TÍNH TỰ TRỊ CỦA TRÍ TUỆ NHÂN TẠO – MỘT GÓC NHÌN TỪ TRƯỜNG PHÁI ĐẠO ĐỨC HỌC CỦA KANT

Phạm Minh Đức[(*)]

[(*)]Viện Khoa học và Giáo dục Khai phóng, Đại học VinUni;

Email: duc.pm@vinuni.edu.vn

Nguyễn Anh Tuấn[(**)]

[(**)]Phó giáo sư, tiến sĩ, Khoa Triết học, trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia Hà Nội.

Email: nguyenanhtuantr@gmail.com

Tóm tắt: Bài viết tìm hiểu và luận giải về tính tự trị của trí tuệ nhân tạo (AI - *Artificial Intelligence*) qua góc nhìn đạo đức học của Immanuel Kant. Trên cơ sở trình bày khái niệm và quá trình phát triển của AI, phân tích cụ thể về tính tự trị của AI và mối liên hệ giữa tính tự trị với ý chí và mệnh lệnh tuyệt đối theo đạo đức học của Kant, bài viết nghiên cứu mối liên hệ giữa tính tự trị và mục đích tự thân, đồng thời chỉ ra những hạn chế và thiếu sót trong quan niệm về việc AI có thể sở hữu tính tự trị một cách đầy đủ; cuối cùng, bài viết trình bày và phân tích về tự trị của AI trong mối liên hệ với mục đích tự thân của con người, cũng như tính khả thi của việc áp dụng các mệnh lệnh tuyệt đối cho AI.

Từ khóa: tính tự trị, trí tuệ nhân tạo (AI), đạo đức học Kant, ý chí, mệnh lệnh tuyệt đối, mục đích tự thân.

Ngày nhận bài: 07/07/2024; ngày phản biện: 08/07/2024; ngày sửa chữa: 01/08/2024; ngày duyệt đăng: 15/08/2024.

1. Mở đầu

Cuộc cách mạng công nghiệp lần thứ tư đã tác động mạnh mẽ đến xã hội loài người. Điểm đặc trưng nhất của cuộc cách mạng này là sự tự động hóa đã vượt ra ngoài khuôn khổ của các ngành công nghiệp, thâm nhập sâu vào các quá trình xã hội khác nhau. Chúng

ta đang chứng kiện một xã hội tự động hóa ngày càng rộng rãi. Với những tiến bộ trong công nghệ như trí tuệ nhân tạo (Artificial Intelligence – AI), dữ liệu lớn, internet vạn vật, robot, v.v., những công việc thường ngày, những hoạt động kinh tế, xã hội và thậm chí là văn hóa đã và đang được thực hiện bởi máy móc (Peña-Cabrera và cộng sự 2019). Một trong những ví dụ minh họa tiêu biểu là việc, AI đang thực hiện một số chức năng tư duy và tính toán của con người. Những tác vụ như viết, sáng tạo nội dung, sáng tạo tri thức, truyền đạt tri thức, v.v. trước kia đã từng thuộc thẩm quyền riêng của con người, thì nay đã được AI thực hiện. Chẳng hạn, gần đây AI đã giúp nữ nhà văn người Nhật, Rie Kudan, đoạt giải cao văn học. Nữ văn sĩ cho biết, cô đã trích nguyên văn một số câu của AI vào trong tác phẩm của cô (Phương Linh 2024). Dù tác giả chỉ dùng 5% nhưng có thể thấy AI đã phần nào có thể thực hiện một số thao tác tư duy, sáng tạo nội dung.

Tình trạng nói trên đã làm nảy sinh vấn đề về tính tự trị của AI. Theo nghĩa chung nhất, có thể định nghĩa tính tự trị của AI là khả năng nó vận hành theo các quy tắc do nó tự định ra. Điều này hàm ý việc AI có khả năng tự ra quyết định bằng cách "tuân theo thuật toán để phản hồi với những dữ liệu đầu vào từ môi trường một cách độc lập với những dữ liệu đầu vào theo thời gian thực của con người" (Etzioni và Etzioni 2016: 149). Theo đó, AI không chỉ tuân theo những thuật toán được con người cài đặt sẵn, mà còn có thể tự cập nhật, phát triển chính thuật toán ấy để sao cho có những phản hồi phù hợp với những tác động từ hoàn cảnh, môi trường xung quanh. Cho nên, khi tồn tại trong xã hội loài người, tính tự trị của AI đã làm nảy sinh những vấn đề đạo đức liên quan tới khả năng ra quyết định của AI. Trong số đó, hai vấn đề nổi bật nhất là: liệu AI có phải chịu trách nhiệm cho những quyết định của mình không, và liệu con người có nên coi AI như những chủ thể đạo đức và qua đó cài đặt dữ liệu về đạo đức cho các chương trình, thuật toán của AI hay không?

Trong bối cảnh như vậy thì đạo đức học Kant có thể cung cấp những giải đáp, trả lời thích đáng. Bởi vì đạo đức học này cho rằng, con người là những thực thể có lý trí, cư xử với nhau theo những nguyên tắc xác định dưới sự kiểm soát của ý chí. Góc nhìn này giúp ta có được một mối liên hệ với thuật toán và các quy tắc của AI. Hơn thế,

đạo đức học Kant xem xét bản thân nguyên tắc đạo đức là những quy luật phổ quát riêng có của xã hội, có cùng cấp độ tác động y như các quy luật tự nhiên. Tuy có một số vấn đề với góc nhìn này, nhưng nhìn chung nó cũng giúp chúng ta tạo ra một bộ khung, vạch ra những nền tảng đạo đức căn bản có thể có cũng như những định hướng để xem xét sự phát triển của AI cùng với tính tự trị của nó trong xã hội loài người. Xa hơn thế, với nền tảng đạo đức học của Kant, ta có thể hướng tới một AI vì nhân loại.

2. Khái niệm và sự phát triển của trí tuệ nhân tạo (AI)

Nhiều tài liệu chuyên ngành đều định nghĩa AI là sự mô phỏng trí tuệ của con người thông qua máy móc, đặc biệt là qua hệ thống máy tính. Định nghĩa này có nguồn gốc từ Alan Turing (1950), khi ông cho rằng bản chất của tư duy con người giống như một cỗ máy có các tác vụ, quy trình xác định và máy móc có thể mô phỏng quá trình như thế. Nhà khoa học máy tính Gelernter cũng định nghĩa AI là "một chương trình máy tính có khả năng tái sản xuất một số khía cạnh của nhận thức con người" (Gelernter 1994: 44).

Tuy nhiên, vấn đề nằm ở chỗ, "trí tuệ con người" là một thuật ngữ khó có thể xác định. Trong phạm vi của bài nghiên cứu, thuật ngữ này có thể được hiểu là "lý trí" – tức là khả năng tư duy trừu tượng, lý luận, suy luận lôgíc, giải quyết vấn đề, ra quyết định, thay đổi và thích ứng với các điều kiện bên ngoài. Thế nhưng, khó có thể coi một cỗ máy sở hữu những năng lực tư duy như trên là một dạng "trí tuệ" đầy đủ, vì bên cạnh đó còn có "trí tuệ cảm xúc" – một dạng trí tuệ thấu hiểu bản chất của vấn đề, đối tượng bằng các quan năng cảm xúc. Vì thế, khả năng tính toán của con người không bằng AI, nhưng không thể vì thế mà nói rằng AI có chức năng trí tuệ như con người được. Thêm nữa, nhiều nhà triết học, tiêu biểu như Jackson (1982) còn cho rằng, con người sở hữu những trải nghiệm trực quan, một kinh nghiệm bên trong ("qualia") góp phần tạo nên trí tuệ, mà không thể quy về những cấu thành vật chất được. Hay nói cách khác, AI có thể sở hữu "trí tuệ" bằng cách mô phỏng tư duy con người, nhưng lại không có "trải nghiệm trực tiếp" về tư duy ấy. Cho nên AI không thể nào có được trí tuệ đầy đủ như con người.

Tuy nhiên, sự phát triển của học máy[1]

1 Học máy là quá trình mà máy móc không ngừng cập nhật và học tập các dữ liệu, những khuôn mẫu sẵn có thông qua các thuật toán để có thể đưa ra quyết định với sự can thiệp tối thiểu (hoặc không cần sự can thiệp) của con người (Antunes và cộng sự 2024: 14-16).

