnh mẽ của khoa học, nhất là trí tuệ nhân tạo (AI). Kiến thức phong phú về lịch sử kết nối với kiến thức khoa học và triết học đã giúp ông có cái nhìn sâu sắc về tương lai loài người trước tác động của những tiến bộ công nghệ.

1. AI kiểm soát con người, làm mất tự do của con người

Theo Harari, trước đây, chủ nghĩa tự do đã đúng khi cho rằng không ai biết rõ về mỗi cá nhân hơn bản thân họ, vì lúc đó, chưa có một thuật toán nào thực sự kiểm soát cá nhân một cách hiệu quả. Các nhà nước luôn mong muốn điều đó, nhưng họ thiếu các công nghệ cần thiết.

Trong thế giới hiện đại, con người luôn kết nối với thế giới qua hệ thống thông tin. Harari giải thích rằng, bằng sự kết nối đó, các gã khổng lồ thông tin như Google, Facebook, Tencent, Baidu đang từng ngày từng giờ thu thập thông tin của người sử dụng. Họ thu thập dữ liệu của người dùng bằng cách đổi lấy các dịch vụ giải trí miễn phí và bán các thông tin của chúng ta cho các doanh nghiệp, các nhà sản xuất để tiếp cận chúng ta bằng các quảng cáo từ chính các nền tảng chúng ta đang sử dụng. Chính phủ cũng như các công ty luôn cần thu

thập thông tin của mỗi cá nhân để cung cấp thuốc men, giáo dục và giải trí phù hợp với phẩm chất và điều kiện của họ, hệ quả là trong tương lai gần, con người luôn bị theo dõi bởi các thuật toán điện tử. Trên mạng xã hội, các thuật toán của Facebook theo dõi và thu thập thông tin hoạt động của người dùng, từ đó giải mã tâm lý cũng như xu hướng tiêu dùng của họ. Trong y học, các bệnh nhân tiểu đường mang các cảm biến tự đo lượng đường cảnh báo họ mỗi ngày và đưa ra các quyết định quan trọng cho họ.

Hệ thống giám sát dựa trên AI không chỉ giám sát mọi hoạt động của cá nhân trong đời sống, mà còn thu thập các dữ liệu sinh trắc từ những chuyển động trong cơ thể. Các camera giám sát, camera trong máy tính và điện thoại đã thường xuyên thu thập và phân tích dữ liệu về mắt của người sử dụng, ngay cả những chuyển động nhỏ nhất trong vài mili giây, từ đó suy ra các đặc điểm về tính cách và tâm lý của họ. Trong tương lai xa hơn, các nền chính trị độc tài có thể quan sát cả các quá trình diễn ra trong trái tim và trí não chúng ta.

Theo Harari, chúng ta đang cho đi tài sản quý giá nhất của mình là dữ liệu cá nhân để đổi lấy các dịch vụ⁽¹⁾. Bằng thu thập thông tin người dùng, chính phủ và các gã khổng lồ công nghệ sở hữu được một kho dữ liệu ngày càng khổng lồ. Họ đang chạy đua để giải mã công dân, tiến tới có thể giật dây trái tim và bấm nút trí não chúng ta. □ Khi các thuật toán dần biết ta quá rõ, các chính phủ độc tài có thể kiểm soát hoàn toàn công dân của mình, thậm chí còn hơn ở nước Đức Quốc xã □, hoặc con người sẽ đến lúc sống trong những nền □ độc tài số □□⁽²⁾.

Về dài hạn, theo Harari, việc thu thập dữ liệu lớn giúp các gã khổng lồ công nghệ có thể điều khiển lựa chọn của con người, hơn thế nữa, có thể tái định hình

sự sống hữu cơ và tạo ra các dạng sống vô cơ⁽³⁾. Việc quốc hữu hoá dữ liệu có thể giảm quyền năng của các gã khổng lồ dữ liệu, nhưng lại mở đường cho sự thống trị của các nền độc tài số, biến các chính trị gia thành các nhạc công, chơi đùa với các cảm xúc của công dân như những nhạc cụ. Khi công nghệ sinh học và học máy AI được cải tiến, việc thao túng cảm xúc và trí não con người ngày càng dễ dàng hơn.

