

RANH GIỚI GIỮA NGƯỜI VÀ MÁY TRONG KỶ NGUYÊN TRÍ TUỆ NHÂN TẠO – GÓC NHÌN TỪ TRIẾT LÝ HIỆN SINH JEAN-PAUL SARTRE

TRẦN THỊ THẢO*

Song song với việc tạo ra bước tiến vượt bậc mang tính cách mạng về khoa học kỹ thuật hướng tới phục vụ loài người, trí tuệ nhân tạo cũng đặt ra những câu hỏi lớn về ranh giới giữa con người và máy móc trong việc sử dụng và phát triển công nghệ này. Điều đó hoàn toàn mâu thuẫn với triết lý hiện sinh của Jean-Paul Sartre khi nhấn mạnh tầm quan trọng của tự do cá nhân và trách nhiệm của con người trong quá trình định hình nhân vị. Thông qua góc nhìn hiện sinh không tách rời tính biện chứng đa ngành, nghiên cứu xem xét những tác động ảnh hưởng trực tiếp đến góc cạnh dường như được cho là bất khả xâm phạm về quyền riêng tư của mỗi cá nhân, những chuyển hóa nhân quả về phương diện đạo đức xã hội khi con người dần phụ thuộc vào máy móc và trí tuệ nhân tạo.

Từ khóa: tự do, trí tuệ nhân tạo (AI), hiện sinh, Jean-Paul Sartre

Nhận bài ngày: 30/10/2023; *đưa vào biên tập:* 03/11/2023; *phản biện:* 08/12/2023; *duyet đăng:* 11/01/2024

1. DẪN NHẬP

Từ những phát minh mang tính đột phá khai thác sức mạnh của hơi nước để điều khiển những khối kim loại vô tri phục vụ cuộc sống, con người dần thể hiện quyền năng làm chủ thế giới vạn vật thông qua nhiều cuộc cách mạng công nghiệp liên tiếp. Với trí tuệ nhân tạo (AI), con người xây dựng hệ

sinh thái tạo bởi máy móc tự động và công nghệ siêu kết nối, vượt qua những giới hạn về vật chất và không gian; khẳng định vai trò chủ thể với những công nghệ có thể hiểu và phục tùng. Tuy nhiên, sự ra đời của AI đang hình thành một bối cảnh mới, trong đó mối quan hệ giữa sự tự do, tính chủ thể của con người với thế giới AI dần trở nên phức tạp. Điều đó đặt ra nghi vấn về vai trò hỗ trợ và những nguy cơ tiềm ẩn đe dọa đến chủ thể lịch sử,

* Trường Đại học Kinh tế - Tài chính Thành phố Hồ Chí Minh.

đến quyền riêng tư và sự tự do cá nhân; về những thách thức đối với trách nhiệm và ý thức đạo đức trong việc sử dụng, phát triển AI.

2. TRÍ TUỆ NHÂN TẠO VÀ SỰ THÁCH THỨC ĐẾN TỰ DO CÁ NHÂN

Theo Tổ chức Hợp tác và Phát triển Kinh tế (OECD) chưa có một định nghĩa về AI mang tính khái quát được chấp nhận một cách rộng rãi (OECD, 2019: 6). AI được các học giả định nghĩa từ một số khía cạnh tiếp cận như “khả năng của một hệ thống kỹ thuật số hoặc robot được điều khiển bởi tập hợp thuật toán để thực hiện các nhiệm vụ, thường được liên tưởng với những sinh vật thông minh (như con người) về khả năng suy luận, khám phá, khái quát hóa,... rút ra từ kinh nghiệm quá khứ hoặc hệ cơ sở dữ liệu được nạp” (Copeland, 2023); “AI là một lĩnh vực nghiên cứu nhằm giải thích và mô phỏng hành vi thông minh dưới dạng các quá trình tính toán” (Schalkoff, 1990: 2). Về mặt chức năng, khái niệm AI đề cập đến công nghệ thực hiện những nhiệm vụ đòi hỏi yêu cầu xử lý thông minh, tương tự như cách mà con người sử dụng tư duy để thực hiện chúng (Boden, 2016: 1); “Nhiều hoạt động tinh thần của con người như viết chương trình máy tính, làm toán, tham gia vào các lý luận thông thường, hiểu ngôn ngữ và thậm chí lái xe ô tô được cho là đòi hỏi “trí thông minh” [...] Chúng ta có thể nói rằng những hệ thống như vậy sở hữu trí tuệ nhân tạo ở một mức độ nào đó” (Nilsson, 1998:

1). Theo các chuyên gia cao cấp về AI của Liên minh Châu Âu (EU) thì “Trí tuệ nhân tạo (AI) là thuật ngữ dùng để chỉ các hệ thống được con người thiết kế, và chúng, khi được giao một nhiệm vụ phức tạp, thực hiện trong thế giới vật lý hoặc số hóa bằng cách cảm nhận môi trường xung quanh, hiểu và phân tích dữ liệu đã thu thập, dựa vào kiến thức được rút ra từ dữ liệu này, quyết định các hành động tốt nhất để thực hiện (theo các tham số được định nghĩa trước) nhằm đạt được mục tiêu đã đề ra. Các hệ thống AI cũng có thể được thiết kế để học và thích nghi với hành vi của chúng bằng cách phân tích cách môi trường bị ảnh hưởng bởi các hành động mà chúng đã thực hiện trước đó” (AI HLEG, 2018: 7).

Do đó, AI là một thuật ngữ chung mô tả các quy trình, về bản chất, ủy quyền một phần hoặc toàn bộ các hoạt động đưa ra và thực hiện các quyết định từ con người sang hệ thống phần mềm. Nó được sử dụng trong hệ thống phần mềm hoạt động trong thế giới ảo và bao gồm phần mềm công cụ tìm kiếm, phân tích hình ảnh, trợ lý ảo, giọng nói hoặc hệ thống nhận dạng khuôn mặt. AI cũng có thể được cài đặt trong các hệ thống phần cứng như ô tô tự hành, robot, hoặc máy bay không người lái (AI HLEG, 2018: 1).