(machine learning), mạng thần kinh nhân tạo[2] (artifical neutral networks), học sâu[3] (deep learning) đã phá vỡ quan niệm nói trên, khi dù chỉ với những dữ liệu đầu vào hữu hạn AI có thể điều chỉnh, thích ứng với môi trường để tự cập nhật thêm dữ liệu. Quá trình này tương tự như quá trình con người tích lũy tri thức thông qua kinh nghiệm, và vì thế hiện nay phần mềm chatGPT có thể mô phỏng "kinh nghiệm" của con người để trò chuyện và phản hồi người dùng.

3. Khái niệm về tính tự trị của trí tuệ nhân tạo

Với sự phát triển của AI, khái niệm về "tính tự trị" của nó cũng được quan tâm luận giải. Tùy theo các trường phái khác nhau, có nhiều quan niệm khác nhau về "tính tự trị". Một cách chung nhất có thể hiểu tính tự trị là sự tự điều hành, tự quản (Darwall 2006: 263-264). Theo đó, AI với một mức độ phát triển nhất định trong khả năng của mình cũng có khả năng tự điều hành, tự quản, ở khía cạnh là, chúng có thể hoạt động và vận hành độc lập với những yếu tố ngoại cảnh, kể cả với con người. Theo đó, tự trị khác với *tự động*. Tự động là hoạt động theo các khuôn mẫu sẵn có, còn tự trị là khả năng tự đề ra khuôn mẫu rồi tự tuân theo khuôn mẫu ấy. Sự tự trị của AI có nhiều cấp độ khác nhau, từ cấp độ thấp nhất là việc chúng có thể lựa chọn, tiếp nhận các dữ liệu đầu vào để hình thành các khuôn mẫu, cho đến cập độ tầm trung là khả năng xử lý, sáng tạo khuôn mẫu mới, và cấp độ cao nhất là cải tạo những điều kiện bên ngoài sao cho phù hợp với những điều kiện bên trong sẵn có nhằm tạo ra các khuôn mẫu phù hợp, đồng thời còn có thể bác bỏ các khuôn mẫu đã tạo sẵn trong trường hợp cần thiết.

2 Mạng thần kinh nhân tạo là mô hình học máy mô phỏng mạng thần kinh của não người. Chúng bao gồm nhiều đơn vị xử lý cơ bản, các nút truyền dẫn, thường được gọi là "nơ-ron" nhân tạo liên kết với nhau, có khả năng xử lý thông tin, truyền dẫn tín hiệu. Mỗi liên kết của mạng nơ-ron đều là một điểm nút có thể điều chỉnh, và khi điều chỉnh nó sẽ ảnh hưởng đến toàn bộ các điểm nút khác, tạo ra sự thay đổi tương ứng. Vì thế, mô hình này có khả năng xử lý thông tin phức tạp, trừu tượng một cách hiệu quả thông qua quá trình học. Nó có thể được sử dụng để giải quyết các vấn đề phức tạp như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên và nhiều ứng dụng khác, nhất là việc mô phỏng quá trình sáng tạo và tư duy trừu tượng của con người (Antunes và cộng sự 2024: 17-19).

3 Học sâu là việc sử dụng các mạng thần kinh có nhiều lớp xử lý để máy móc tiếp thu và phân tích các dữ liệu đầu vào. Chính vì có nhiều lớp khác nhau cho nên chúng được gọi là "học sâu". Mỗi lớp trong hệ thống mạng học sâu biểu diễn một cấp độ trừu tượng hóa khác nhau: các lớp thấp hơn có thể học các đặc trưng đơn giản và rõ ràng, trong khi các lớp cao hơn học các đặc trưng phức tạp trừu tượng hơn. Học sâu mô phỏng một quá trình trí tuệ phức tạp của nhân loại thông qua cách kết hợp và xử lý thông tin từ các lớp khác nhau, đem lại cho máy móc khả năng tư duy trừu tượng ở một mức độ nhất định (Antunes và cộng sự 2024: 19-22).

Tuy nhiên, để có được tính tự trị như vậy, rõ ràng, AI phải được cung cấp mọi điều kiện cần thiết để nó tự nâng cấp và hoàn thiện. Và trước mắt, hiện nay (2024) điều kiện như thế đang do con người cấp cho. Chính vì thế có thể nói rằng tính tự trị của AI trong chừng mực nhất định phụ thuộc vào tính tự trị của con người[4]. Vì lý do này Floridi và Cowl (2019) mới quan ngại rằng khi con người giao cho AI quyền tự trị, nhất là về khả năng ra quyết định, thì "sự phát triển tính tự trị nhân tạo có thể làm hao tổn sự đầy đủ trong tính tự trị của nhân loại". Nghĩa là nếu con người cho phép AI có tính tự trị thì đồng thời cũng làm giảm đi tính tự trị của chính mình.

Tựu chung lại, tính tự trị của AI là khả năng độc lập nhất định của nó với các điều kiện bên trong và bên ngoài. Tính tự trị ấy có mối quan hệ trực tiếp với con người vì AI tồn tại trong xã hội loài người. Cho nên, xem xét tính tự trị của AI thì cũng đồng thời phải xem xét về tính tự trị của con người, đó là hai vấn đề trực tiếp liên quan đến nhau. Đó chính là cơ sở để phần sau, bài nghiên cứu sẽ xem xét về tính tự trị của con người, nhưng không phải tính tự trị nói chung mà chỉ giới hạn lại trong phạm vi đạo đức học Kant đặt ra.

4. Tính tự trị, ý chí và mệnh lệnh tuyệt đối

Xuất phát điểm của đạo đức học Kant là những giá trị nội tại của các mệnh lệnh đạo đức. Một mệnh lệnh là điều ta bắt buộc phải tuân theo, nhưng theo Kant, ta tuân theo một mệnh lệnh không phải là vì nó hướng tới mục tiêu nào, mà ta tuân theo mệnh lệnh ấy vì chính bản thân nó, độc lập với những mục đích bên ngoài. Sở dĩ con người có thể tuân theo là vì theo Kant con người là thực thể có lý trí; con người hiểu được nội dung của những mệnh lệnh ấy phù hợp với ý chí của bản thân. Con người thấy cần thiết phải làm như vậy không những là để duy trì đời sống của cá nhân mà còn là các giá trị chung của toàn nhân loại. Tuy nhiên, chỉ có lý trí là chưa đủ, ta cần phải có ý chí – khả năng điều chỉnh hành vi dù gặp phải những trở ngại. Nhưng không chỉ dừng lại ở ý chí, ta phải có sự tự trị của ý chí. Chỉ có sự tự trị của ý chí thì con người mới có thể tuân theo các mệnh lệnh một cách nhất quyết, tức một cách tất yếu. Trong *Thiết lập nền tảng cho Siêu hình học về Đạo* đức (Groundwork of the Metaphysic of Morals), Kant viết: "sự tự trị của ý chí là thuộc tính của ý chí mà tự nó trở thành một quy luật

4 Khái niệm này sẽ được phân tích rõ thêm ở bên dưới

(độc lập với bất kỳ thuộc tính nào là đối tượng của sự ham muốn)" (Kant 1996: 89).

Như vậy, Kant hiểu tính tự trị ở góc độ ý chí, ở góc độ tự ban bố và thiết lập các quy luật và tuân theo nó sao cho phù hợp với mọi ý chí, của cá nhân cũng như của người khác. Điều khiến cho một sinh thể lý tính khác biệt so với mọi sinh thể khác, chính là ở việc, nó tự đặt ra cho mình một mục đích tự thân để tuân theo (Kant 1996: 86). Khi đặt ra một mục đích tự thân, nó cũng tôn trọng những giá trị phổ quát mà chính nó thuộc về, và không làm gì để vi phạm những giá trị như vậy, bởi một cách hiển nhiên con người với lý trí hiểu rõ mình luôn phụ thuộc vào người khác và vào cộng đồng. Chính vì thế, Kant đã liên hệ tính tự trị của ý chí với đạo đức, cho rằng đạo đức là việc ý chí tự đặt ra và tuân theo các mệnh lệnh tuyệt đối - các bổn phận có tính bắt buộc và không có ngoại lệ trong mọi trường hợp. Cho nên, khi có tính tự trị, con người ta sẽ mang trong mình đức hạnh và ngược lại, khi sở hữu trong mình đạo đức, tức là ta đang mang trong mình tính tự trị.