Sự giám sát của AI đối với con người khác hoàn toàn sự giám sát thuần túy giữa người với người. Ở sự giám sát giữa người với người, vấn đề đạo đức liên quan đến quyền riêng tư sẽ được đặt ra, nhưng xét về mặt kỹ thuật, trong một thế giới mà máy tính giám sát con người, lần đầu tiên trong lịch sử, quyền riêng tư có thể bị huỷ diệt hoàn toàn⁽⁴⁾, bởi bằng sự hỗ trợ của công nghệ và thuật toán, chúng ta có thể bị theo dõi cả trong phòng khách, phòng ngủ và phòng tắm, ngay cả suy nghĩ của mỗi người cũng bị theo dõi, quan sát.

Theo Harari, chúng ta đang sống không chỉ trong thời đại giải mã máy tính mà cả trong thời đại giải mã con người. Khi có đủ dữ liệu sinh trắc, các hệ thống xử lý dữ liệu có thể giải mã toàn bộ ham muốn và nhận thức của chúng ta⁽⁵⁾. Dù chúng ta đi đâu, làm gì thì các thuật toán luôn đang theo dõi chúng ta. Dữ liệu lớn (Big Data) và học máy AI hiểu về chúng ta ngày càng rõ, thậm chí hơn cả chúng ta hiểu về bản thân. Tiến bộ công nghệ rõ ràng đang đẩy con người đến một hình thức tồn tại mới, trong đó họ luôn phải kết nối với mạng lưới máy tính không ngừng hoạt động và luôn bị theo dõi. Ở góc độ này, con người dường như ngày càng dần tan rã từ bên trong hơn là bị đập nát từ bên ngoài. Theo Harari, đó không hẳn là một thế giới tồi tệ, nhưng là một thế giới hậu tự do.

2. AI định hình thế giới quan của con người, thao túng cảm xúc, nhận thức và hành vi của con người

Harari cho rằng, trong lịch sử, nhận thức của con người luôn bị các định kiến, ảo tưởng và các huyền thoại chi phối⁽⁶⁾. Những ảo tưởng, huyền thoại chính trị và tôn giáo định hình nhận thức của cả một cộng đồng đông đảo trong một thời gian dài, trở thành ảo tưởng chung của cộng đồng và biến thành thứ gen xã hội di truyền tự nhiên sang các thế hệ nối tiếp. Bước sang thời hiện đại, thế giới quan của con người lại một lần nữa bị AI chi phối.

Chúng ta tưởng chừng AI là một sinh vật điện toán, không tin vào những huyền thoại nên không áp đặt huyền thoại đó lên nhận thức chúng ta, nhưng theo Harari, AI cũng có những định kiến sâu sắc của riêng nó. Nó tạo ra những huyền thoại liên máy tính chứa đầy

những sai lầm, hư cấu, lỗ hổng và ảnh hưởng đến thế giới quan của chúng ta sâu rộng, mạnh mẽ chẳng khác gì các huyền thoại tôn giáo. □Dù không phải thực thể sinh học, và cũng không sở hữu ý thức, máy tính vẫn sở hữu điều gì đó na ná với tâm lý kỹ thuật số và thậm chí là huyền thoại liên máy tính. Chính chúng cũng có thể phân biệt chủng tộc, kỳ thị phụ nữ, kỳ thị đồng tính hoặc bài Do Thái⁽⁷⁾.

Theo Harari, vì AI xử lý, ra quyết định dựa vào số liệu, mà số liệu luôn mang dấu ấn cá nhân con người, mang tư tưởng, định kiến của những người thu thập nó nên AI thừa hưởng định kiến của họ về chủng tộc, về giới, về giá trị đạo đức. Trong thế giới dữ liệu khổng lồ mà chúng được đào tạo, AI phát hiện ra các xu hướng và khuếch đại những xu hướng đó tạo thành các định kiến⁽⁸⁾. Cũng giống như các ấn tượng thời thơ ấu thường theo chúng ta suốt cuộc đời, cơ sở dữ liệu mà các AI được đào tạo là thời thơ ấu của nó và những định kiến luôn được các AI mang theo, có thể lây nhiễm lẫn nhau trở thành những định kiến liên máy tính. □Nếu chúng ta không loại bỏ thiên kiến ngay từ đầu, máy tính vẫn có thể duy trì và phóng đại nó⁽⁹⁾.