Lịch sử phát triển AI xuất phát từ những năm 1940, khi các nhà khoa học đầu tiên bắt đầu nghiên cứu về khái niệm máy tính thông minh. Trong

thập niên 1950, John McCarthy đặt ra thuật ngữ “trí tuệ nhân tạo” (Artificial Intelligence - AI) để mô tả việc phát triển các hệ thống máy tính có khả năng xử lý và tư duy như con người (McCarthy, 2006: 13). Đến những năm 1960 và 1970, các nhà khoa học đã phát triển các thuật toán và công nghệ mới, đưa AI trở thành một lĩnh vực liên ngành. Trong thập niên 1980, 1990, AI đã bị đánh giá thấp khi các nhà khoa học gặp nhiều khó khăn trong việc giải quyết các thuật toán liên quan đến khả năng suy luận và tự học của hệ thống AI. Tuy nhiên, những năm tiếp theo, nền tảng công nghệ và các thuật toán phục vụ AI đã được phát triển và đa dạng hóa ở mức độ cao hơn. Từ các hệ thống nguyên lý cơ sở (rule-based systems)⁽¹⁾, AI đã có bước tiến vượt bậc với sự ra đời của các khái niệm như máy tự học (machine learning)⁽²⁾, mạng lưới nơ-ron nhân tạo (neural networks)⁽³⁾ và các thuật toán tiến hóa (evolutionary algorithms). Điều này tạo ra cơ sở cho việc xây dựng các ứng dụng AI đa dạng, từ hệ thống nhận dạng giọng nói, xe tự hành, hệ thống gợi ý, cho đến robot thông minh và trợ lý ảo. Mục đích chính của AI là “bắt chước chức năng nhận thức của con người và thực hiện các hoạt động thường được thực hiện bởi con người, bằng cách sử dụng kinh nghiệm của chính nó để giải quyết vấn đề” (Sadiku, 1989: 39).

AI đã đạt được những thành tựu đáng kể và trở thành một trong những khía

cạnh quan trọng nhất của sự tiến bộ nhân loại. Trước đây, con người là loài có khả năng nhận thức bậc cao nhất, nhưng hiện nay “AI đã phát triển đến một giai đoạn mà nó có thể sánh ngang và thậm chí vượt qua con người trong nhiều nhiệm vụ liên quan đến nhận thức” (Nowak và các cộng sự, 2018: 26).

Sự trỗi dậy của AI trong những năm gần đây đã đưa đến những quan điểm trái chiều. Trong khi Giám đốc điều hành của IBM, Ginni Rometty (2017), cho rằng AI là “công nghệ gia tăng trí tuệ con người... Chúng ta đang chứng kiến một thế giới nơi con người và máy móc hợp tác, mang lại những lợi ích và cho phép chúng ta thực hiện tốt nhất những gì mà con người có thể làm”; Stephen Hawking lại nhận xét rằng “sự phát triển của AI hoàn chỉnh có thể đánh dấu sự kết thúc của loài người” (Cellan-Jones, 2014). Đồng quan điểm, Bill Gates cũng nói rằng con người nên lo lắng về sự đe dọa của trí tuệ nhân tạo (Rawlinson, 2015). Những dự báo trên của các chuyên gia hàng đầu về công nghệ đã buộc chúng ta phải suy nghĩ một cách thận trọng về những can dự quá sâu của AI đối với cuộc sống con người.

Trong khi đó, tự do lại là một khía cạnh quan trọng đối với mỗi cá thể con người. Từ lâu, khái niệm phức tạp này đã thu hút sự quan tâm của toàn xã hội, đặc biệt là sự tranh luận của nhiều triết gia qua các giai đoạn lịch sử. Vì vậy, khi quyền con người có nguy cơ bị xâm phạm, con người cần

phải nhận thức và điều chỉnh hành vi để bảo toàn trạng thái. Điều này đã được Jean - Paul Sartre⁽⁴⁾ cảnh báo từ giữa thế kỷ XX khi khoa học công nghệ ngày càng phát triển, con người ngày càng phụ thuộc vào máy móc mà đánh mất dần tự do của bản thân.

3. SỰ ẢNH HƯỞNG CỦA TRÍ TUỆ NHÂN TẠO ĐẾN TỰ DO CON NGƯỜI QUA LĂNG KÍNH TRIẾT HỌC HIỆN SINH

3.1. Sự can thiệp của trí tuệ nhân tạo vào quá trình lựa chọn và ra quyết định của con người

Triết lý hiện sinh của Sartre (1992a: 530) đề cao con người, trao quyền ra quyết định lựa chọn của con người dựa trên quyền tự do cá nhân và ý chí riêng của mỗi người. Đối với chủ nghĩa hiện sinh, con người là cá thể tự chủ có khả năng định hình và chọn lựa cuộc sống của mình theo ý muốn và giá trị cá nhân đồng thời chịu trách nhiệm cho sự lựa chọn đó. “Con người không là gì khác ngoài những gì mà nó tự tạo nên. Đó là nguyên tắc đầu tiên của thuyết hiện sinh” (Sartre, 2015: 26). Con người muốn tồn tại phải đặt mọi sự phát minh dưới nguyên tắc này. Đối với lĩnh vực AI, trong chiến lược phát triển ban đầu, loài người chủ trương đề xuất những nghiên cứu khai thác nó để hỗ trợ cho con người ra quyết định tối ưu hơn. Mc Corduck và các cộng sự (2004) đã tổng kết: “Sự tiến bộ của máy móc trong một số nhiệm vụ đã vượt trội hơn con người chỉ trong vài năm qua, và điều này đã thể hiện rõ qua các