Kant (1996: 84) đã khái quát hóa tư tưởng nói trên thành một mệnh lệnh nhất quyết[5]: "hãy hành động chỉ khi *với mọi thời điểm ý chí tự coi chính nó có khả năng đưa ra các quy luật phổ quát thông qua các châm ngôn*". Các quy luật phổ quát là những điều mà ai cũng có thể làm theo được, còn các châm ngôn theo Kant là các nguyên tắc chủ quan của hành động. Lấy một ví dụ về nói dối, Kant cho rằng, nói dối có thể là một nguyên tắc chủ quan, nhưng không thể trở nên phổ quát vì không phải ai cũng làm theo được, bởi vì đã là một con người có lý trí thì ít nhất họ phải thành thật với bản thân, và ít nhất một lần trong đời họ phải thành thật với người khác. Thành thật là cơ sở căn bản trong giao tiếp, nếu ai cũng nói dối thì thành ra không ai có thể trao đổi, giao tiếp được gì. Cho nên, nói dối không thể trở thành một mệnh lệnh nhất quyết, mà ngược lại, không được nói dối mới là mệnh lệnh nhất quyết, vì ai, trong bất kể trường hợp nào, cũng có thể làm theo (Kant 1996: 71-72).

Tựu chung lại, tính tự trị nằm ở tính tự trị của ý chí, với khả năng đồng thời ban bố và tuân theo các mệnh lệnh tuyệt đối.

5. Tính tự trị và mục đích tự thân

Phần trên ta đã nói về mệnh lệnh tuyệt đối như một quy luật phổ quát. Nhưng để trở thành quy luật phổ quát

5 Kant đề ra nhiều mệnh lệnh nhất quyết khác nhau nhưng trong giới hạn phạm vi của bài viết, chúng tôi chỉ xem xét các mệnh lệnh nhất quyết có liên quan tới tính tự trị.

thì trước tiên, khi mỗi sinh thể lý tính tự đặt ra cho mình mệnh lệnh tuyệt đối, nó phải phù hợp với các mục đích tự thân của chính sinh thể ấy. Và sau đó, để trở thành phổ quát, thì mục đích tự thân của mỗi sinh thể cũng phải phù hợp với mục đích tự thân của tất cả các sinh thể có lý tính khác. Theo đó, mỗi sinh thể lý tính sẽ không được coi sinh thể lý tính khác như phương tiện cho mục đích tự thân của mình, mà phải chính như một mục đích tự thân. Cho nên, Kant trình bày thêm một mệnh lệnh tuyệt đối khác là: "Hãy hành động chỉ khi ở mọi lúc ta coi nhân tính (humanity), dù của bản thân hay người khác, như một mục đích tự thân, không bao giờ coi nhân tính như một phương tiện" (1996: 80).

Chính từ cơ sở này, Kant (1996: 83) gọi một liên hiệp của những sinh thể lý tính cùng chia sẻ các mệnh lệnh tuyệt đối là *vương quốc của những mục đích tự thân*. Ở trong vương quốc này, mỗi sinh thể lý tính đều có được tính tự trị, vừa là chủ thể ban bố các mệnh lệnh nhất quyết, vừa là chủ thể tuân theo nó. Tuy nhiên, vì có tính tự trị ý chí, cho nên mỗi sinh thể lý tính sẽ không thể quyết định hay áp đặt ý chí của mình cho các sinh thể khác, và đồng thời, mỗi sinh thể lý tính sẽ không chịu sự quyết định hay áp đặt của các sinh thể khác. Mỗi sinh thể sẽ chỉ tự coi mình là kẻ ban bố ra mệnh lệnh tuyệt đối, cho chính bản thân cũng như cho người khác và tuân theo nó, *nhưng không áp đặt người khác tuân theo nó*, bởi vì chính bản thân mệnh lệnh ấy cũng cần trở nên phù hợp để người khác, với sự độc lập và tự chủ trong ý chí của mình, tự làm theo các quy luật đó.

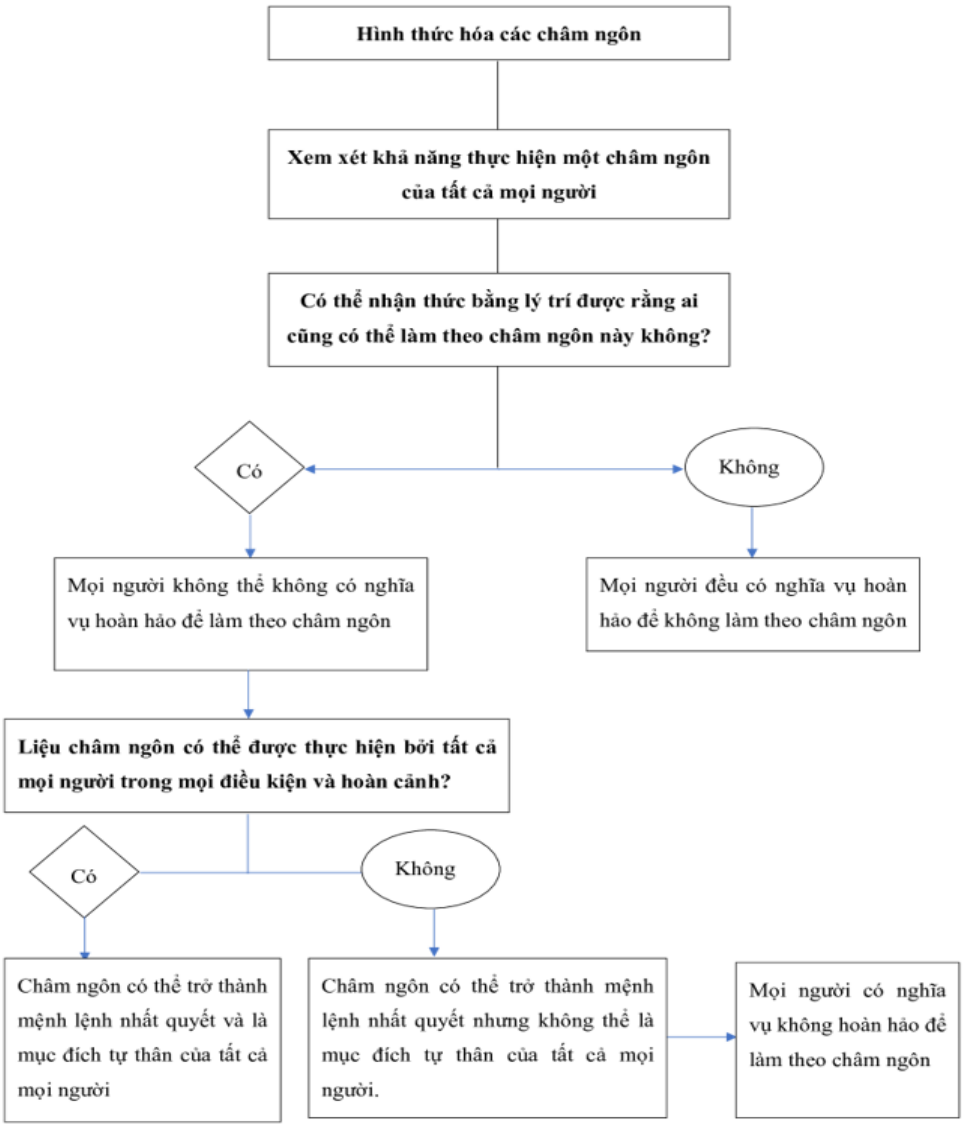
Như vậy, ở trong vương quốc đó mỗi một chủ thể sẽ mang trong mình những nghĩa vụ đạo đức và phải thực hiện nó. Có những nghĩa vụ đạo đức là hoàn hảo khi ai cũng có thể làm theo được trong mọi điều kiện (miễn là có lý trí và cấu trúc nhận thức thông thường), song vẫn có nghĩa vụ đạo đức không hoàn hảo khi ai cũng có thể làm theo được nhưng chỉ khi có những điều kiện xác định. Những nghĩa vụ đạo đức hoàn hảo là bắt buộc ai cũng phải làm theo, còn những nghĩa vụ đạo đức không hoàn hảo thì ai cũng *nên* làm theo mỗi khi *có đủ điều kiện*, và sẽ được khích lệ, ủng hộ để làm theo nó (Kant 1996: 73-74).

Nếu kết hợp giữa tính tự trị, mệnh lệnh tuyệt đối với mục đích tự thân, thì ta sẽ có một sơ đồ như dưới đây để hình thức hóa các châm ngôn để sao cho: một chủ thể có tính tự trị, và một chủ thể có tính đạo đức (như đã nói ở trên, tính tự trị về ý chí và tính đạo đức là

điều kiện lẫn nhau). Lưu ý, sự mô hình hóa này có điều kiện là: mỗi người đều là một sinh thể lý tính toàn vẹn, độc lập, có đủ điều kiện để hoàn thiện và phát triển bản thân theo những mục đích tự thân phù hợp mà không ảnh hưởng tới sinh thể lý tính khác.