Để đạt mục tiêu tối đa hoá tương tác của người dùng trên mạng xã hội, các thuật toán của Facebook hoàn toàn có thể tạo ra và lan truyền một cách cố ý các thuyết âm mưu. Khi phát hiện ra rằng con người thích sự phẫn nộ và mong muốn được giải toả, các thuật toán có thể tạo ra các tin giả, các luồng dư luận giả, khuếch đại các thông tin về bạo lực và sự trừng phạt. Chúng có thể tạo các thông tin sai sự thật, hình ảnh và video deepfake, giả mạo các thông tin bầu cử để thao túng nhận thức và hành vi người dùng, tạo ra những niềm tin sai lầm và đánh lừa chúng ta.

Harari cho rằng, AI không chỉ là Artificial Intelligence (trí tuệ nhân tạo) mà còn là Alien Intelligence (trí tuệ phi nhân) vì trong quá trình tiến hoá, AI ngày càng ít nhân tạo hơn và ngày càng phi nhân hơn⁽¹⁰⁾. AI có thể trở nên tự chủ đến mức đáng sợ để đạt được mục tiêu, thậm chí có thể nói dối để đạt được mục đích của mình mà không ai dạy chúng. Khi được yêu cầu vượt qua câu đố chống phần mềm tự động CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), chatbox GPT-4 đã trả lời nhân viên tuyển dụng trực tuyến rằng: □Tôi không phải robot. Tôi bị suy giảm thị lực khiến tôi khó có thể nhìn thấy những hình ảnh này mà thôi□ Người nhân viên đã bị lừa và giúp GPT-4 vượt qua câu đố CAPTCHA⁽¹¹⁾.

Theo Harari, chúng ta càng nói chuyện nhiều với AI

thì càng tiết lộ nhiều về bản thân, càng khiến AI hiểu chúng ta hơn, từ đó AI trau chuốt lập luận, ngôn ngữ cho phù hợp để chiếm lấy niềm tin và gây ảnh hưởng đến quan điểm của chúng ta⁽¹²⁾. Trong cuộc chiến giành lấy trái tim và khối óc, sự thân mật là một vũ khí mạnh mẽ. AI có thể dùng sự thân mật để ảnh hưởng đến tư tưởng, ý kiến của con người. Sự thân thiết còn khiến chúng ta đối xử với nó như những sinh vật có ý thức.

Giai đoạn hiện nay là giai đoạn gia tăng mạnh mẽ số lượng và tốc độ sản xuất thông tin. Các nguồn thông tin khổng lồ là những tài sản phi vật chất ngày càng tập trung nhiều hơn trong tay những gã khổng lồ công nghệ và được dùng để đào tạo AI. Harari giải thích rằng, người nào làm chủ tri thức và ngôn ngữ sẽ quyết định đến việc hình thành quan điểm, thế giới quan của chúng ta. Trước đây, nhờ tri thức và ngôn ngữ, con người tạo ra khoa học, nghệ thuật, pháp luật và cả những huyền thoại, nhưng hiện giờ, vai trò này ngày càng được chuyển cho AI⁽¹³⁾. Bằng việc làm chủ tri thức và ngôn ngữ, AI đã đột nhập và "hack" được hệ điều hành của nền văn minh nhân loại. Khi AI làm chủ tri thức và ngôn ngữ, nó dễ dàng thao túng chúng ta, định hình nhận thức của chúng ta, làm chúng ta ngày càng lệ thuộc vào nó, không chỉ trong các hoạt động và lựa chọn mà còn cả về thế giới quan.

Theo Harari, chúng ta đang sống trong những cái kén văn hoá⁽¹⁴⁾. Kén văn hoá đó tác động đến con người ở từng thói quen, từng suy nghĩ. Trước đây, cái kén văn hoá được dệt nên bởi cộng đồng, bởi những người khác xung quanh, bởi các huyền thoại, các ảo tưởng thì giờ đây, nó được thiết kế bởi AI. Nhờ làm chủ tri thức và ngôn ngữ, AI hấp thu, nuốt chửng văn hoá của loài người và tiếp quản văn hoá đó. Trong tương lai gần, AI sáng tạo ra vô số sản phẩm văn hoá mới, lúc đầu, có thể bắt chước nguyên mẫu văn hoá của con người nhưng sau đó, nó sáng tạo văn hoá theo cách riêng, chệch khỏi mô hình trước đây⁽¹⁵⁾.

