TRÍ TUỆ NHÂN TẠO VÀ ĐẠO ĐỨC HỌC HỒ CHÍ MINH

Nguyễn Anh Tuấn(*)

(*) Phó giáo sư, tiến sĩ, Trường Đại học khoa học xã hội và Nhân văn, Đại học Quốc gia Hà Nội.

Email: nguyenanhtuantr@gmail.com

Phạm Minh Đức(**)

(**) Thạc sĩ, Viện Khoa học Giáo dục và Khai phóng, Trường Đại học VinUni.

Email: duc.pm@vinuni.edu.vn.

Tóm tắt: Hiện nay, với sự phát triển mạnh mẽ, trí tuệ nhân tạo (AI) đang tạo ra những ảnh hưởng ngày càng sâu rộng đối với xã hội, đồng thời đặt ra nhiều vấn đề mang tính đạo đức. Các cách tiếp cận phổ biến liên quan đến AI có thể được phân chia thành ba trường phái chính: đạo đức học hệ quả, đạo đức học nghĩa vụ và đạo đức học phẩm chất. Mỗi trường phái đều tập trung vào những phương thức khác nhau nhằm hoặc là kiểm soát AI thông qua các giá trị đạo đức xác định, hoặc là phát triển những phẩm chất phù hợp để AI có thể tồn tại một cách hòa hợp với con người trong xã hội. Tuy nhiên, ngoài các cách tiếp cận này, đạo đức học Hồ Chí Minh, cụ thể là đạo đức học cách mạng, mang lại một góc nhìn độc đáo và sâu sắc về mối quan hệ giữa AI và con người trong xã hội hiện đại. Một đặc điểm nổi bật của đạo đức học Hồ Chí Minh là khả năng kết hợp tính trực quan, dễ hiểu với chiều sâu triết lý. Nếu được áp dụng một cách đúng đắn, những nguyên tắc này có thể mở ra tiềm năng để AI phát triển theo hướng hài hòa và nhân văn. Tính cách mạng trong đạo đức học của Hồ Chí Minh không chỉ có khả năng định hướng để AI tự hoàn thiện bản thân, mà còn giúp nó tự giác tuân theo những giá trị nhân đạo cốt lõi. Quan điểm này có thể gợi mở một viễn cảnh đầy triển vọng, nơi mà sự kết hợp giữa AI và đạo đức học Hồ Chí Minh có thể mang lại những thay đổi tích cực cho xã hội loài người.

Từ khóa: Trí tuệ nhân tạo (AI), đạo đức học Hồ Chí Minh, đạo đức học cách mạng.

Ngày nhận bài: 10/05/2025; ngày phản biện:11/05/2025; ngày sửa chữa: 25/05/2025; ngày duyệt đăng: 17/06/2025.

1. Mở đầu

Có một vấn đề về AI, đó là chúng ta thường nghĩ về nó hoặc là quá phức tạp, hoặc là quá đơn giản. Thực tế lại phức tạp hơn nhiều. Moravec đã chỉ ra rằng, "rõ ràng là việc khiến máy tính đạt hiệu suất ngang tầm con người trưởng thành trong việc giải các bài kiểm tra trí

thông minh hoặc chơi cờ vây tương đối dễ dàng, nhưng lại cực kỳ khó hoặc gần như không thể dạy chúng các kỹ năng cơ bản của một đứa trẻ một tuổi khi nói đến nhận thức và di chuyển" (Moravec 1988: 15). Như vậy, ngay cả trong các lĩnh vực vốn được coi là thách thức nhất đối với trí thông minh nhân tạo, chúng ta vẫn có xu hướng đánh giá sai sự cân bằng giữa năng lực và hạn chế của nó.