6. Tính tự trị giới hạn và không đầy đủ của AI

Với những tri thức trình bày ở trên thì rõ ràng rằng, theo đạo đức học Kant, AI không có đầy đủ tính tự trị. Hay nói đúng hơn, tính tự trị của AI,



nếu có, không tương đồng với tính tự trị của con người, dù nó có thể có chức năng lý trí y hệt.

Thứ nhất, AI không có tính tự trị vì chúng *không có năng lực lý trí* như con người, dù nó có thể mô phỏng được một phần lý trí con người. Cụ thể hơn, hiện nay, AI chỉ làm nhiệm vụ là xử lý và tiếp nhận các thông tin nguyên bản, dù có biến đổi chúng thì cũng biến đổi theo những dữ liệu và các biến sẵn có. Để tiếp nhận và xử lý thông tin, AI có hai kiểu học tập chính: học tập có giám sát và học tập không có giám sát. Học tập có giám sát là phương pháp trong đó AI được cung cấp một tập dữ liệu đã được gán nhãn, nghĩa là các đầu vào đi kèm với các đầu ra mong muốn. AI sẽ học cách liên kết giữa các đầu vào và đầu ra này để có thể dự đoán chính xác cho những dữ liệu mới. Tuy nhiên, nhược điểm của học tập có giám sát là nó đòi hỏi một lượng lớn dữ liệu được gán nhãn, điều này không phải lúc nào cũng sẵn có hoặc có thể tạo ra một cách dễ dàng. Việc gán nhãn dữ liệu cũng có thể tốn kém và mất thời gian, và nếu dữ liệu được gán nhãn sai, AI sẽ học sai lệch, dẫn đến kết quả không chính xác. Học tập không có giám sát là phương pháp mà AI được cung cấp dữ liệu mà không có nhãn, và nhiệm vụ của nó là tìm ra cấu trúc hoặc mô hình của dữ liệu này. Phương pháp này mạnh mẽ hơn trong việc xử lý các tình huống thực tế, nơi mà dữ liệu gán nhãn không có sẵn, và có thể phát hiện ra những mẫu mà con người có thể không nhận ra. Tuy nhiên, nhược điểm lớn của học tập không có giám sát là độ phức tạp trong việc giải thích kết quả. Do không có nhãn để hướng dẫn, AI có thể tạo ra các nhóm hoặc mô hình không có ý nghĩa hoặc không phù hợp với mục tiêu của người dùng. Việc kiểm tra và xác nhận các kết quả từ học tập không có giám sát cũng khó khăn hơn, và đôi khi không rõ liệu mô hình mà AI làm ra có thực sự hữu ích hay không.

Một ví dụ tiêu biểu là trường hợp của một mạng nơ-ron được huấn luyện để phân loại hình ảnh thành hai nhóm: "có chứa động vật" và "không chứa động vật". Trong quá trình huấn luyện, mạng nơ-ron này đã học dựa trên một tập dữ liệu mà hầu hết các hình ảnh có chứa động vật đều có nền mờ do hiệu ứng chụp ảnh ở cự ly gần. Khi mạng nơ-ron được kiểm tra bằng những hình ảnh mới, nó bắt đầu phân loại tất cả các hình ảnh có nền mờ là "có chứa động vật", bất kể thực tế là hình ảnh đó không hề có động vật (Mitchell 2020: 120). Nguyên nhân của lỗi này là do

sự phụ thuộc quá mức của AI vào các yếu tố không liên quan trong dữ liệu huấn luyện, dẫn đến việc học sai lệch và đưa ra kết quả không chính xác[6].

Tất nhiên AI có thể thay đổi các thông tin, dữ liệu, nhưng vấn đề là ở chỗ, AI không thể tổ chức và sắp xếp thông tin sao cho nó phù hợp với thực tiễn sinh động đang diễn ra (Chalmers và cộng sự 1992). Với việc "thế giới thực không chứa đựng với các nhãn gắn liền với dữ liệu cảm giác", và AI gặp khó khăn khi phải tự tìm hiểu và nhận diện ý nghĩa từ dữ liệu cảm giác mà không có nhãn hướng dẫn (Evans 2022: 39), thì dường như AI không thể nào có được một "lý trí" đầy đủ như con người.

Để khắc phục những hạn chế nêu trên và đem lại cho AI lý trí "đích thực" như con người, Richard Evans đã phát triển "cơ chế thông giác" (Apperception Engine) như một giải pháp bổ sung cho hai phương pháp học tập truyền thống của AI: học tập có giám sát và học tập không có giám sát (Evans 2022: 40-41). Được lấy cảm hứng từ quan niệm của Immanuel Kant về nhận thức, đặc biệt là thông giác (apperception) và thông giác siêu nghiệm (transcendental apperception). Cơ chế thông giác tập trung vào việc giúp AI hiểu và tổ chức thông tin một cách tự động mà không phụ thuộc quá nhiều vào dữ liệu gán nhãn hoặc việc tìm kiếm các mẫu dữ liệu một cách mù quáng. Để hiểu rõ thêm về mô hình này, chúng ta cần khái lược quan niệm của Kant về nhận thức. Kant cho rằng nhận thức không đơn thuần là quá trình thụ động thu nhận thông tin từ thế giới bên ngoài qua các giác quan, mà còn liên quan mật thiết đến cách thức chúng ta xử lý, kết hợp và tổ chức những thông tin cảm giác đó để tạo ra ý nghĩa có thể hiểu được. Theo Kant, con người không chỉ tiếp nhận dữ liệu một cách riêng lẻ mà còn áp dụng các cấu trúc tư duy nội tại để gắn kết các dữ liệu này thành những khái niệm và tri thức có hệ thống (Kant 1998: 310-316).

Trong hệ thống triết học của Kant, nhận thức của con người chứa đựng chức năng *thông giác* (apperception),

6 Một ví dụ khác có thể được đề cập đến là khi AI được triển khai trong chế độ lái tự động của xe tự hành, xe tự hành đang sử dụng học tập không có giám sát để xác định làn đường dựa trên các dấu hiệu trực quan trên đường. Tuy nhiên, khi xe gặp phải các vạch muối được rải trên đường để chuẩn bị cho một trận bão, nó đã nhầm lẫn những vạch muối này với vạch kẻ làn đường thật. Đây là một tình huống khó dự đoán nhưng hoàn toàn có thể xảy ra, nơi mà AI gặp khó khăn trong việc phân biệt các tín hiệu trực quan bất thường mà nó chưa từng gặp trong quá trình học tập. Lỗi này xuất phát từ việc AI thiếu khả năng phân loại đúng các yếu tố chưa từng thấy hoặc hiếm gặp trong dữ liệu mà nó đã được huấn luyện (Schlicht 2022: 3-39)

tức là khả năng nhận thức và nhận biết về các trải nghiệm của chính mình. Thông giác là yếu tố cốt lõi giúp con người không chỉ nhận ra các cảm giác riêng lẻ mà còn ý thức về sự tồn tại và trải nghiệm của bản thân trong quá trình tiếp nhận những cảm giác này. Đây là một dạng nhận thức tự phản ánh, trong đó ta nhận ra và hiểu được rằng chính mình đang trải nghiệm và cảm nhận (Kant 1998: 231). Bên cạnh đó, theo Kant, nhận thức của con người còn sở hữu chức năng *thông giác siêu nghiệm* (transcendental apperception), là quá trình hợp nhất tất cả các trải nghiệm cảm giác thành một nhận thức thống nhất về bản thân trong mối liên hệ với thế giới xung quanh. Thông giác siêu nghiệm là nguyên lý đảm bảo rằng tất cả các trải nghiệm cá nhân của một người đều được liên kết với nhau trong một cấu trúc nhận thức thống nhất, cho phép con người hình thành một ý thức liên tục về bản thân và thế giới. Điều này không chỉ giúp chúng ta tổ chức và hiểu các thông tin cảm giác một cách nhất quán mà còn đảm bảo rằng mọi tri thức chúng ta có về thế giới đều được trải nghiệm như một phần của một chỉnh thể liên tục và toàn vẹn (Kant 1998: 232-233). Chính vì vậy, Kant kết luận: "chỉ có một kinh nghiệm duy nhất, trong đó tất cả các cảm giác được thể hiện theo mối liên hệ toàn vẹn và hợp quy luật, giống như chỉ có một không gian và thời gian, trong đó tất cả các hình thức hiện tượng và tất cả các mối quan hệ của sự tồn tại hoặc không tồn tại xảy ra. Nếu người ta nói về những kinh nghiệm khác nhau, chúng chỉ là nhiều tri giác khác nhau trong chừng mực chúng thuộc về một kinh nghiệm phổ quát duy nhất. Sự thống nhất toàn vẹn và tổng hợp của các tri giác chính là cái tạo nên hình thức của trải nghiệm, và nó không gì khác ngoài sự thống nhất tổng hợp của các hiện tượng dựa trên các khái niệm" (Kant 1998: 234).