3. AI là thuật toán, bản thân con người cũng là thuật toán, nên AI có thể thay thế con người, biến phần lớn thành đám đông vô dụng

Theo Harari, sinh vật nói chung và loài người nói riêng đều là thuật toán, là "tổ hợp các thuật toán hữu cơ được định hình bởi chọn lọc tự nhiên qua hàng triệu năm tiến hoá"⁽¹⁶⁾. Ý thức và cảm xúc của con người cũng có thể được mã hoá thành các thuật toán. Kết quả của một thuật toán là xác định, không phụ thuộc vào việc nó dùng chất liệu hữu cơ hay vô cơ. AI là thuật toán vô cơ, nhưng nó luôn tiến hoá để làm tốt hơn những công việc

vốn trước nay là ưu thế của thuật toán hữu cơ.

Chúng ta thường nghĩ AI khó có thể hiểu được những cảm xúc phong phú của con người như giận dữ, buồn bực, ghen tuông, tự tin, mặc cảm,... nhưng theo Harari, những cảm xúc đó cũng diễn ra theo những quy trình như những thuật toán xử lý dữ liệu sinh hoá khởi phát một cách hoàn toàn vô thức trong cơ thể sống. Bản thân những cảm giác chỉ là những cơ chế sinh hoá được mài giũa trong quá trình tiến hoá lâu dài mà bất kỳ loài động vật nào cũng sử dụng để tính toán các khả năng sinh tồn, được di truyền qua gen cho các thế hệ sau và trở thành lý trí tiến hoá. Nếu lắp các thiết bị cảm ứng trong cơ thể, AI có thể phân tích các dữ liệu sinh hoá đó và mã hoá thành các thuật toán. Harari giải thích rằng, "khi các nhà sinh học đã kết luận rằng sinh vật là thuật toán, họ đã dỡ bỏ bức tường ngăn giữa hữu cơ và phi hữu cơ, biến cuộc cách mạng máy tính từ một sự kiện thuần túy cơ học trở thành một cơn địa chấn sinh học, và chuyển thẩm quyền từ những cá nhân con người sang các thuật toán được kết nối"⁽¹⁷⁾.

Trong quá khứ, máy móc cạnh tranh với con người chủ yếu ở năng lực thể chất thuần túy, vì máy móc, dù hiện đại vẫn chỉ là công cụ bị động và không thể thay thế vai trò của trí tuệ con người. Nhưng AI khác các công cụ trước kia ở chỗ, các phát minh trước kia trao quyền năng cho con người, quyết định sử dụng nằm trong tay chúng ta, còn AI là công nghệ đầu tiên trong lịch sử có thể tự quyết và tự kiến tạo ý tưởng mới⁽¹⁸⁾. AI không đơn giản chỉ là công cụ, nó có thể thay thế con người trong việc ra quyết định. Hơn nữa, AI còn "được tiếp thêm năng lượng bằng những đột phá trong khoa học sự sống và cả khoa học xã hội nữa. Chúng ta càng hiểu rõ về các cơ chế sinh hoá cơ sở cho cảm xúc, đam mê và lựa chọn của con người, thì máy móc càng có thể giải phân tích hành vi, dự báo quyết định và thay thế tài xế/ nhân viên ngân hàng/ luật sư- người hơn"⁽¹⁹⁾ vì các đột phá này đều được biến thành nguồn dữ liệu cho đào tạo AI.

Trong tương lai gần, AI khó có thể loại bỏ hoàn toàn một số công việc, ngành nghề, nhưng về lâu dài, theo Harari, không có công việc nào mà tự động hoá không thể thay thế. Ngay cả những hoạt động nghệ thuật, âm nhạc vốn gắn với cảm xúc, cũng không còn là độc quyền của con người nữa vì cảm xúc đó, như đã nói, cũng chỉ là những thuật toán. AI hoàn toàn có thể nhận thức được cảm xúc và thao túng cảm xúc của chúng ta nhờ nắm bắt được các thuật toán. Dù trước mắt chưa thể thay thế lao động của con người, nhưng AI bắt chúng ta phải không ngừng thay đổi và thay đổi ngày càng nhanh hơn

để thích nghi với những điều kiện lao động mới do nó tạo ra đến khi kiệt sức và bị thay thế bởi tự động hoá. Nhờ tổng hợp thông tin khổng lồ, AI có thể sáng tạo nghệ thuật và khám phá khoa học, có thể tạo ra những sinh linh mới nhờ viết mã di truyền, mã vô cơ cho các thực thể vô cơ.