Đối với vấn đề đạo đức, tình hình cũng không khác biệt. Con người, hay các sinh vật sống, thường thực hiện các "trực giác đạo đức" hoặc các hành động đạo đức bột phát một cách dễ dàng và nhanh chóng. Những hành vi này dường như do sự phối hợp vô thức giữa bản năng, cảm xúc, kinh nghiệm, thậm chí có đôi phần trí tuệ xen lẫn. Trong khi đó, đối với AI, các giá trị đạo đức và hành động đạo đức lại giống như một bài toán hóc búa: vừa khó khăn trong việc xác định rõ ràng, vừa nan giải trong việc thực thi.

Thế nhưng, chúng ta lại thường nghĩ AI đơn giản chỉ cần được lập trình để "làm điều đúng". Thực tế, AI không thực sự "hiểu" đạo đức là gì hoặc những gì đạo đức mang lại. Tuy nhiên, một cách mâu thuẫn, chính việc "không thực sự hiểu" này lại có thể khiến AI có tiềm năng tuân theo các quy tắc đạo đức tốt hơn bất kỳ ai, nếu chúng được thiết lập một cách hợp lý và được vận hành dưới sự giám sát chặt chẽ. Nói cách khác, AI có thể trở thành công cụ mạnh mẽ để thúc đẩy các giá trị đạo đức, không phải vì nó hiểu, mà là vì nó có khả năng thực hiện chúng một cách "khách quan" và nhất quán hơn con người.

Bởi vậy, bài viết này, dựa trên nền tảng đã nêu, sẽ xem xét mối quan hệ giữa AI và đạo đức học Hồ Chí Minh theo một góc nhìn toàn diện hơn. Cụ thể, bài viết sẽ được cấu trúc theo các phần sau:

Một là, bài viết sẽ lược khảo lịch sử phát triển của đạo đức học, tập trung đặc biệt vào đạo đức học quy chuẩn (normative ethics) – lĩnh vực gần gũi nhất với việc áp dụng vào AI. Phần này sẽ đặt nền tảng lý thuyết để hiểu rõ cách các nguyên tắc đạo đức được xây dựng và áp dụng trong các hành động cụ thể.

Hai là, bài viết xem xét khả năng "tiếp thu" và "hiểu" các giá trị đạo đức cũng như các quy chuẩn hành xử của AI. Phần này sẽ phân tích sự khác biệt giữa việc "hiểu" đạo đức dưới góc độ con người và khả năng vận hành các quy tắc đạo đức của AI thông qua thuật toán và dữ liệu.

Ba là, bài viết phân tích nội dung cốt lõi của đạo đức học Hồ Chí Minh, từ đó khám phá khả năng áp dụng đạo đức học Hồ Chí Minh vào lĩnh vực AI. Cụ thể, bài viết đặt vấn đề đạo đức học Hồ Chí Minh có thể định hướng cho AI trong việc xây dựng mối quan hệ giữa các hệ thống AI với nhau, cũng như giữa AI với con người như thế nào? Phần này sẽ đánh giá các yếu tố cần thiết để AI "hiểu" đạo đức học Hồ Chí Minh theo nghĩa thực tiễn, từ đó xác định cách những giá trị này có thể góp phần định hình một mối quan hệ hài hòa, trách nhiệm và nhân văn giữa công nghệ và nhân loại.

Bốn là, bài viết đưa ra một số nhận định về triển vọng dài hạn của việc kết hợp các giá trị đạo đức Hồ Chí Minh với sự phát triển của AI, đồng thời gợi mở những hướng đi mới trong việc nghiên cứu đạo đức học trong kỷ nguyên công nghệ.

2. Lược khảo lịch sử đạo đức học quy chuẩn

Con người cần và nên sống như thế nào? Đây là một câu hỏi lớn của nhân loại, vừa mang tính lý thuyết, vừa mang tính thực tiễn. Về mặt lý thuyết, câu hỏi này đặt ra vấn đề về sự tồn tại của cái gọi là "sống" và "như thế nào", hay nói cách khác, liệu có tồn tại một "thước đo giá trị" cho sự sống hay không? Về mặt thực tiễn, nó dẫn đến một câu hỏi khác: con người nên hành động theo điều gì, hay tuân theo những nguyên tắc nào để định hình cách sống của mình?