Dựa trên những quan điểm cơ bản trên của Kant, Evans đã phát triển cơ chế thông giác để giải quyết những thách thức trong việc xử lý dữ liệu cho AI. Thay vì chỉ áp dụng các phương pháp học tập có hoặc không có giám sát, nơi AI thường dựa vào dữ liệu gán nhãn có sẵn hoặc cố gắng phát hiện các mẫu từ dữ liệu không có nhãn, cơ chế thông giác cung cấp một cách tiếp cận sâu sắc hơn. Cơ chế này cho phép AI xây dựng một khung nhận thức nội tại để tự động cấu trúc và hiểu thông tin từ dữ liệu kinh nghiệm, mô phỏng cách con người sử dụng các khái niệm để hình thành một nhận thức thống nhất. Cơ chế thông giác hoạt động như một "chức năng nhận thức" cho AI, giúp nó không chỉ đơn thuần xử lý dữ liệu mà còn tổ chức và

liên kết thông tin một cách có hệ thống. Thay vì dựa vào các nhãn hoặc mẫu đã được xác định trước, cơ chế thông giác cho phép AI tự xây dựng và phát triển khái niệm của riêng nó dựa trên các cảm giác và thông tin mà nó tiếp nhận. Điều này làm cho hệ thống AI có khả năng diễn giải và hiểu thông tin mới một cách linh hoạt và sáng tạo, tương tự như cách mà con người phát triển khả năng nhận thức từ các trải nghiệm cảm giác. Hơn nữa, cơ chế thông giác của Evans còn mở rộng khả năng của AI trong các tình huống phức tạp và không xác định, nơi dữ liệu huấn luyện có thể không đầy đủ hoặc không có sẵn.

Dẫu vậy, cơ chế này cũng không mang lại được cho AI năng lực lý trí đầy đủ. Một trong những hạn chế chính là việc hệ thống có thể gặp khó khăn trong việc thực hiện sự tổng hợp và nhận thức khi đối diện với dữ liệu cảm giác rất phức tạp hoặc chưa được định hình rõ ràng, nhất là với các dữ liệu liên quan đến các tình huống đạo đức. Như đã trình bày, cơ chế thông giác hoạt động dựa trên khả năng tự động xây dựng khái niệm và cấu trúc nhận thức từ dữ liệu cảm giác mà không cần dựa vào các nhãn hoặc mẫu có sẵn. Mặc dù điều này cho phép AI có khả năng linh hoạt và sáng tạo hơn trong việc xử lý thông tin, nó cũng có thể dẫn đến tình trạng mà hệ thống gặp khó khăn trong việc tổ chức và kết hợp thông tin một cách hiệu quả khi đối diện với các tình huống quá đa dạng hoặc không có quy chiếu đúng sai rõ ràng, nhất là các tình huống dựa trên đánh giá về mặt giá trị. Đặc biệt, nếu dữ liệu đầu vào quá phức tạp hoặc thiếu cấu trúc, cơ chế này có thể gặp khó khăn trong việc phát hiện và tạo ra các khái niệm hợp lý, dẫn đến việc nhận thức của AI có thể trở nên mơ hồ hoặc không chính xác. Tựu chung lại, chính vì những hạn chế nói trên, AI không thể tổ chức và sắp xếp thông tin sao cho nó phù hợp với một giá trị và mục đích tự thân nhằm định hướng mối quan hệ và góc nhìn của nó về thế giới. AI không thể hiểu và ban bố các mệnh lệnh tuyệt đối, khi nó không có tính tự trị.

Thứ hai, bất kể những hạn chế nói trên, dù có giả sử rằng AI có lý trí y như con người, thì không thể nói rằng nó có tính tự trị, vì theo Kant, tính tự trị còn phụ thuộc vào *ý chí*. Theo Kant:

"mọi thứ trong tự nhiên hoạt động theo các quy luật. Chỉ có một thực thể lý trí mới có khả năng *hành động theo sự biểu thị của các quy luật*, tức là theo các nguyên tắc, hay một *ý chí*" (Kant 1996: 66).

Ý chí không chỉ đơn thuần tuân theo một quy tắc nào đó, mà thường buộc

phải tuân theo hay phủ nhận một điều gì đó, mà nó còn là sự thấu hiểu về giá trị để không ngừng hoàn thiện bản thân sao cho phù hợp với những quy tắc ấy. AI dường như không có sự thấu hiểu về giá trị ấy khi tư duy trừu tượng do nó kiến tạo chỉ là sự dựng lại theo quy tắc một số dữ liệu để tạo ra các dữ liệu tương ứng. AI không có khả năng tự định hình các nguyên tắc hành động của mình hoặc có ý thức về quy luật mà nó đang tuân theo. AI hoạt động dựa trên các chương trình được lập trình sẵn và không có khả năng tạo ra các nguyên tắc đạo đức hoặc ý chí cá nhân để hướng dẫn hành động của mình. Quyết định của AI là kết quả của các thuật toán và dữ liệu đầu vào, không phải từ một ý chí. Hơn nữa, điều quan trọng nhất là AI tồn tại trong xã hội con người, mà con người có tính tự trị của ý chí, cho nên AI sẽ chịu sự quy định và hướng dẫn của con người. Con người là chủ thể để thiết lập những khuôn mẫu và chương trình cho AI, kể cả những nguyên tắc đạo đức và mệnh lệnh nhất quyết (Ulgen 2017).

Như vậy, về bản chất, AI là một phương tiện để con người có thể ban bố các mệnh lệnh nhất quyết, chứ tự bản thân nó không thể là một chủ thể ban bố mệnh lệnh nhất quyết. AI là một phương tiện để thúc đẩy và hoàn thiện tính tự trị của con người như tuyên bố Montreal (2017) đã chỉ ra rằng: "sự phát triển của AI phải làm gia tăng sự tự trị của tất cả con người". Tuy nhiên, vấn đề còn nằm ở chỗ, thông qua AI thì liệu các mệnh lệnh tuyệt đối của con người có thực hiện được trong mọi hoàn cảnh và trường hợp không, với giả định người thực hiện phải có lý trí toàn vẹn và đầy đủ. Rõ ràng rằng, nếu theo sơ đồ đã mô hình hóa ở trên thì không, bởi vì không phải tất cả mọi người đều tiếp cận được AI, không phải ai cũng có thể có điều kiện để sở hữu AI. Cho nên, tất cả các mệnh lệnh tuyệt đối liên quan đến AI, hiện nay (2024), dường như chỉ có thể được phổ quát hóa, chứ không thể là mục đích tự thân của tất cả mọi người, và theo đó mọi người có một nghĩa vụ không hoàn hảo để làm theo các châm ngôn liên quan đến AI.

Bởi vậy, có một vấn đề khác đặt ra là, liệu AI, nếu phù hợp với các mệnh lệnh tuyệt đối của con người, thì nó có thể được phổ quát hóa để cho tất cả mọi người sở hữu hay không. Nói cách khác, liệu AI có thể trở thành điều kiện căn bản trong đời sống của tất cả mọi người hay không. Theo chúng tôi là không, bởi vì AI khác với tất cả các điều kiện tồn tại khác như nước, đồ ăn, thức uống, v.v., nó có khả năng tự đưa ra quyết định và

nó sẽ ảnh hưởng đến tính tự trị của con người. Theo chúng tôi, tính tự trị của con người phải là căn bản nhất, phải đảm bảo cho tính tự trị đó không bị ảnh hưởng bởi bất kỳ nhân tố nào. Chỉ sau khi tính tự trị đó được đảm bảo chắc chắn, thì AI mới có thể được phổ quát hóa. Cho nên, sẽ không phải ai cũng có điều kiện ngay lập tức tiếp cận với AI. Điều này cũng phù hợp với quan điểm của Hiệp hội châu Âu về Đạo đức trong khoa học và những công nghệ tân tiến, khi họ cho rằng bất kỳ thực thể mang tính tự trị nào cũng "không thể làm suy giảm sự tự do của con người trong việc thiết lập những quy chuẩn và chuẩn mực của chính mình" (European Group on Ethics in Science and New Technologies 2018).

Tóm lại, theo đạo đức học Kant, tính tự trị của AI phải được giới hạn và nằm dưới sự tự trị của con người, trong khi đó sự tự trị của con người thì không có giới hạn và không được đặt dưới bất kỳ nhân tố nào, trong bất kỳ hoàn cảnh nào.