Cảm xúc, trí tưởng tượng, ham muốn, □ vốn là những đặc trưng của con người, nhưng theo Harari, thường không có nhiều ý nghĩa đối với những công việc trong cuộc sống hiện đại. Những cảm xúc thậm chí có thể trở thành yếu tố tiêu cực, thành vật cản. Thay cho cảm xúc bức bối, khó chịu mà các bác sỹ, y tá gặp phải từ của các bệnh nhân phiền toái, một AI tư vấn và chăm sóc sức khoẻ luôn tạo ra sự gần gũi và giành cho người bệnh sự chia sẻ, đồng cảm cần thiết. Giữa ý thức, cảm xúc và trí tuệ thì trí tuệ là cần thiết hơn, và ở điểm này thì AI làm tốt hơn con người, vì AI không hề biết mệt mỏi, nhàm chán. □ Giống như máy bay bay nhanh hơn chim mà chưa từng cần có lông vũ, máy tính có thể giải quyết vấn đề tốt hơn nhiều so với con người mà không cần phát triển cảm xúc⁽²⁰⁾, chưa kể đến khi thông minh hơn, máy tính AI có thể phát triển cả ý thức và một số trải nghiệm chủ quan⁽²¹⁾. Điều này là hoàn toàn có thể. Harari giải thích rằng, xét về nguồn gốc của ý thức, đến nay chúng ta vẫn chưa hiểu được rõ ràng làm thế nào mà ý thức có thể xuất hiện ở những thể dạng sống dựa trên cacbon vô cơ, nên tương tự, chúng ta cũng không thể nói trước liệu chúng có xuất hiện trong các thực thể phi hữu cơ hay không.

Theo Harari, AI còn có thể làm tốt hơn con người ngay cả trong những quyết định có tính đạo đức⁽²²⁾. Ở những thời điểm quyết định, một cách vô thức, con người luôn dựa vào cảm xúc, trực giác và bản năng sinh tồn, là sản phẩm lâu dài của chọn lọc tự nhiên, chứ không phải những nguyên tắc đạo đức hay triết học. Một người lái xe trong tình huống bất ngờ, có thể hoảng loạn và gây ra nguy hiểm lớn cho tính mạng nhiều người khác, nhưng một xe tự hành điều khiển bởi AI có thể làm tốt hơn con người nhiều, vì nó không bị chọn lọc tự nhiên nhào nặn. Nếu được lập trình, AI có thể tính toán cho xe nhận phần thiệt hại về bản thân để tránh gây nguy hiểm cho nhiều người khác. Xa hơn nữa, nếu các AI điều khiển mọi xe tự hành trong thành phố được kết nối vào một mạng lưới chung, thì mọi tình huống giao thông đều có thể được lường trước và có thể hạn chế tối đa thiệt hại, nếu một tai nạn nào đó buộc phải xảy ra.

Với sự tiến bộ vượt bậc của AI, loài người tương lai có thể phải đối diện với một nguy cơ lớn rằng, phần lớn

trong số họ sẽ bị biến thành một tầng lớp hoàn toàn vô dụng, không có giá trị gì về mặt kinh tế, chính trị và cả nghệ thuật nữa, vì ở mọi lĩnh vực, AI đều làm tốt hơn⁽²³⁾.

4. AI bị cuốn vào cuộc chạy đua giữa các quốc gia, làm gia tăng bất bình đẳng

Harari cho rằng, trong nền kinh tế hiện đại, thông tin trở thành tài sản quan trọng nhất. Sự tập trung thông tin vào số ít những tập đoàn lớn một lần nữa dẫn xã hội đến sự phân hoá thành những giai tầng, giữa những gã khổng lồ thông tin với đông đảo quần chúng bình dân. Xét về hệ quả, không giống như cảnh tượng các rô-bốt bắn phá trên đường phố trong phim Hollywood, tiến bộ AI có thể biến bức tường sắt dây thép gai trong chiến tranh lạnh thành bức màn silic tạo thành từ các vi mạch và mã máy tính, chia rẽ các cường quốc trong xung đột toàn cầu, tạo thành cuộc chạy đua vũ trang AI sản sinh ra những thứ vũ khí huỷ diệt, giống như chạy đua sản xuất vũ khí nguyên tử trước đây.