thành tựu thực tế được dự đoán rộng rãi”. Theo Bader, J và các cộng sự (1988: 271), AI sử dụng trong các hệ thống có chức năng xử lý thông minh được phân loại theo sáu vai trò khác nhau: trợ lý, nhà phê bình, ý kiến tham khảo, chuyên gia cố vấn, người hướng dẫn và hệ thống tự động. Từ các vai trò đó, “trí tuệ nhân tạo được áp dụng rộng rãi trong việc hỗ trợ ra quyết định và trở thành một trong những ứng dụng quan trọng nhất của lịch sử AI” (Duan và các cộng sự, 2019: 67). Các ứng dụng của AI được sử dụng trong việc giúp con người lựa chọn để ra quyết định hành động trong thời gian thực, đặc biệt các quyết định có tính chất phức tạp, giảm tải việc xử lý thông tin, tạo tương tác và quản lý rủi ro. Nhờ những tiến bộ này, AI mang lại nhiều cơ hội và tiềm năng trong việc tối ưu hóa quá trình ra quyết định của con người (Phillips và Jain, 2006: 531-536). Hàng loạt các công cụ AI có thể được kể tên như Neural networks⁽⁵⁾, Fuzzy Logic⁽⁶⁾, Intelligent Agents⁽⁷⁾, Case-Based Reasoning⁽⁸⁾,... khi được kết hợp với các thuật toán hỗ trợ ra quyết định sẽ giúp con người đánh giá và lựa chọn quyết định tối ưu hay các giải pháp thay thế (Phillips, 2012: 2).

Tuy nhiên, cùng với những bước tiến vượt bậc về khoa học công nghệ, xuất hiện nguy cơ can thiệp của AI vào các quyết định của con người. Một trong những rủi ro nghiêm trọng nhất là con người dần mất đi sự kiểm soát đối với hệ thống và phai mờ trách nhiệm cá

nhân. Khi AI ngày càng thông minh và có khả năng tự học, có thể tự đưa ra quyết định mà không cần tới sự can thiệp của con người, thì “ngày càng có nhiều hệ thống AI vượt ra ngoài vai trò là người đại diện cho con người mà nó hoạt động theo cách thực sự tự chủ” (Candrian và Scherer, 2022: 1). “Ngày càng có nhiều AI có tính tự chủ cao. Nghĩa là, chúng được thiết kế và được trao quyền để đưa ra quyết định của riêng mình dựa trên hệ dữ liệu mà chúng thu nhận, phân tích và diễn giải” (ServiceNow, 2020). Theo Shrestha và các cộng sự (2022: 2): “Gần đây, nghiên cứu về hiệu suất của trí tuệ nhân tạo (AI) đã chỉ ra sự tự thúc đẩy trong quá trình phát triển của nó, đặc biệt là do sự phát triển nhanh chóng của khả năng tự chủ suy xét. Thay vì chỉ đơn giản hỗ trợ con người, các hệ thống AI này có khả năng tự động thực hiện nhiều nhiệm vụ khác nhau. Con người ngày càng phải phụ thuộc vào các thuật toán AI khi đưa ra các quyết định quan trọng”.

Việc AI can thiệp vào quá trình ra quyết định đang làm giảm trách nhiệm của con người. Một khi con người quá phụ thuộc vào AI và mất tự tin trong việc đánh giá và tự suy xét, gây tình trạng thụ động, giảm khả năng đánh giá và tính công bằng trong quyết định. Mặc dù “sử dụng trí tuệ nhân tạo trong việc ra quyết định cuối cùng, con người vẫn phải chịu trách nhiệm về kết quả của quyết định. Tuy nhiên, có nhiều bằng chứng từ các nghiên cứu gần đây cho thấy việc dựa vào AI

trong quá trình ra quyết định có thể tạo ra một loạt các thành kiến và thách thức nghiêm trọng, tiềm ẩn rủi ro liên quan đến tính công bằng, trách nhiệm cá nhân và tính minh bạch trong quá trình giải trình lựa chọn quyết định” (Shrestha và các cộng sự, 2019: 8).

Hơn nữa, can thiệp của AI trong quyết định của con người có thể tạo ra các rào cản xã hội. AI, khi sử dụng dữ liệu lịch sử, có thể gây ra thiên lệch và phân biệt dựa trên giới tính, chủng tộc, hoặc các thuộc tính khác, đe dọa tính công bằng và đa dạng trong xã hội, gây ra phân cách và leo thang các vấn đề xã hội hiện có. Elliott và Soifer (2022: 8) khẳng định: “Các nghiên cứu đã chỉ ra rằng các thuật toán máy tự học có thể thu thập và tái tạo những định kiến ngấm của con người liên quan đến chủng tộc và giới tính từ dữ liệu văn bản trực tuyến mà họ sử dụng, để rút ra những kết luận quan trọng và đưa ra quyết định cuối cùng”.

Như vậy, mặc dù AI có tiềm năng lớn trong việc tối ưu hóa quá trình ra quyết định của con người, nhưng tham chiếu đến quan điểm về tự do của Jean-Paul Sartre, nó cũng khiến cho nguy cơ đánh mất khả năng tự quyết định của con người và làm giảm tính công bằng trở nên rõ ràng. Điều này đối lập với nguyên tắc cơ bản của chủ nghĩa hiện sinh, vốn tôn trọng quyền tự do và ý chí của con người.

3.2. Tác động của AI đối với việc bảo toàn nhân vị⁽⁹⁾

Theo quan điểm của chủ nghĩa hiện sinh, sự bảo toàn nhân vị là một khía cạnh quan trọng trong định hình danh tính và sự tồn tại của con người. Mỗi cá nhân trong xã hội đều là một chủ thể độc lập có quyền và trách nhiệm, cần có ý thức và tự do. Tuy nhiên, trong một thế giới đang phát triển với nhiều tiến bộ của công nghệ và AI, sự bảo toàn nhân vị có thể đối mặt với nhiều thách thức. Sự can thiệp của công nghệ, đặc biệt là việc thu thập thông tin cá nhân đã đe dọa trực tiếp đến quyền riêng tư và sự độc lập của từng cá thể. Có thể khẳng định, “... trong số những mối quan tâm nổi bật nhất của mọi người về công nghệ AI là những vấn đề liên quan đến quyền riêng tư” (Elliott và Soifer, 2022: 1).