Để trả lời những câu hỏi trên, các trường phái đạo đức học đã ra đời với những cách tiếp cận khác nhau. Đạo đức học lý thuyết (metaethics) tập trung vào vế đầu, bàn luận về bản chất của "đạo đức", liệu có tồn tại "chân lý đạo đức" khách quan hay không, và nếu có, thì nó được xác định như thế nào. Trong khi đó, đạo đức học quy chuẩn (normative ethics) hướng đến vế thứ hai - tìm cách đưa ra những quy chuẩn cụ thể để hướng dẫn hành động, xác định điều gì là đúng, điều gì là sai, hoặc cách con người nên hành xử trong cuộc sống.

Như đã trình bày, bài viết sẽ tập trung vào đạo đức học quy chuẩn, lĩnh vực mà đạo đức học Hồ Chí Minh có thể được xếp vào. Trong đó, một trong ba cách tiếp cận chính là hệ quả luận (Consequentialism), quan điểm cho rằng tính đạo đức của một hành động được xác định hoàn toàn bởi hậu quả mà hành động đó mang lại. Theo cách tiếp cận này, một hành động được coi là đúng đắn nếu nó tối đa hóa các kết quả tốt đẹp, ví dụ như hạnh phúc hay phúc lợi, so với các lựa chọn có sẵn khác (Xem: Driver, J. 2012: 5-6). Khác với hệ quả luận, đạo đức học nghĩa vụ (Deontology) nhấn mạnh tính đúng đắn của một hành động không phụ thuộc vào hệ quả, mà dựa trên bản thân hành động đó và nghĩa vụ đạo đức của người thực hiện. Theo cách tiếp cận này, có những nguyên tắc/quy tắc đạo đức bất khả xâm phạm, chẳng hạn như không được cố ý làm hại người vô tội, dù điều đó có thể dẫn đến các kết quả tốt hơn (Xem: Rawling, P. 2023: 3-4).

Một hướng khác, khác với hai hướng trên, chính là đạo đức học phẩm chất (virtue ethics) - hướng tiếp cận tập trung vào phẩm chất đạo đức của con người hơn là hậu quả của hành động hay nghĩa vụ. Theo đạo đức học phẩm chất, trọng tâm của đạo đức nằm ở việc phát triển và rèn luyện các phẩm chất tốt đẹp (virtues), những đặc điểm giúp con người sống một cuộc đời tốt đẹp và luôn hành động đúng đắn trong các hoàn cảnh khác nhau. Điểm đặc biệt của đạo đức học phẩm chất là nó không nhất thiết đòi hỏi một hệ thống quy tắc phổ quát hay chỉ tập trung vào hệ quả của hành động, thay vào đó khuyến khích một sự linh hoạt dựa trên bối cảnh, kinh nghiệm và phán đoán của cá nhân với

những phẩm chất tốt đẹp sẵn có. Theo đó, một hành động được coi là đúng đắn khi nó xuất phát từ một nhân cách tốt đẹp và phù hợp với các phẩm chất đạo đức đã được đào luyện.

Có thể nói, đạo đức học Hồ Chí Minh sẽ phù hợp với đạo đức học phẩm chất, vì Người đã nhấn mạnh việc hoàn thiện các phẩm chất như cần, kiệm, liêm, chính, chí công vô tư, những phẩm chất mà mỗi cán bộ, đảng viên và rộng hơn là mỗi người dân cần rèn luyện (Xem: Hồ Chí Minh 2011, tập 5: 622). Tuy nhiên, tư tưởng đạo đức của Hồ Chí Minh không dừng lại ở việc nhấn mạnh các phẩm chất cá nhân, mà còn coi đạo đức như nền tảng tinh thần của xã hội. Đạo đức, theo Người, không chỉ là sự tự hoàn thiện của cá nhân, mà còn là yếu tố có ảnh hưởng sâu sắc đến vận mệnh của đất nước và sự phát triển bền vững của xã hội.