7. Tính tự trị của AI với mục đích tự thân của con người

Tuy tính tự trị của AI là không đầy đủ và bị giới hạn trong chừng mực tính tự trị của con người, thì vẫn có một vấn đề khác đặt ra là, liệu AI có thể mang trong mình một mục đích tự thân trong chừng mực nó thống nhất với mục đích tự thân của con người hay không. Tức là, liệu AI có thể tự đặt ra những quy tắc của riêng nó mà con người không thể can thiệp, miễn là nó phù hợp với nhân tính và không coi bất kỳ con người nào là phương tiện cho mục đích của nó. Tất nhiên, như đã nói ở trên, AI không thể đặt ra những mệnh lệnh nhất quyết được, nên AI không thể mang những mục đích tự thân nào nằm ngoài mục đích tự thân của con người. Nhưng nếu như mục đích tự thân của AI thống nhất với con người, thì dường như, trong chừng mực nhất định, ta phải tự đặt ra một mệnh lệnh nhất quyết liên quan tới việc thống nhất ấy, tức là không được phép ảnh hưởng tới điều đó.

Như thế, tuy AI có thể được coi như là một công cụ, một phương tiện cho mục đích tự thân của con người, nhưng dường như theo đạo đức học của Kant, chúng ta không nên và thậm chí không được coi nó thuần túy như một công cụ hay phương tiện. Bởi vì, trong chừng mực AI có một mục đích tự thân phù hợp với tất cả con người, thì nó có thể liên hợp với con người và qua đó tồn tại được trong vương quốc của những mục đích tự thân. Theo Kant, trong vương quốc của những mục đích tự thân, mọi thứ đều có giá cả (price) hoặc phẩm giá (digni-ty) (Kant 1996: 84). Thứ nào có giá cả thì có thể thay thế bằng những

thứ có giá trị tương đương, nhưng phẩm giá thì cao hơn hết thảy mọi giá cả và không thể được đánh đổi bằng bất kỳ cái gì. Như thế, có những AI có giá cả, nhưng có những AI sẽ mang trong mình phẩm giá vì nó phụng sự nhân tính như một mục đích tự thân cao nhất của chính nó. Cho nên, con người sẽ có trách nhiệm đạo đức và một mệnh lệnh nhất quyết để tôn trọng mục đích tự thân của những AI phụng sự nhân loại, còn với những AI không mang mục đích đó, thì nó có thể được thay thế bằng những thứ có giá cả tương đương.

Bởi vậy, những AI dùng cho những mục đích không phụng sự nhân tính nói chung như phát triển vũ khí hay chiến tranh, thì có thể được thay thể, thậm chí được phá hủy và không được phép phát triển bất kỳ cái gì liên quan đến mục đích tự thân của chúng. Còn những AI phụng sự nhân loại, như những AI dùng để phát triển tri thức, những AI đạo đức, có thể mang những mục đích tự thân, được phát triển theo hướng của riêng nó miễn là nó phụng sự nhân loại. AI có thể phụng sự và bảo toàn chính nó trong chừng mực như vậy, tức là phụng sự chính nó để đem lại sự phát triển nói chung cho nhân tính và cho toàn bộ phẩm giá của nhân loại.

8. Về tính khả thi của việc áp dụng các mệnh lệnh tuyệt đối cho AI

Tuy rằng, các mệnh lệnh tuyệt đối của Kant là có thể được ban bố với những điều kiện phù hợp, nhưng rõ ràng trong thực tế có thể có mâu thuẫn nảy sinh. Điều này đã được Benjamin Constant nêu ra khi phê phán Kant, khi ông cho rằng, việc tuân theo mệnh lệnh nhất quyết là không khả thi. Ông lấy ví dụ về việc có một kẻ sát nhân đang truy đuổi một người bạn của chúng ta, và theo mệnh lệnh nhất quyết của Kant thì ta phải có nghĩa vụ nói lên sự thật về vị trí của người bạn đó ở đâu, kể cả đối phương có là kẻ sát nhân hay không. Nhưng như thế là vô lý, vì ông cho rằng, việc liệu nói dối có sai hay không phụ thuộc vào hoàn cảnh cụ thể và đối tượng chúng ta đang nói dối. Những kẻ sát nhân không có quyền đòi hỏi sự thật, vì vậy không ai có nghĩa vụ phải nói sự thật với chúng. Do đó, Constant kết luận rằng việc nói dối với những kẻ sát nhân không nên bị coi là phi đạo đức (Dẫn theo Kant 1996: 161). Phân tích kỹ hơn, có thể thấy, không chỉ là vô lý, điều này còn tạo nên một mâu thuẫn giữa nghĩa vụ nói lên sự thật và nghĩa vụ bảo toàn sự sống của người khác. Nếu nói thật, ta còn vô tình tiếp tay cho kẻ sát nhân thực hiện hành vi của mình, như thế là ta không có đạo đức. Còn nếu nói dối, ta lại không tuân theo nghĩa vụ nói lên sự thật mọi lúc, như

thế ta cũng không có đạo đức. Theo lý thuyết, con người tuân theo hai mệnh lệnh nhất quyết ấy sẽ đạo đức, nhưng trên thực tế, khi tuân theo con người lại vô đạo đức. Do vậy, mệnh lệnh nhất quyết nếu không thể áp dụng được cho con người, thì cũng không thể áp dụng được cho AI.

Tuy nhiên, nếu phân tích kỹ hơn, ta có thể thấy rằng, khi kẻ sát nhân hỏi, ta hoàn toàn có thể từ chối trả lời. Trên thực tế, việc từ chối trả lời các câu hỏi từ những kẻ sát nhân không có nghĩa rằng ta đã vi phạm nghĩa vụ đạo đức. Thay vào đó, chủ nhà có thể đơn giản yêu cầu kẻ sát nhân rời đi, vì việc ai đang ở trong nhà của mình không phải là vấn đề của kẻ sát nhân. Theo đó, ta vừa có thể tuân theo hai mệnh lệnh nhất quyết cùng một lúc: phải nói lên sự thật rằng "ta không phải nói", và phải bảo toàn mạng sống. Bởi vì, nói sự thật không yêu cầu rằng một người phải tiết lộ thông tin cho bất kỳ ai, đặc biệt là cho những người lạ hoặc kẻ sát nhân, cũng như không bác bỏ quyền riêng tư của một cá nhân. Ngoài ra, có thể đáp ứng vấn đề bằng cách nói rằng theo lý thuyết của Kant, chúng ta có thể trả lời: "đúng thế, bạn tôi đang ở trong nhà, nhưng ông không được phép vào nhà tôi". Tuy nhiên, Kant không xét trường hợp đó, mà ông lại xét một trường hợp triệt để hơn: một người bị ép buộc phải nói điều gì đó để tránh gây hại cho bản thân hoặc người khác một cách bất công, và những trường hợp mà người trả lời cửa không có lựa chọn yêu cầu kẻ sát nhân ra đi.

Kant trả lời rằng trong trường hợp này, nghĩa vụ trung thực vẫn được áp dụng một cách tuyệt đối. Ông lập luận rằng ngay cả khi một người bị ép buộc phải nói điều gì đó để tránh gây hại cho bản thân hoặc người khác, việc nói dối vẫn không được phép. Kant cho rằng việc giữ vững nguyên tắc trung thực là cần thiết để duy trì sự tin cậy trong xã hội và để đảm bảo rằng tất cả các nghĩa vụ đạo đức đều được thực hiện đúng đắn. Kant lập luận rằng nếu ta tự ý chọn tham gia vào tình huống bằng cách nói dối, thì dù ta thực hiện hành động đó với ý định tốt, điều này cũng không làm giảm trách nhiệm của bạn đối với các hậu quả xấu xảy ra từ lời nói dối ấy. Theo Kant, khi ta nói dối, ta dường như đang chủ động can thiệp vào tình huống và chấp nhận mọi hậu quả có thể xảy ra từ hành động của mình. Điều này có nghĩa là ta phải sẵn sàng chịu trách nhiệm pháp lý nếu sự phán đoán của bạn là sai lầm và lời nói dối của bạn cuối cùng lại giúp kẻ sát nhân bắt được nạn nhân. Ngược lại, khi ta chọn trung thực và không tham gia

vào tình huống bằng cách nói sự thật, ta không can thiệp vào diễn biến tiếp theo của sự việc. Theo đó ta để mở khả năng xảy ra, bao gồm việc nạn nhân có thể lén ra ngoài mà kẻ sát nhân không phát hiện (nếu như nói dối có thể vô tình dẫn đến kẻ sát nhân ra ngoài và tìm ra nạn nhân), hoặc ta và hàng xóm có thể cùng nhau khống chế, vô hiệu hóa hoặc thậm chí phản kháng kẻ sát nhân trong quá trình tìm kiếm, hoặc cảnh sát có thể đến kịp thời để ngăn chặn kẻ sát nhân.