Quyền riêng tư đòi hỏi bảo vệ thông tin cá nhân và ngăn truy cập trái phép. Đây là một quyền cơ bản, đảm bảo quyền kiểm soát dữ liệu cá nhân và cách chúng được sử dụng. Với sự phát triển của AI, quyền riêng tư trở nên ngày càng quan trọng hơn, khi sự thu thập và phân tích dữ liệu cá nhân gia tăng không ngừng. Mặc dù AI có thể mang lại lợi ích trong việc tối ưu hóa quyết định của con người, nhưng việc thu thập và chia sẻ lượng lớn dữ liệu hàng ngày cũng tiềm ẩn nguy cơ làm tiết lộ thông tin nhạy cảm mà cá nhân không muốn tiết lộ.

Quyền riêng tư liên quan đến hình ảnh cá nhân mà mỗi người xây dựng. Mọi người quan tâm đến cách người khác đánh giá họ, như Jean-Paul Sartre đã đề cập đến trong quan niệm “Cái

nhìn”⁽¹⁰⁾. Sartre (2003: 246) nói: “Tôi tồn tại như cách người khác nhìn nhận tôi”. Tức là, con người có quyền lựa chọn hình ảnh họ muốn người khác thấy, được xây dựng thông qua quyền riêng tư và quyền thông tin cá nhân. Những quyền này cho phép họ kiểm soát thông tin mà họ muốn chia sẻ hoặc ẩn giấu để tạo dựng hình ảnh cá nhân theo ý muốn. Từ quan điểm này, con người có quyền tự chọn hình ảnh họ muốn hiển thị trước mắt người khác thông qua quyền riêng tư và quyền thông tin cá nhân. Những quyền này cho phép họ lựa chọn thông tin mà họ muốn chia sẻ và thông tin mà họ muốn bảo mật để xây dựng một hình ảnh cá nhân theo ý muốn.

Tuy nhiên, AI và công nghệ giám sát hàng loạt (Mass Surveillance)⁽¹¹⁾ đang tạo ra mối đe dọa nghiêm trọng đối với quyền riêng tư cá nhân. Theo Macnish (2020: 10) “Phương pháp giám sát này cho phép tự động thu thập và xử lý dữ liệu mà không cần xem xét trách nhiệm và quyền sở hữu dữ liệu”. Quá trình này không phân biệt giữa các đối tượng, mục tiêu và áp dụng cho hầu hết mọi người sử dụng các thiết bị kết nối mạng và các ứng dụng thông minh. Nó diễn ra một cách bí mật và hoàn toàn tự động. Nghĩa là nhiều người không biết, không hiểu rõ về lượng và phạm vi dữ liệu được thu thập, “... quá trình thu thập dữ liệu này rất lớn và toàn diện hơn nhiều so với việc giám sát phi kỹ thuật số của cá nhân trước đây” (Elliott và Soifer, 2022: 4).

Ngoài ra, nhân vị của con người còn thể hiện ở khía cạnh tự do hình thành và phát triển ý kiến cá nhân. Những gì con người phát biểu, phát ngôn thể hiện cá tính, nhân vị của mỗi con người. Ngày nay, với sự phát triển của AI, giao tiếp chủ yếu diễn ra trực tuyến, các phương tiện truyền thông truyền thống đang được thay thế bởi các mạng xã hội, các trang tổng hợp tin tức và công cụ tìm kiếm, những thứ thống trị cách thức cung cấp thông tin cho cá nhân. Xã hội phụ thuộc rất nhiều vào các nền tảng trực tuyến để thực hiện quyền tự do ngôn luận và truy cập tin tức. Tìm kiếm trên internet là một trong những hình thức cá nhân hóa nội dung phổ biến nhất được thực hiện bởi AI (Kaye, 2018: 11). Vì lý do này, các công cụ tìm kiếm AI đóng vai trò rất quan trọng đối với các cá nhân trong việc tìm kiếm, nhận hoặc truyền đạt thông tin. AI ngày càng hướng đến sự cá nhân hóa, càng chi tiết hóa trải nghiệm trực tuyến cá nhân của mỗi người dùng. Điều này có nghĩa là nó cố gắng hiển thị thông tin và nội dung dựa trên sở thích và lịch sử trực tuyến của mỗi người. Lợi ích là con người thuận tiện trong việc truy cập nội dung thông tin cần thiết và quan tâm. Tuy nhiên, sự cá nhân hóa dựa trên AI có thể làm suy giảm khả năng tiếp cận các quan điểm khác nhau và giới hạn lựa chọn của cá nhân trong việc tìm kiếm nội dung. Điều này ảnh hưởng việc hình thành và phát triển ý kiến của cá nhân.

3.3. Đạo đức, trách nhiệm và sự tha hóa của con người trong kỷ nguyên trí tuệ nhân tạo

Sự tiến bộ của AI mang lại nhiều tiềm năng và lợi ích cho xã hội, nhưng đồng thời cũng đặt ra những thách thức đạo đức mới mà con người phải đối mặt. K. Schwab (2016: 94) đã nhấn mạnh rằng: “Những tiến bộ công nghệ đang đẩy chúng ta đến những biên giới mới về luân lý. Liệu chúng ta chỉ nên sử dụng những tiến bộ đáng kinh ngạc của sinh học để chữa bệnh hoặc làm lành vết thương, hay nên dùng chúng để biến bản thân thành những con người hoàn hảo hơn? Nếu chấp nhận phương án sau, chúng ta có nguy cơ biến việc làm cha/mẹ thành một phần của xã hội tiêu dùng, nơi trẻ em trở thành hàng hóa, những vật thể được đặt hàng theo nguyện vọng của chúng ta? Chúng ta phải đối mặt với những câu hỏi phức tạp và ranh giới luân lý tương tự khi nói về AI”.