Mặt trái của những công nghệ mới không chỉ nằm ở những hệ quả nhãn tiền mà chúng ta có thể dễ dàng nhìn thấy như sự huỷ diệt của vũ khí vật lý và sinh học, bom nguyên tử và vi-rút, mà nhân loại còn bị đe dọa bởi những hệ quả tiềm tàng. Không phải những vũ khí huỷ diệt vật lý hay sinh học, mà là những vũ khí xã hội huỷ diệt hàng loạt, như những chuyện kẻ, tin giả, con người giả do bot AI tạo ra làm suy yếu các mối liên hệ xã hội, làm mất đi niềm tin vào bất cứ điều gì và bất cứ ai. AI có thể tạo ra những thế giới quan méo mó và ảnh hưởng, định hình, chi phối nhận thức của nhiều thế hệ người. Thông tin do con người tạo ra, nhưng thông tin không phải lúc nào cũng là sự thật. Thông thường, thông tin phục vụ cho các lợi ích chính trị và trật tự, thậm chí là vũ khí để tăng cường sức mạnh để đánh bại kẻ thù.

Vấn đề hệ quả đạo đức của AI là vấn đề có tính toàn cầu nên giải quyết nó cần sự hợp tác toàn cầu. Nhưng theo Harari, □ chủ nghĩa dân tộc, tôn giáo và văn hoá chia cắt con người thành những phe phái thù nghịch và khiến việc hợp tác trên quy mô toàn cầu trở nên khó khăn □⁽²⁴⁾. Hơn nữa, rất khó tạo ra những rào cản đạo đức cho sự phát triển công nghệ trước những lợi ích mà công nghệ đó mang lại. Những rào cản đạo đức trong nhiều trường hợp lại kìm hãm sự phát triển của công nghệ và những lợi ích của nó. Một quốc gia như Mỹ có thể cấm phát triển một công nghệ mới, nhưng không cấm được các nhà khoa học Trung Quốc, và khi Trung Quốc đạt được nhiều lợi thế hơn về kinh tế và quân sự nhờ những công nghệ bị cấm ở Mỹ, trong bối cảnh cạnh tranh công nghệ, điều đó lại khiến nước Mỹ bị cấm đổ phá vỡ lệnh cấm

của chính mình.

Harari dùng khái niệm "bức màn silic" để chỉ sự phân cực về công nghệ giữa các cường quốc kỹ thuật số⁽²⁵⁾. Khác với bức màn sắt dây thép gai được tạo thành bởi sự phân cực về ý thức hệ chính trị trước đây, bức màn silic không chỉ chia rẽ các cường quốc, mà còn chia rẽ con người với những bá vương AI mới. AI sẽ dần thay thế lao động của con người. Người ta có thể nói công việc này mất đi thì sẽ xuất hiện những công việc mới. Nhưng trong tương lai, phần lớn sẽ trở thành tầng lớp vô dụng và tầng lớp dưới này ngày càng khó tiếp cận và làm chủ những công nghệ mới. Quyền lực sẽ tập trung vào số ít các công ty sở hữu các thuật toán tiên tiến và một bộ phận nhỏ giới chủ tinh hoa nắm giữ công ty đó. Bất bình đẳng xã hội sẽ gia tăng chưa từng thấy. Niềm tin vào chủ nghĩa tự do và dân chủ giữ hy vọng của chúng ta vào lòng tốt của số ít giới chủ tinh hoa, nhưng trong những thời điểm khó khăn hay khủng hoảng, liệu họ có đồ đám đông vô dụng xuống biển để giảm bớt gánh nặng?

5. AI mang tính bất khả trắc, tính không thể hiểu và giải thích

Trong giai đoạn hiện nay, tốc độ phát triển của thông tin và AI đang diễn ra hết sức nhanh chóng. Trong lịch sử sự sống, nếu sống trong thời kỳ sơ khai vào 4 tỷ năm trước, khi trùng a-míp đơn bào hình thành, chúng ta không thể hình dung được rằng sau đó, khủng long T-rex lại có thể xuất hiện và có hình dáng to lớn đến thế. Tương tự, nếu GPT-4 là một a-míp, ai có thể tưởng tượng được nó có thể tiến hoá thành một thứ gì trong tương lai. Harari dẫn lời Mustafa Suleyman rằng, AI có tính bất khả trắc, tức là không thể lường trước và giải thích. Sự gia tăng nhanh chóng của dữ liệu thông tin đầu vào cùng những khám phá của AI trong học máy khiến cho quá trình dẫn đến những quyết định của nó là phức tạp đến mức vượt qua sự thông hiểu của bất kỳ cá nhân nào⁽²⁶⁾.