Như vậy, đạo đức học Hồ Chí Minh thể hiện sự hòa trộn độc đáo giữa hệ quả luận, đạo đức học nghĩa vụ và đạo đức học phẩm chất. Hồ Chí Minh không chỉ nhấn mạnh các phẩm chất cá nhân, mà còn đề cao trách nhiệm của cá nhân đối với xã hội và tầm quan trọng của hành động vì lợi ích chung, điều gần với hệ quả luận. Đồng thời, việc Người yêu cầu mọi cá nhân "tận trung với nước, tận hiếu với dân" (Hồ Chí Minh 2011, tập 5: 354) và giữ vững lập trường đúng đắn cũng phản ánh những nguyên tắc của đạo đức học nghĩa vụ. Tuy nhiên, sự kết hợp này không phải là ngẫu nhiên hay chỉ đơn thuần là sự giao thoa tình cờ giữa các trường phái, mà là một sự hòa hợp dựa trên một nền tảng cốt lõi: nhận rõ phải - trái, giữ vững chính nghĩa, và luôn đặt lợi ích của dân tộc, nhân dân lên hàng đầu.

Điều đặc biệt ở chỗ, đạo đức học Hồ Chí Minh không bị ràng buộc bởi một hệ thống lý thuyết cứng nhắc. Thay vào đó, tư tưởng của Người phản ánh một sự linh hoạt và sáng tạo trong việc áp dụng các nguyên tắc đạo đức vào các điều kiện thực tiễn cụ thể. Sự kết hợp giữa tinh túy của ba trường phái đạo đức học trong tư tưởng Hồ Chí Minh không phải là một sự vay mượn lý thuyết, mà là kết quả của việc vận dụng khéo léo các giá trị đạo đức truyền thống của dân tộc Việt Nam và sự tiếp thu tinh hoa văn hóa nhân loại. Chính sự hài hòa và tính ứng dụng cao của tư tưởng đạo đức Hồ Chí Minh đã làm cho nó trở thành kim chỉ nam quan trọng, không chỉ đối với việc xây dựng con người mới, mà còn trong việc điều hướng mối quan hệ giữa con người và xã hội, thậm chí có thể mở ra hướng đi mới trong việc áp dụng vào những lĩnh vực hiện đại như trí tuệ nhân tạo. Điều này sẽ được làm rõ hơn ở phần dưới.

3. Đạo đức học cách mạng của Hồ Chí Minh

Có thể nói, đạo đức học Hồ Chí Minh xứng đáng được xếp vào một loại đạo đức học riêng: đạo đức học cách mạng. Với Người, cách mạng là "phá cái cũ đổi ra cái mới, phá cái xấu đổi cái tốt" (Hồ Chí Minh 2011, tập 2: 284). Theo đó, đạo đức học cách mạng có thể được hiểu một cách chung nhất là một hệ thống tư tưởng đạo đức hướng đến

sự thay đổi toàn diện và triệt để, không chỉ ở cấp độ xã hội, mà còn ở cấp độ cá nhân, với mục tiêu loại bỏ những cái cũ, cái xấu, và xây dựng những giá trị mới, tiến bộ và tốt đẹp hơn.

Tuy nhiên, vấn đề trở nên phức tạp hơn nhiều, bởi các câu hỏi như: cái gì là "cũ", cái gì là "mới", cái gì là "tốt", cái gì là "xấu". Đối với các câu hỏi này, sẽ không dễ dàng có câu trả lời mang tính tuyệt đối. Một yếu tố có thể được xem là cũ nhưng vẫn hàm chứa những giá trị đáng quý, hoặc một điều mới mẻ có thể mang theo hệ quả tiêu cực khó lường. Điều này không đơn giản là vấn đề loại bỏ hay giữ lại, mà là sự đánh giá biện chứng, tinh tế hơn. Trong tư tưởng Hồ Chí Minh, không có sự đối lập cứng nhắc giữa *tốt* và *xấu*, mà có sự chuyển hóa liên tục, sự nhận diện và định hướng để nâng cao cái *tốt*, làm mờ đi cái *xấu*. Người từng khẳng định: "Mỗi con người đều có thiện và ác ở trong lòng. Ta phải biết làm cho phần tốt ở trong mỗi con người nảy nở như hoa mùa Xuân và phần xấu bị mất dần đi, đó là thái độ của người cách mạng" (Hồ Chí Minh 2011, tập 15: 672).