Trong thời đại mà robot ngày càng tiến bộ, yêu cầu về khía cạnh đạo đức trở nên ngày càng quan trọng. Con người đòi hỏi robot phải thể hiện trí tuệ và khả năng cảm xúc tương tự như con người, mặc dù việc xây dựng một khung chuẩn đạo đức cho robot vẫn còn đang trong quá trình định hình. Điều cần phải nhấn mạnh là sự quan trọng của việc phân biệt rõ sự khác biệt giữa robot và con người. Abney (2012: 46) đã thể hiện quan điểm rằng: “Hệ thống suy nghĩ liên quan đến khả năng của con người xây dựng các tương lai khả thi khác nhau

thông qua quá trình tư duy, và sau đó lựa chọn hành động dựa trên tư duy mà con người muốn biến thành trải nghiệm thực tế.” Hệ thống suy nghĩ này đại diện cho quá trình ra quyết định và chứa trong nó sự hiểu biết, thông tin và cảm xúc cá nhân, trong khi robot chỉ có thể đưa ra quyết định dựa trên thông tin được cung cấp hoặc thu thập từ con người. Abney (2012: 37) cũng nêu rõ rằng “Khía cạnh đạo đức liên quan đến bản sắc con người, không chỉ đơn thuần là việc đánh giá đúng hoặc sai về từng hành vi”. Đạo đức, được chia thành nhiều hình thức khác nhau từ quy tắc đạo đức, xác định nhân cách và cuộc sống của con người. Đạo đức, đức hạnh đặt câu hỏi về ai là chủ thể của cuộc sống, có trách nhiệm với hành động và sự lựa chọn của mình. Chỉ con người mới có thể trả lời câu hỏi đó, tức là robot không thể tự suy nghĩ. Dù cho robot có bộ não thông minh hơn con người trong một số lĩnh vực, nhưng chúng không có hệ thống suy nghĩ gắn liền với yếu tố đạo đức và cảm xúc giống con người.

Sartre cho rằng, đạo đức không thể được tách rời khỏi quyền tự do và trách nhiệm cá nhân mà trong đó, đạo đức tồn tại xuyên suốt quá trình con người tự xác định và tự chịu trách nhiệm về hành động của mình. Ngay cả khi con người đối xử với con người như nô lệ, thì những ông chủ cũng yêu cầu “sự chủ động từ người nô lệ trong giới hạn của nhiệm vụ là yêu cầu bắt buộc; nghĩa là họ phải tự

khám phá và đảm nhận một số trách nhiệm” (Sartre, 1992b: 331). Ông cũng khẳng định rằng: “Con người không được đánh giá là vĩ đại hay khốn khổ dựa trên ý kiến của người khác. Tuy nhiên, trong ngữ cảnh xã hội và theo các tiêu chuẩn cụ thể, sự vĩ đại hay khốn khổ của một người không thể tránh khỏi sự ảnh hưởng của một hoàn cảnh khốn khổ hoặc may mắn. Bởi vì con người có quan điểm, giới hạn và ngẫu nhiên, anh ta tạo nên một thế giới của riêng mình, tức là anh ta đồng thời chịu trách nhiệm với bản thân và với vũ trụ” (Sartre, 1992b: 493).

Theo Sartre, khi con người không hiểu để chịu trách nhiệm cho sự tự do của mình thì tha hóa xảy ra⁽¹²⁾. Sartre cho rằng con người thường phải đối mặt với sự tha hóa trong cuộc sống hàng ngày, đặc biệt trong môi trường xã hội và công việc. Trong bối cảnh AI ngày càng phát triển, nó sẽ vượt ra khỏi tầm kiểm soát của con người, trong khi khung đạo đức của AI còn chưa rõ ràng đối với con người thì tất yếu dẫn đến tình trạng tha hóa. Sartre (1992b: 294) chỉ rõ: “Một khía cạnh của sự tha hóa là sự không hiểu biết. Đó là một cách mà con người tương tác với nhau. Nó có thể được hiểu là một sự từ chối, trong nghĩa rằng bị người khác đánh giá là không biết điều gì đó ảnh hưởng đến sự tự do của tôi”.

Vì vậy, trong thời đại của AI, việc duy trì tính cá nhân và tự quyết định trở nên cực kỳ quan trọng. Con người cần phải đánh thức ý thức của họ, tự

quyết định và hành động theo giá trị và ý nghĩa cá nhân của họ để vượt qua sự tha hóa và đảm bảo rằng AI không làm mất đi sự tự quyết định và đạo đức của con người.

4. HƯỚNG TỚI KHUNG TRIẾT LÝ ĐẠO ĐỨC VÀ TÔN TRỌNG TỰ DO CÁ NHÂN TRONG KỶ NGUYÊN TRÍ TUỆ NHÂN TẠO

Đối diện với những tác động sâu rộng về các quyền cơ bản của con người như quyền riêng tư, an ninh và cơ hội bình đẳng, cũng như những nguyên tắc đạo đức như sự công bằng và tôn trọng, việc đặt ra vấn đề đạo đức liên quan đến AI trở nên quan trọng. Những thách thức đạo đức này yêu cầu một cái nhìn toàn diện, từ cấp độ cá nhân, xã hội đến toàn cầu. Nhìn từ góc độ triết học, việc đặt ra trách nhiệm đạo đức của AI sẽ phải giải quyết một loạt các vấn đề mâu thuẫn về thế giới quan. Bởi chỉ có con người mới là chủ thể của mọi hoạt động thực tiễn, máy móc là công cụ lao động của con người do đó không thể bắt buộc chúng phải có đạo đức và trách nhiệm. J. Storrs Hall (2011: 29) đã khẳng định: “Đạo đức nằm trên vai con người, và nếu máy móc thay đổi sự thuận tiện trong việc thực hiện mọi thứ, chúng không thay đổi trách nhiệm về việc thực hiện chúng. Luôn luôn chỉ con người mới là ‘những chủ thể đạo đức’”. Tương tự, con người chủ yếu là đối tượng của trách nhiệm. Đang có một cuộc tranh luận ngày càng gay gắt về trách nhiệm của chúng ta đối với các sinh vật sống khác hoặc các loài trong

số chúng... Tuy nhiên, chúng ta chưa bao giờ tự coi mình có nghĩa vụ ‘đạo đức’ đối với máy móc của chúng ta hoặc của chúng đối với chúng ta”.

Thật vậy, con người chịu trách nhiệm thiết kế, phát triển và triển khai công nghệ. Vì vậy, người thiết kế, nhà sản xuất hoặc người sử dụng là người phải chịu trách nhiệm cho công nghệ và những hành động cuối cùng được thực hiện (hoặc không thực hiện) với nó. Ngoài ra, những quyền đặc quyền, quyền lợi, quyền lực và những yêu cầu liên quan đến việc sử dụng hoặc áp dụng công nghệ cần được tôn trọng.