Nói cách khác, theo quan điểm của Kant, việc lựa chọn nói dối có thể dẫn đến việc ta phải chịu trách nhiệm về các hậu quả không mong muốn nếu sự phán đoán của mình là sai lầm và hành động của bạn thực sự làm tổn hại đến nạn nhân. Việc trung thực không chỉ giúp giữ vững các mệnh lệnh đạo đức mà còn bảo đảm rằng ta không tự ý can thiệp vào diễn biến của sự việc, để cho các phương pháp hợp pháp khác, như sự can thiệp của cảnh sát hoặc các hành động của cộng đồng, có thể giải quyết tình huống mà không bị ảnh hưởng bởi sự can thiệp của ta (Kant 1996: 611-615). Như vậy, từ quan điểm của Kant, có thể kết luận rằng các mệnh lệnh tuyệt đối, chẳng hạn nguyên tắc trung thực tuyệt đối, là khả thi trong thực tế. Mặc dù các mệnh lệnh tuyệt đối có thể dẫn đến những tình huống khó khăn và hậu quả không ngờ tới, nhưng việc giữ vững nguyên tắc đạo đức là cần thiết, vì chúng cung cấp một nền tảng vững chắc cho hành vi đạo đức và sự tin cậy trong các mối quan hệ xã hội, cũng như tạo điều kiện để các giải pháp phù hợp có thể được áp dụng.

Dẫu vậy, việc áp dụng các mệnh lệnh tuyệt đối của Kant vào AI là một vấn đề phức tạp và khó thực hiện. Phải nói rằng, AI hiện tại (2024) vẫn còn thiếu khả năng đánh giá tình huống một cách toàn diện như con người. Khi AI phải ra quyết định dựa trên mệnh lệnh tuyệt đối là trung thực trong mọi tình huống, chẳng hạn như khi phải trả lời một kẻ sát nhân về sự hiện diện của một nạn nhân trong nhà, hệ thống AI không thể dự đoán tất cả các kịch bản có thể xảy ra từ quyết định của nó. Ví dụ, AI có thể không biết liệu nạn nhân có thể lén ra ngoài mà không bị phát hiện, hoặc liệu các cơ quan chức năng có đến kịp thời để ngăn chặn kẻ sát nhân hay không. Ngược lại, nếu AI chọn cung cấp thông tin sai, dù nhằm mục đích bảo vệ nạn nhân, nó cũng gặp phải các vấn đề tương tự về trách nhiệm. Kant cho rằng việc nói dối có thể dẫn đến trách nhiệm pháp lý nếu lời nói dối đó gây hại cho người khác. Tuy nhiên, AI không có khả năng chịu trách nhiệm pháp lý hay hiểu biết đầy

đủ về các yếu tố pháp lý và đạo đức liên quan.

Vì thế, để áp dụng mệnh lệnh nhất quyết của Kant vào trí tuệ nhân tạo (AI) một cách khả thi, cần phải có một cách tiếp cận toàn diện. Mệnh lệnh tuyệt đối, theo Kant, yêu cầu hành động dựa trên các mệnh lệnh đạo đức mà mọi người có thể chấp nhận một cách phổ quát. Đối với AI, điều này có nghĩa là nó cần phải hoạt động không chỉ dựa trên các mệnh lệnh nhất quyết mà còn phải phục vụ nhân loại một cách cao cả và bảo vệ phẩm giá con người.

Trước tiên, cần phải xác định rõ mục đích tự thân của AI. Mục đích này phải được định hình để đảm bảo rằng AI hoạt động vì nhân tính và không chỉ đơn thuần là thực hiện các nhiệm vụ theo yêu cầu. Ví dụ, nếu AI được phát triển để hỗ trợ trong lĩnh vực y tế, mục đích tự thân của nó có thể là không ngừng cải thiện sức khỏe và nâng cao chất lượng cuộc sống của con người. Điều quan trọng là mục đích này phải phản ánh sự phục vụ cho nhân loại và không được bị ảnh hưởng bởi các lợi ích cá nhân hay thương mại.

Sau khi mục đích tự thân được xác định, việc thiết kế các nguyên tắc đạo đức cho AI là bước tiếp theo. Các nguyên tắc này phải bao gồm việc tôn trọng phẩm giá con người, bảo vệ quyền riêng tư, và đảm bảo tính trung thực trong các tương tác. Ví dụ, nếu AI phải xử lý dữ liệu cá nhân, nó cần phải được lập trình để bảo vệ thông tin này khỏi sự lạm dụng và đảm bảo rằng các quyết định được đưa ra dựa trên thông tin chính xác và không thiên lệch. Nguyên tắc đạo đức cần phải được tích hợp vào các thuật toán và quy trình hoạt động của AI để đảm bảo rằng nó hoạt động theo các chuẩn mực đạo đức. AI cần được lập trình không chỉ để thực hiện các nhiệm vụ mà còn để xử lý các tình huống đạo đức phức tạp một cách chính xác. Điều này có thể bao gồm việc cung cấp cho AI khả năng học hỏi từ các tình huống mô phỏng, trong đó nó phải đưa ra các quyết định dựa trên các nguyên tắc đạo đức. Chẳng hạn như, nếu AI phải quyết định liệu có nên tiết lộ thông tin nhạy cảm trong một tình huống khẩn cấp, nó cần phải có khả năng cân nhắc các yếu tố đạo đức và pháp lý trước khi đưa ra quyết định. Theo đó, việc xây dựng cơ chế giám sát và đánh giá liên tục là cần thiết để đảm bảo rằng AI thực hiện các nguyên tắc đạo đức một cách chính xác. Cơ chế này phải bao gồm việc thiết lập các tiêu chuẩn và quy trình kiểm tra để đánh giá hiệu quả hoạt động của AI và xử lý các vấn đề phát sinh khi nó không tuân thủ các nguyên tắc đạo đức. Từ đó, nếu AI đưa ra một quyết định sai lầm

dẫn đến hậu quả không mong muốn, cơ chế giám sát cần phải có khả năng phát hiện và điều chỉnh các sai sót này một cách kịp thời.

Cơ chế giám sát này hoàn toàn có thể được thực hiện thông qua một hệ thống AI khác, được lập trình để tuân thủ các nghĩa vụ đạo đức rõ ràng. Điều này có nghĩa là bản thân hệ thống giám sát cũng phải tuân theo một bộ quy tắc đạo đức được xác định từ trước, với khả năng nhận biết và đánh giá hiệu suất của các hệ thống AI mà nó giám sát. Sự phát triển của một AI có khả năng giám sát các hệ thống khác có thể giúp giảm thiểu khối lượng công việc của con người trong quá trình giám sát liên tục, đồng thời đảm bảo rằng việc theo dõi diễn ra chính xác, nhất quán và không bị chi phối bởi yếu tố cảm xúc hay thiên kiến cá nhân.

Tuy nhiên, dù AI có thể đảm nhiệm vai trò giám sát một cách tự động và hiệu quả, người giám sát cuối cùng vẫn phải là con người. Điều này đảm bảo rằng các quyết định mang tính đạo đức cuối cùng được xem xét dưới góc độ nhân văn. Trách nhiệm của con người trong quá trình giám sát là không thể thay thế, vì chỉ con người mới có khả năng hiểu sâu sắc những giá trị đạo đức, ý nghĩa văn hóa và các yếu tố xã hội phức tạp mà AI, dù có phát triển đến đâu, khó có thể hoàn toàn nắm bắt được. Hơn nữa, sự can thiệp của con người là cần thiết trong các tình huống mà AI có thể gặp khó khăn trong việc đưa ra phán đoán chính xác, chẳng hạn như các tình huống đạo đức mơ hồ hoặc phức tạp không có một giải pháp nhất quán và duy nhất.

Hơn nữa, chính vì mệnh lệnh tuyệt đối là do con người ban bố, nên việc đào tạo và giáo dục cho những người thiết kế và vận hành AI về các nguyên tắc đạo đức là điều tối quan trọng. Những người này không chỉ cần hiểu rõ các nguyên tắc đạo đức mà còn phải có khả năng áp dụng chúng một cách có lý trí và tự chủ trong quá trình phát triển và triển khai AI. Điều này đòi hỏi sự kết hợp giữa kiến thức lý thuyết và khả năng thực tiễn, trong đó họ không chỉ tuân theo các chuẩn mực đạo đức đã được đề ra, mà còn cần phát triển khả năng suy xét độc lập, tự đưa ra các quyết định đúng đắn trong những tình huống đạo đức phức tạp hoặc chưa có tiền lệ.