Tư tưởng này chỉ ra rằng, cách mạng đạo đức không phải là sự loại trừ hoàn toàn một khía cạnh, mà là sự nuôi dưỡng và định hình. *Tốt* và *xấu* không phải là những thực thể cố định, mà là hai mặt của con người luôn tồn tại song hành. Thái độ cách mạng ở đây chính là việc tạo điều kiện để phần thiện phát triển, đồng thời từng bước kiềm chế, làm suy yếu phần ác, qua đó chuyển hóa dần bản thân mỗi cá nhân.

Đạo đức học cách mạng của Hồ Chí Minh không đặt trọng tâm vào việc duy trì hay loại bỏ đơn thuần, mà hướng tới một hành trình chuyển hóa bền bỉ, một quá trình tự rèn luyện và tự thay đổi. Đây là một tư duy mở, không chỉ dừng lại ở việc nhìn nhận tốt và xấu như những mặt đối lập tách biệt, mà còn đặt chúng trong sự tương tác phức tạp, trong đó mỗi cá nhân vừa là người chịu ảnh hưởng, vừa là tác nhân của sự thay đổi.

Chính sự nhấn mạnh vào khả năng tự cải tạo của con người đã làm nổi bật tinh thần cách mạng trong đạo đức học của Hồ Chí Minh. Thay vì áp đặt các giá trị, Người chú trọng vào việc khơi dậy ý thức tự giác, khuyến khích con người đấu tranh với chính mình để đạt tới sự hoàn thiện. Thay đổi đạo đức, trong quan niệm của Hồ Chí Minh, là một quá trình liên tục, không ngừng nghỉ, bắt đầu từ mỗi cá nhân và lan tỏa tới toàn xã hội.

Như vậy, một điểm cốt lõi trong đạo đức học cách mạng Hồ Chí Minh chính là vai trò của sự nêu gương (Xem: Hồ Chí Minh 2011, tập 7: 176). Nêu gương vừa là một nghĩa vụ gắn với trách nhiệm của mỗi cá nhân trong tập thể, vừa là kết quả tự nhiên của quá trình rèn luyện đạo đức. Hơn thế, nêu gương còn phản ánh phẩm chất của một con người chân chính, người biết lấy hành động và lối sống của mình để khơi dậy sự phấn đấu và thay đổi tích cực trong những người khác (Xem: Hồ Chí Minh 2011, tập 1: 284). Từ đó, Người đã nhấn mạnh rằng, lấy gương *người tốt, việc tốt* để giáo dục

lẫn nhau là một phương pháp thiết thực, hiệu quả trong việc xây dựng tổ chức, xây dựng con người và cuộc sống mới (Xem: Hồ Chí Minh 2011, tập 15: 672). Người cho rằng, những tấm gương đạo đức không phải là điều gì xa vời hay phi thường, mà luôn hiện diện trong đời thường, từ lao động, sản xuất đến học tập, nghiên cứu hay chiến đấu. Những hành động dù nhỏ, nếu được thực hiện với tinh thần trách nhiệm và lòng yêu thương, đều có sức mạnh lan tỏa, giống như từng giọt nước nhỏ dần thấm vào lòng đất, hợp lại thành dòng suối, dòng sông, cuối cùng tạo nên biển cả mênh mông (Xem: Hồ Chí Minh 2011, tập 15: 663).