Theo Martin Heidegger (1977: 4-5), “khi chúng ta đặt câu hỏi về công nghệ, có hai câu trả lời chính. Một là, công nghệ là một phương tiện để đạt đến một mục tiêu. Hai là, công nghệ là một hoạt động của con người. Hai định nghĩa này liên quan chặt chẽ với nhau, vì đặt mục tiêu và sử dụng phương tiện để đạt được mục tiêu là một hoạt động của con người”. Heidegger gọi đặc điểm này là “định nghĩa công cụ” và cho rằng nó là cách hiểu đúng về bất kỳ sáng tạo công nghệ nào.

Do đó, lý thuyết công cụ “đưa ra quan điểm được chấp nhận rộng rãi nhất về công nghệ. Nó dựa trên ý tưởng thông thường rằng công nghệ là “công cụ” sẵn sàng phục vụ mục đích của người dùng”. Và vì một công cụ hoặc dụng cụ “được coi là “trung lập”, không có nội dung có giá trị riêng của nó” (Feenberg, 1991: 5), một tác phẩm

công nghệ được đánh giá không phải ở bản thân nó mà dựa trên cơ sở công dụng của nó được quyết định bởi con người thiết kế nó hoặc người dùng. Do đó, về vấn đề này, rào cản để xây dựng khung đạo đức đối với một công cụ máy móc, như robot hoặc AI là rất lớn. Nó đòi hỏi xem xét trách nhiệm ở từng mức độ tương tác giữa AI và đạo đức. Ngày 25/11/2023, Hoa Kỳ, Vương quốc Anh và 16 quốc gia khác đã cùng nhau công bố một thỏa thuận quốc tế toàn diện nhằm đảm bảo an toàn cho các hệ thống AI ngay từ giai đoạn thiết kế. Đây là một bước quan trọng đánh dấu sự hợp tác để tạo ra một môi trường an toàn và có trách nhiệm hơn cho sự phát triển của công nghệ AI trên phạm vi toàn cầu. Thảo luận thống nhất quan điểm về lợi ích cũng như nguy cơ của AI, đồng thời đưa ra các hướng dẫn để hướng tới một kỷ nguyên AI an toàn: Hệ thống AI có tiềm năng mang lại nhiều lợi ích cho xã hội. Tuy nhiên, để các cơ hội của AI được hiện thực hóa đầy đủ, nó phải được phát triển, triển khai và vận hành một cách an toàn và có trách nhiệm. Các hệ thống AI phải đối mặt với các lỗ hổng bảo mật mới cần được xem xét cùng với các mối đe dọa an ninh mạng tiêu chuẩn. Khi tốc độ phát triển cao – như trường hợp của AI – bảo mật thường có thể được xem xét thứ yếu. Bảo mật phải là yêu cầu cốt lõi, không chỉ trong giai đoạn phát triển mà trong suốt vòng đời của hệ thống. Vì lý do này, các hướng dẫn được chia thành bốn lĩnh vực chính trong vòng đời phát triển hệ thống AI:

thiết kế an toàn, phát triển an toàn, triển khai an toàn cũng như vận hành và bảo trì an toàn (NCSC, CISA, 2023: 5).

Tuy nhiên, dưới góc độ triết học, quan niệm tự do của Sartre không chỉ đơn thuần là khả năng lựa chọn, mà còn là việc chấp nhận trách nhiệm và hậu quả của những quyết định mình đưa ra. Khi áp dụng quan điểm này vào mối quan hệ giữa người và máy, chúng ta cần tạo ra những hệ thống AI không chỉ dựa trên logic và dữ liệu mà còn phải có khả năng đối mặt với những giới hạn và sự mâu thuẫn của thế giới con người.

Trong bối cảnh xã hội hiện đại, nơi mà sự giao thoa giữa con người và máy móc đã tạo ra một cuộc cách mạng về cách chúng ta sống, làm việc và giao tiếp, thì vấn đề đạo đức và tôn trọng tự do cá nhân trở nên càng phức tạp hơn bao giờ hết. Tư duy của Sartre nhấn mạnh rằng mọi hành động đều mang trong mình sự trách nhiệm và sự lựa chọn của bản thân. Vì vậy, khi chúng ta đối diện với việc phát triển và sử dụng AI, việc đảm bảo rằng những quyết định và hành động của máy móc vẫn đảm bảo sự tôn trọng đối với quyền tự do và cá nhân của con người trở nên quan trọng. Điều này đòi hỏi chúng ta phải cân nhắc sự tương tác phức tạp giữa con người và AI. Tôn trọng tự do cá nhân không chỉ đề cập đến việc bảo vệ quyền lựa chọn và quyết định của con người, mà còn bao gồm việc đảm bảo rằng AI không trở thành một yếu tố áp đặt hoặc làm mất đi tính cá nhân và

quyền tự chủ. Khi xây dựng các hệ thống AI, việc xem xét và ứng phó với tác động xã hội, đạo đức và tâm lý là không thể thiếu. Hệ thống AI nên phải có khả năng “đối thoại” với người dùng và xem xét những hậu quả của hành động mình để đảm bảo tính đạo đức và tôn trọng cho con người. Sartre đã giải thích rằng tự do không chỉ là khả năng lựa chọn, mà còn là việc tạo ra ý nghĩa và giá trị. Trong việc phát triển và sử dụng AI, chúng ta cần đảm bảo rằng các máy móc thông minh không chỉ hoạt động dựa trên các thuật toán và dữ liệu, mà còn phải có khả năng “tạo nên” ý nghĩa thông qua tương tác và hoạt động độc lập.

Tóm lại, tư tưởng tự do của Sartre mang lại một góc nhìn quan trọng về triết học đạo đức và sự tôn trọng đối với tự do cá nhân trong kỷ nguyên đang thay đổi do sự tương tác phức tạp giữa con người và máy móc. Việc áp dụng tư tưởng này không chỉ cung cấp một khung triết lý để đối mặt với thách thức của AI, mà còn đề cao sự tôn trọng đạo đức và sự tự chủ trong việc phát triển công nghệ và xã hội.