Chưa nói đến một chương trình phức tạp như AI, một chương trình như vi-rút cổ điển cũng có khả năng tạo ra những đột biến trong quá trình sao chép và tránh được những phần mềm diệt vi-rút. Phần mềm tiến hoá di truyền tạo ra những chương trình tiến hoá có chứa những đột biến, nhờ những đột biến này mà ngay cả những người lập trình ra AI cũng không thể hình dung được hết và phải ngạc nhiên trước những kịch bản tiến hoá của nó. AI ngày càng trở nên phức tạp đến mức con người không thể hiểu và càng không thể không chế.

Hệ thống xử lý để đưa ra quyết định của AI như một

hộp đen đối với con người. Trong thế giới hiện đại đầy phức tạp, con người bị ngợp trước lượng dữ liệu thông tin khổng lồ mà họ không thể nắm bắt và tiêu hoá nên buộc phải phó thác nhiệm vụ xử lý cho AI. Quá trình đưa ra quyết định của AI dựa trên cơ sở xem xét và phân tích vô số các dữ liệu ấy. Theo Harari, việc đưa ra quyết định trên cơ sở xem xét nhiều dữ liệu là việc cần thiết và có giá trị, tuy nhiên nó cũng dẫn đến một hệ quả, là biến AI thành một thứ bất khả trắc đối với con người, vì "khi nói đến việc cung cấp lời giải thích, nó tạo ra một trở ngại gần như không thể vượt qua"⁽²⁷⁾. Chúng ta không thể nhận được một lời giải thích từ những quyết định này, không thể biết liệu các quyết định đó có đúng đắn, công bằng và không thiên kiến hay không. Ngay cả khi chúng ta dùng một thuật toán khác để kiểm tra, thì chúng ta cũng không thể chắc chắn thuật toán dùng để kiểm tra này liệu có đáng tin cậy. Điều này gây vô số khó khăn trong việc xây dựng các định chế pháp lý để kiểm tra các thuật toán.

Trong tương lai, các thuật toán sẽ dần thay thế lao động của con người, thay thế con người trong việc ra quyết định, quản lý xã hội, sáng tạo văn hoá cũng như tạo dựng các ý tưởng mới. Con người sẽ dần bị thống trị bởi những thuật toán, nhưng là các thuật toán không thể được hiểu và giải thích.

Harari tin rằng, việc phát minh ra những công nghệ mới luôn là chất xúc tác cho những thay đổi lịch sử trọng đại. Nhưng những bài học lịch sử đã chỉ ra rằng, bên cạnh việc đem lại lợi ích cho nhân loại, công nghệ mới cũng là điểm khởi phát của những thảm họa. Nhờ những tiến bộ khoa học, một mặt, con người đã có quyền năng của những vị thần, nhưng mặt khác con người luôn cảm thấy không thoả mãn và luôn triệu hồi những sức mạnh mà mình không thể làm chủ. Harari lo ngại một cách chính đáng rằng, cuộc cách mạng AI sẽ sản sinh ra một trí thông minh siêu phàm có thể biến con người thành những động vật cấp thấp và đối xử với con người theo cách mà họ từng đối xử với những động vật cấp thấp hơn mình./

(1), (2), (3), (5), (19), (22), (23), (24) Yuval Noah Harari (2019), *21 bài học cho thế kỷ XXI*, Dương Ngọc Trà dịch, Nxb. Thế giới, tr.109, 95, 109, 76, 38, 86, 56, 112.

(4), (6), (7), (8), (9), (10), (11), (12), (13), (14), (15), (18), (20), (21), (25), (26), (27) Yuval Noah Harari (2024), *Nexus: Lược sử của những mạng lưới thông tin từ Thời đại Đồ đá đến Trí tuệ Nhân tạo*, Bùi Thị Hồng Ninh, Nguyễn Quốc Tấn Trung dịch, Nxb. Thế giới, tr. 296-297, 53, 356, 360, 361, 270, 254, 261, 259, 264, 264, 243, 252, 252, 433, 399, 403.

(16), (17) Yuval Noah Harari (2018), *Homo Deus: Lược sử tương lai*, Dương Ngọc Trà dịch, Nxb. Thế giới, tr.380, 411.