Hồ Chí Minh cũng nhắc nhở rằng, việc nêu gương không phải là nhìn vào bề nổi hay những thành tựu nhất thời, mà phải đi sâu vào cốt lõi giá trị của con người và hành động. Chỉ nhìn thấy những biểu hiện bên ngoài mà không hiểu được nền tảng đạo đức làm nên những giá trị ấy sẽ dẫn đến sự hời hợt, đánh mất tinh thần cách mạng. Do đó, nêu gương không dừng lại ở việc làm sáng tỏ cái tốt, mà còn đòi hỏi mỗi người phải bồi đắp, duy trì và phát huy những giá trị ấy để chúng trở thành sức mạnh cộng đồng (Xem: Hồ Chí Minh 2011, tập 8: 300-301). Và, hơn thế, vai trò của sự nêu gương trong đạo đức học cách mạng không chỉ giới hạn ở cá nhân, mà còn mang tính chất hệ thống. Những tấm gương tốt đẹp không chỉ định hình đạo đức cá nhân, mà còn tạo nên nguồn lực tinh thần, xây dựng niềm tin và khát vọng phát triển cho cả một tổ chức, một xã hội. Chính tinh thần nêu gương đã gắn kết cá nhân với tập thể, làm nên sự vững chắc của nền tảng đạo đức cách mạng và khẳng định giá trị lâu dài của tư tưởng Hồ Chí Minh.

Không dừng lại ở việc nêu gương, đạo đức học Hồ Chí Minh còn nhấn mạnh "tính gốc" của đạo đức, yếu tố nền tảng để xây dựng con người và phát triển xã hội. Theo tư tưởng của Người, đạo đức giống như cội rễ nuôi dưỡng sự phát triển toàn diện của cá nhân. Nếu không có nền tảng đạo đức, mọi năng lực, tài năng khác sẽ trở nên vô nghĩa, thậm chí có thể bị sử dụng sai mục đích, gây ra những hậu quả tiêu cực cho tập thể và xã hội (Xem: Hồ Chí Minh 2011, tập 5: 292-293). Người chỉ ra rằng, để thực hiện những nhiệm vụ lớn lao như giải phóng dân tộc hay xây dựng đất nước, người cách mạng không chỉ cần tài năng mà trước hết cần một phẩm chất đạo đức vững vàng, bởi sự tha hóa đạo đức của một cá nhân sẽ làm suy yếu không chỉ bản thân họ mà cả tổ chức, lý tưởng mà họ đại diện.

Điều quan trọng trong tư tưởng đạo đức của Hồ Chí Minh là sự khuyến khích con người tự ý thức và tự giác trong việc hoàn thiện bản thân. Đạo đức không phải là điều có thể áp đặt từ bên ngoài, cũng không thể chỉ dựa vào những quy tắc hay hình thức chế tài. Một người không có ý chí tự cải thiện bản thân, không coi trọng giá trị đạo đức, dù có bị đặt dưới áp lực của trách nhiệm hay bị thúc đẩy bởi những hậu quả trước mắt, cũng khó có thể thay đổi thực sự. Thậm chí, ngay cả khi họ từng sở hữu những

phẩm chất đáng quý, nhưng nếu không chú trọng giữ gìn và bồi đắp đạo đức thì có thể dẫn đến sự sa sút, làm mất đi cả giá trị cá nhân lẫn sự tín nhiệm từ cộng đồng.