5. KẾT LUẬN

Trong bối cảnh nhiều thách thức và không ít cơ hội của kỷ nguyên xã hội giao thoa giữa người và máy, triết lý

hiện sinh của Jean-Paul Sartre đã cung cấp cho chúng ta những cảnh báo về mối liên hệ phức tạp giữa tự do cá nhân và AI.

Sự tương tác ngày càng gia tăng giữa người và máy móc đã đặt ra những vấn đề về quyền riêng tư và độc lập. Sự tương đồng và mâu thuẫn giữa triết lý hiện sinh của Sartre và thế giới hiện đại đặt ra nhiều vấn đề cấp thiết về tích hợp giữa cá nhân và cộng đồng trong việc sử dụng và phát triển công nghệ. Mặc dù AI mang lại tiện ích và hiệu suất vượt trội, nhưng nó cũng đặt ra những vấn đề nan giải về sự riêng tư và quyền tự do của mỗi người. Sự can thiệp của AI vào quyết định và hành vi con người có thể ảnh hưởng đến tính cá nhân và tạo nên một cách nhìn mới về quyền riêng tư.

Ngoài ra, sự phụ thuộc ngày càng lớn của con người vào AI cũng đẩy chúng ta đối mặt với những biến đổi đạo đức xã hội. Sự tương tác liên tục này có thể tạo nên sự chuyển đổi về trách nhiệm và hậu quả đạo đức của hành động con người. Khía cạnh này chứng tỏ rằng triết lý của Sartre không chỉ là một lý thuyết cổ điển mà còn có giá trị thực tiễn trong bối cảnh hiện đại, giúp chúng ta hiểu rõ hơn về cách thức con người và AI tương tác và tác động lẫn nhau. □

CHÚ THÍCH

⁽¹⁾ Rule-based system là một hệ dựa trên quy tắc được sử dụng để lưu trữ và vận dụng kiến thức nhằm giải thích thông tin một cách hữu ích.

⁽²⁾ Machine learning là một nhánh của trí tuệ nhân tạo (AI) và khoa học máy tính, đi sâu vào sử dụng dữ liệu và thuật toán để bắt chước và tinh chỉnh việc thực hiện các hành động của con người. Machine learning còn là một thành phần quan trọng của lĩnh vực khoa học dữ liệu.

- (3) Neural Network là chuỗi thuật toán nhằm tìm kiếm quan hệ trong tập hợp dữ liệu hệ thống dựa theo cách thức hoạt động não bộ con người.
- (4) Jean-Paul Sartre, nhà triết học hiện sinh người Pháp với quan điểm về tự do của con người. Tư tưởng tự do của Sartre tập trung vào ý niệm rằng tự do không phải là một trạng thái tồn tại sẵn trong con người, mà là một trạng thái mà mỗi cá nhân phải tự tạo dựng và xác định. Ông cho rằng chúng ta không định trước sẵn bản chất hay ý nghĩa của cuộc sống, mà tự quyết định thông qua các lựa chọn, các hành động của mình và chịu trách nhiệm trước những lựa chọn đó. Tư tưởng tự do của Jean-Paul Sartre không chỉ là một phần quan trọng trong lĩnh vực triết học mà còn có tầm ảnh hưởng lớn đến cách chúng ta đối diện với cuộc sống và xã hội hiện đại.
- (5) Mạng neural nhân tạo hay thường gọi ngắn gọn là mạng neural (tiếng Anh là artificial neural network - ANN hay neural network) là một mô hình toán học hay mô hình tính toán được xây dựng dựa trên các mạng neural sinh học.
- (6) Fuzzy Logic là một phương pháp lập luận tương tự khả năng lý luận của con người. Hệ thống này có thể ra quyết định tương tự như con người, dựa vào tất cả các khả năng trung gian giữa các giá trị kỹ thuật số “có” và “không”.
- (7) Intelligent Agents là một hệ thống phần mềm máy tính có khả năng hoạt động độc lập nhằm đạt được các mục tiêu nhất định và phản hồi những đối tượng đang diễn ra xung quanh.
- (8) Case-Based Reasoning là phương pháp tìm lời giải dựa trên việc sử dụng lại các ca suy luận trong quá khứ.
- (9) Nhân vị (personal) là khái niệm then chốt của triết học hiện sinh, dùng để chỉ bộ mặt đặc hữu của cá nhân. Trong triết học hiện sinh, nhân vị đề cập đến khả năng tự quyết định của con người, quyền tự do trong việc lựa chọn và hình thành cuộc sống của mình dưới sự tác động của ý thức và ý muốn cá nhân. Khái niệm này nhấn mạnh sự độc đáo và quan trọng của mỗi cá nhân trong xã hội và cuộc sống cá nhân, đồng thời đề cao quyền tự do và tôn trọng quyền tự quyết định của mỗi người.
- (10) Một trong những khía cạnh về quan niệm cái nhìn của Sartre là “mắt của người khác” (the gaze of the other). Ông cho rằng khi chúng ta bị người khác nhìn thấy, chúng ta trải qua một sự thay đổi tâm trạng và nhận thức về bản thân. Người khác có thể đánh giá, phê phán và tạo ra một hình ảnh về chúng ta, và điều này ảnh hưởng đến cách chúng ta nhìn nhận bản thân mình. Sự hiện diện của người khác qua cái nhìn có thể làm cho chúng ta cảm thấy bị kiểm soát, mất đi tính tự do và tự chủ.
- (11) Mass surveillance sử dụng các hệ thống hoặc công nghệ thu thập, phân tích và/hoặc tạo dữ liệu về số lượng người lớn hoặc không xác định thay vì giám sát hạn chế đối với những cá nhân cụ thể. Dưới các hình thức giám sát hàng loạt hiện có, các chính phủ có thể nắm bắt hầu như tất cả các khía cạnh trong cuộc sống của công dân.
- (12) Quan niệm “Tha hóa” (alienation) trong triết học hiện sinh của Jean-Paul Sartre thường được hiểu là mất đi tính cá nhân và tự do của con người do áp lực xã hội và tình huống cụ thể. Tha hóa có thể diễn ra khi con người mất đi sự tự quyết định và kiểm soát về cuộc sống của họ và trở thành đối tượng của các lực lượng bên ngoài.