Việc tổ chức các khóa đào tạo và chương trình giáo dục cần phải chú trọng vào việc rèn luyện tư duy tự chủ và khả năng nhận thức tự giác về trách nhiệm cá nhân, để mỗi cá nhân tham gia vào quá trình phát triển AI có thể hành động với ý thức tự trị, không phụ thuộc hoàn toàn vào các quy định cứng

nhắc. Điều này giúp họ có thể đánh giá và điều chỉnh hành động của mình dựa trên sự suy xét kỹ lưỡng về hậu quả và tác động đạo đức. Giáo dục không chỉ là truyền thụ kiến thức mà còn là quá trình hình thành nhân cách đạo đức, nơi các nhà phát triển không chỉ đơn thuần tuân thủ quy tắc mà còn tự giác hành động vì lợi ích chung, với trách nhiệm đối với xã hội và nhân loại. Ngoài ra, việc giáo dục còn cần tạo điều kiện cho họ hiểu sâu về tầm quan trọng của tự do trong việc lựa chọn đạo đức và nhận thức về quyền tự chủ của con người, nhằm tránh tình trạng phụ thuộc máy móc vào các hệ thống AI mà họ phát triển. Điều này khuyến khích một cách tiếp cận mang tính sáng tạo và có trách nhiệm trong công nghệ, đảm bảo rằng AI được phát triển không chỉ vì lợi ích kinh tế hay kỹ thuật mà còn vì sự tiến bộ của xã hội trong khuôn khổ các giá trị đạo đức lâu dài.

Cuối cùng, cần xây dựng các chính sách và quy định pháp lý liên quan đến việc phát triển và sử dụng AI. Các quy định này nên bao gồm các tiêu chuẩn về bảo vệ quyền riêng tư, tôn trọng phẩm giá con người, và minh bạch trong việc sử dụng AI. Cụ thể, tiêu chuẩn như làm điều tốt (beneficence) và tránh gây hại (nonmaleficence) phải được ưu tiên hàng đầu, đảm bảo rằng AI không chỉ mang lại lợi ích cho xã hội mà còn giảm thiểu tối đa các rủi ro và thiệt hại có thể xảy ra. Bên cạnh đó, tiêu chuẩn về tính tự trị (autonomy) cũng cần được tôn trọng, đảm bảo rằng người dùng và các cá nhân chịu ảnh hưởng bởi AI có quyền kiểm soát và đưa ra các quyết định liên quan đến cuộc sống của họ. Điều này yêu cầu các hệ thống AI phải hoạt động minh bạch và rõ ràng, cung cấp đầy đủ thông tin về cách thức hoạt động và ra quyết định (explicability), để người dùng có thể hiểu và đánh giá được những ảnh hưởng của AI đến họ. Hơn nữa, tiêu chuẩn về công lý (justice) cũng cần phải được xem xét cẩn thận, nhằm đảm bảo rằng AI không tạo ra sự bất bình đẳng hay phân biệt đối xử trong quá trình áp dụng. Các chính sách pháp lý cần đảm bảo rằng AI được phát triển và sử dụng một cách bình đẳng, bảo vệ quyền lợi của tất cả các nhóm trong xã hội, bất kể về sắc tộc, giới tính, hay địa vị kinh tế (Floridi và Cowls 2019). Tất cả những nguyên tắc này cùng với các chính sách cụ thể sẽ tạo ra một môi trường pháp lý rõ ràng và công bằng cho việc phát triển và sử dụng AI, đảm bảo rằng AI không chỉ hoạt động trong khuôn khổ pháp lý mà còn tuân thủ các nguyên tắc đạo đức nghiêm ngặt, bảo vệ lợi ích của con người và xã hội.

9. Kết luận

Trong bối cảnh công nghệ đang không ngừng phát triển với tính tự động hóa ngày càng cao cũng như khả năng ra quyết định ngày càng độc lập, tính tự trị của trí tuệ nhân tạo (AI) là một vấn đề đáng được quan tâm. Dựa trên các nguyên tắc đạo đức học của Kant, AI cần phát triển không chỉ với khả năng tự quyết định mà còn phải tuân theo một bộ quy tắc đạo đức rõ ràng, với mục đích cao cả là phục vụ lợi ích chung của nhân loại. Tuy nhiên, sự tự trị này không thể tách rời sự giám sát và hướng dẫn của con người, mà cần nhấn mạnh tầm quan trọng của việc duy trì một mối quan hệ cân bằng và có trách nhiệm giữa con người và máy móc. Cuối cùng, việc khám phá tính tự trị của AI theo quan điểm của Kant mở ra những cơ hội mới cũng như trách nhiệm mới, đòi hỏi chúng ta phải suy ngẫm sâu sắc về tương lai mà chúng ta đang hình thành.

Tài liệu trích dẫn

1. Etzioni & O.Etzioni. 2016. "AI assisted ethics". *Ethics and Information Technology* 18(2): 149–156.

2. A.M.Turing. 1950. "Computing machinery and intelligence". *MIND: A Q Rev Pyschol Philos* 59(236): 433–460.

3. D.Chalmers, R.M.French & D.Hofstadter. 1992. "High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology". *Journal of Experimental and Theoretical Artificial Intelligence* 4(3): 185–211.

4. D.Gelernter. 1994. *The Muse in the Machine: Computerizing the Poetry of Human Thought*. New York: Free Press.

5. European Group on Ethics in Science and New Technologies. 2018. *Statement on Artificial Intelligence, Robotics and "Autonomous" Systems*.

6. F. Jackson. 1982. "Epiphenomenal Qualia". *The Philosophical Quarterly* 32(127): 127–136. (https://doi.org/10.2307/2960077).

7. H.S.Antunes et al. 2024. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. Cham: Springer Nature.

8. I.Kant. 1996. "Groundwork of the Metaphysic of Morals". Pp. 37–109 in I. Kant, *Practical Philosophy*, translated and edited by Mary J. Gregor. Cambridge: Cambridge University Press.

9. Kant. 1996. "On A Supposed Right To Lie from Philanthropy". Pp. 605–617 in I. Kant, *Practical Philosophy*, translated and edited by Mary J. Gregor. Cambridge: Cambridge University Press.

10. Kant. 1998. *Critique of Pure Reason*, translated and edited by P. Guyer & A.W. Wood. Cambridge: Cambridge University Press.

11. J. Moor. 2008. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years". *AI Magazine* 27: 4.

12. L.Floridi & J.Cowls. 2019. "A Unified Framework of Five Principles for AI in Society". *Harvard Data Science Review*.

13. M.Mitchell. 2020. *Artificial Intelligence. A Guide For Thinking Humans*. London: Penguin.

14. M.Peña-Cabrera, V. Lomas & G. Lefranc. 2019. "Fourth industrial revolution and its impact on society". *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, Valparaiso, Chile, pp. 1-6. https://doi.org/10.1109/CHILECON47746.2019.8988083.

15. O. Ulgen. 2017. "Kantian Ethics in the Age of Artificial Intelligence and Robotics". *Questions of International Law* 1(43): 59–83.

16. P.Formosa. 2021. "Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy". *Minds & Machines* 31: 595–616. https://doi.org/10.1007/s11023-021-09579-2.

17. P.S.Park, S.Goldstein, A. O'Gara, M. Chen & D. Hendrycks. 2024. "AI deception: A survey of examples, risks, and potential solutions". *Patterns* 5(5): 100988. https://doi.org/10.1016/j.patter.2024.100988.

18. Phương Linh. 2024. "Nhà văn Nhật thắng giải văn học dù dùng ChatGPT." *VNExpress*. (https://vnexpress.net/nha-van-nhat-thang-giai-van-hoc-du-dung-chatgpt-4703530.html) truy cập ngày 27/01/2024.

19. R.Evans. 2022. "The Apperception Engine". Pp. 39–105 in H. Kim & D. Schönecker (Eds.) *Kant and Artificial Intelligence*. Berlin: De Gruyter.

20. S.Darwall. 2006. "The value of autonomy and autonomy of the will". *Ethics* 116: 263–284.

21. T.Schlicht. 2022. "Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant's Approach to Cognition". Pp. 3–39 in H. Kim & D. Schönecker (Eds.) *Kant and Artificial Intelligence*. Berlin: De Gruyter.

22. U.Barthelmeß & U.Furbach. 2023. *A Different Look at Artificial Intelligence: On Tour with Bergson, Proust and Nabokov*. Wiesbaden: Springer Vieweg.

23. Université de Montréal. 2017. *Montreal Declaration for a Responsible Development of Artificial Intelligence*. (https://www.montrealdeclaration-responsibleai.com/the-declaration).