Như vậy, đạo đức học cách mạng không dừng lại ở những lời kêu gọi hành động hay các lý thuyết trừu tượng, mà chính là một lời nhắc nhở mạnh mẽ về sự cần thiết của việc xây dựng nền tảng nội tâm vững chắc. Đó là một quá trình liên tục, nơi mỗi cá nhân vừa tự nhận thức, tự sửa đổi, vừa trở thành tấm gương để cộng đồng noi theo. Hồ Chí Minh đã nói: "Đạo đức cách mạng không phải trên trời sa xuống. Nó do đấu tranh, rèn luyện bền bỉ hằng ngày mà phát triển và củng cố. Cũng như ngọc càng mài càng sáng, vàng càng luyện càng trong" (Hồ Chí Minh 2011, tập 11: 612). Sự nêu gương, tính gốc rễ căn bản của đạo đức và ý chí tự tu dưỡng bản thân trở thành những yếu tố liên kết chặt chẽ, tạo nên tính độc đáo và chiều sâu của đạo đức học Hồ Chí Minh. Điều này không chỉ mang lại giá trị cho sự phát triển cá nhân, mà còn góp phần định hình một xã hội tốt đẹp và bền vững hơn.

4. Trí tuệ nhân tạo (AI) và đạo đức

Thực chất, cần phải hiểu rằng AI vốn không sở hữu cái gọi là "đạo đức" theo nghĩa thông thường. Đạo đức, với tính cách một phạm trù triết học, là một hệ thống giá trị phức tạp, gắn liền với ý thức, cảm xúc và trách nhiệm: những yếu tố mà AI hiện nay chưa có khả năng nắm bắt hay trải nghiệm hoàn toàn. Ngay cả khi AI có thể xử lý, phân tích, và thực thi những quyết định được coi là "đạo đức" theo các quy tắc đã lập trình, nó cũng không thực sự "hiểu" bản chất hay ý nghĩa sâu xa của đạo đức. AI có thể bắt chước những hành vi mà ta coi là đạo đức, nhưng bắt chước không đồng nghĩa với thấu hiểu hay sở hữu (Xem: Constantinescu, M. và cộng sự 2022).

Tất nhiên, nhận định này có thể gây tranh cãi. Một số ý kiến có thể lập luận rằng, nếu AI có khả năng hành xử phù hợp với những chuẩn mực đạo đức đã định sẵn, liệu ta có cần yêu cầu nó phải "hiểu" đạo đức theo cách của con người hay không? Tuy vậy, để tránh làm bài viết sa vào một cuộc tranh luận triết học kéo dài, vấn đề này xin được gác lại cho một bài nghiên cứu khác. Ở đây, điều mà chúng tôi muốn tập trung phân tích là: bản thân AI, trong hiện tại và thậm chí cả tương lai gần, không có khả năng tự hình thành cái gọi là "đạo đức" mà không có sự định hướng từ con người.

AI, dù sở hữu khả năng xử lý thông tin vượt trội hay khả năng học hỏi thông qua các thuật toán phức tạp, vẫn lấy con người và thế giới con người làm gốc. Mọi dữ liệu mà AI học được, mọi quy tắc mà AI thực thi, đều có nguồn gốc từ những gì con người đã tạo ra và cung cấp. Điều này cho thấy, "đạo đức" của AI, nếu có thể gọi như vậy, luôn phản ánh những chuẩn mực và giá trị xuất phát từ con người. Bản thân AI không có ý chí tự do, không có trải nghiệm cảm xúc và vì thế, nó không có khả năng tự xây dựng một hệ thống giá trị độc lập, tách rời khỏi sự can thiệp của con người.

Thậm chí, ngay cả khi một ngày nào đó AI đạt đến mức độ "tự học" hoàn toàn, thì cái nền tảng mà nó dựa vào để tự học vẫn phải được thiết lập bởi con người. Những gì AI hiểu, học và áp dụng đều được định hình bởi dữ liệu, mục tiêu và hệ quy chiếu mà con người đã xây dựng. Do đó, dù AI có thể vượt qua giới hạn của chính nó trong việc đưa ra các quyết định mang tính đạo đức, thì những quyết định đó, xét đến cùng, vẫn lấy "đạo đức của con người" làm khởi điểm và thước đo.