TÀI LIỆU TRÍCH DẪN

1. A Conversation with Ginni Rometty, Chairman, President and Chief Executive Officer of IBM, on Business Transformation in the Era of Cognitive Computing. <https://www.youtube.com/watch?v=mNEQsl1-iZs>. Accessed 16 Oct 2023.
2. Abney, K., Bekey, G.A., & Lin, P. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*. London: The MIT Press.
3. Bader, J., Edwards, J., Harris-Jones, C., & Hannaford, D. 1988. "Practical Engineering of Knowledge-Based Systems". *Information and Software Technology*, 30(5), pp. 266-277. doi:10.1016/0950-5849(88)90019-5.
4. Boden, M.A. 2016. *AI: Its Nature and Future*. Oxford University Press.
5. Candrian, C., & Scherer, A. 2022. "Rise of the Machines: Delegating Decisions to Autonomous AI". *Computers in Human Behavior*, 134, 107308 1-16. doi: 10.1016/j.chb.2022.107308.
6. Cellan-Jones, R. 2014. "Stephen Hawking Warns Artificial Intelligence Could End Mankind". *BBC News*. <https://www.bbc.com/news/technology-30290540>. Accessed 16 Oct 2023.
7. Copeland, B.J. 2023. "Artificial Intelligence. Encyclopedia Britannica". <https://www.britannica.com/technology/artificial-intelligence>. Accessed 15 Oct 2023.
8. Duan, Y., Edwards, J.S., & Dwivedi, Y.K. 2019. "Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges and Research Agenda". *International Journal of Information Management*, 48, pp. 63-71. doi: 10.1016/j.ijinfomgt.2019.01.
9. Elliott, D., & Soifer, E. 2022. "AI Technologies, Privacy, and Security". *Frontiers in Artificial Intelligence*, 1(5), pp. 1-8. doi: <https://doi.org/10.3389/frai.2022.826737>.
10. European Commission's High-Level Expert Group on Artificial Intelligence [AI HLEG] 2018. *A Definition of AI: Main Capabilities and Scientific Disciplines*. Brussels: European Union.
11. Feenberg, A. 1991. *Critical Theory of Technology*. New York: Oxford University Press.
12. Hall, J.S. 2011. "Ethics for Machines". In M. Anderson & S.L. Anderson (Eds.). *Machine Ethics*. USA: Cambridge University Press.
13. Heidegger, M. 1977. *The Question Concerning Technology and Other Essays*. USA: Garland.
14. Johnson, D.G. 2006. "Computer systems: Moral Entities but not Moral Agents". *Ethics and Information Technology*, 8(4), pp. 195-204. doi:10.1007/s10676-006-9111-5
15. Kaye, D. 2018. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/73/348)*. New York: United Nations General Assembly.
16. Macnish, K. 2020. Mass Surveillance: A Private Affair? *Moral Philosophy and Politics*, 7(1), pp. 9-27. doi:10.1515/mopp-2019-0025.
17. McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 27, pp. 12-14.

18. McCorduck, P., & Cfe, C. 2004. *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. CRC Press.
19. NCSC, CISA. 2023. *Guidelines for Secure AI System Development*.
20. Nilsson, N.J. 1992. *Principles of Artificial Intelligence*. Palo Alto, CA: Tioga Publishing Company.
21. Nowak, A., Lukowicz, P., & Horodecki, P. 2018. "Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance? or the Biggest Threat" [Opinion]. *IEEE Technology and Society Magazine*, 37(4), pp. 26-34. doi:10.1109
22. OECD. 2019. Artificial Intelligence in Society. OECD EBooks. <https://doi.org/10.1787/eedfee77-en>
23. Phillips-Wren, G. 2012. Ai Tools in Decision Making Support Systems: a Review. *International Journal on Artificial Intelligence Tools*, 21(02), 124000. 1-13. doi:10.1142/s0218213012400052.
24. Phillips-Wren, G., & Jain, L. 2006. "Artificial Intelligence for Decision Making". In *Knowledge-Based Intelligent Information and Engineering Systems: 10th International Conference, KES 2006, Bournemouth, UK, October 9-11, 2006. Proceedings, Part II 10*, pp. 531-536.
25. Rawlinson, K. 2015. *Microsoft's Bill Gates Insists AI is a Threat*. BBC. <http://www.bbc.com/news/31047780>. Accessed 16 Oct 2023.
26. Sadiku, M. 1989. "Artificial intelligence". *IEEE Potentials*, pp. 35-39.
27. Sartre. J.P. 1992a. *Being and Nothingness* (Translated by E.B. Hazel), New York: University of Colorado.
28. Sartre. J.P. 1992b. *Notebooks for an Ethics*. USA: Chicago Press.
29. Sartre. J.P. 2003. *Being and Nothingness: An Essay on Phenomenological Ontology*. London: Routledge.
30. Sartre. J.P. 2015. *L'existentialisme est un humanisme*. France: Gallimard.
31. Schalkoff, R. 1990. *Artificial intelligence: An Engineering Approach*. New York: McGraw-Hill.
32. Schwab. K. 2016. *The Fourth Industrial Revolution*, Switzerland: World Economic Forum.
33. ServiceNow. 2020. Autonomous AI: How Powerful New Networks Will Drive More Businesses. <https://www.forbes.com/sites/servicenow/2020/10/24/autonomous-ai-how-powerful-new-networks-will-drive-more-businesses/?sh=680d90ea50ae>. Accessed 18 Oct 2023.
34. Shrestha, Y.R., Ben-Menahem, S.M., & von Krogh, G. 2019. "Organizational Decision-Making Structures in the Age of Artificial Intelligence". *California Management Review*, 000812561986225, 1-18. doi:10.1177/0008125619862257.