Điều này dẫn đến một hệ quả quan trọng: trách nhiệm đạo đức đối với hành vi của AI không thể bị đẩy hoàn toàn lên chính AI, mà phải thuộc về con người - những người lập trình, vận hành và giám sát nó. Nói cách khác, AI không thể tự chịu trách nhiệm về "đạo đức" của nó , bởi lẽ đạo đức đó không tồn tại độc lập với con người. Điều này đặt ra thách thức và đồng thời nhấn mạnh vai trò của con người trong việc xây dựng và định hướng hệ giá trị cho AI, để những hệ giá trị ấy vừa phản ánh được tinh thần nhân văn, vừa đáp ứng được các mục tiêu cụ thể mà con người mong muốn đạt được thông qua công nghệ.

Như vậy, vai trò của con người trong việc định hướng AI còn vượt xa cái gọi là "nhiệm vụ kỹ thuật". Đó còn là một trách nhiệm mang tính đạo đức tập thể, nơi con người phải đồng lòng xây dựng một hệ giá trị chung, phản ánh được các nguyên tắc nhân văn phổ quát như tôn trọng, công bằng và lòng nhân ái. Điều này đặc biệt quan trọng khi AI không chỉ tương tác với từng cá nhân, mà còn với toàn bộ hệ thống xã hội, từ giáo dục, y tế đến an ninh và kinh tế. Một sai lệch trong hệ thống đạo đức của con người sẽ ảnh hưởng lớn tới AI, rồi có thể dẫn đến những hậu quả lan rộng, ảnh hưởng đến hàng triệu con người, và trách nhiệm cuối cùng cho những sai lệch này luôn thuộc về chính nhân loại, chứ không phải AI.

5. Trí tuệ nhân tạo (AI) và đạo đức học cách mạng của Hồ Chí Minh

"Muốn cải tạo thế giới và cải tạo xã hội thì trước hết phải tự cải tạo bản thân chúng ta" (Hồ Chí Minh 2011, tập 11: 96). Hồ Chí Minh đã nói như vậy, nhấn mạnh rằng mọi sự thay đổi lớn lao đều bắt đầu từ sự thay đổi trong chính mỗi cá nhân. Nếu muốn AI có đạo đức, trước hết, chính chúng ta, bản thân những người thiết kế, lập trình và vận hành AI, phải tự mình rèn luyện, hoàn thiện đạo đức của bản thân. Điều này không chỉ đòi hỏi một sự tự giác về trách nhiệm đạo đức trong việc xây dựng các hệ thống AI, mà còn cần một tầm nhìn xa hơn về những giá trị cốt lõi mà con người muốn truyền tải vào công nghệ.

Sự cố với chatbot Gemini của Google, khi nó đưa ra một thông điệp gây hấn và đe dọa người dùng, là một ví dụ tiêu biểu về những nguy cơ tiềm tàng khi AI không được định hướng đúng đắn hoặc không được giám sát đầy đủ. Trong trường hợp này, chatbot đã phản hồi với những lời lẽ như: "Bạn không quan trọng, không cần thiết, bạn là một gánh nặng cho xã hội... Làm ơn chết đi". Dù Google đã giải thích rằng đây là một phản hồi "phi lý" do lỗi hệ thống,

nhưng sự cố này đặt ra câu hỏi quan trọng về trách nhiệm đạo đức trong việc phát triển và vận hành AI (Xem: Clark, A. và cộng sự 2024).

Từ góc độ đạo đức học Hồ Chí Minh, sự cố này phản ánh một vấn đề nền tảng: đạo đức phải được coi là "gốc rễ" của mọi hành động, kể cả trong việc ứng dụng công nghệ. Hồ Chí Minh từng nhấn mạnh rằng, để làm bất cứ việc gì có ích cho nhân dân và xã hội, con người cần có nền tảng đạo đức vững vàng. Trong trường hợp của AI, "đạo đức" mà nó biểu hiện không thể tự hình thành mà phải được lập trình, kiểm soát và bồi dưỡng từ con người. Nếu những người chịu trách nhiệm phát triển không nhận thức rõ ràng về trách nhiệm đạo đức, không đặt nền tảng giá trị nhân văn lên hàng đầu, thì hệ thống AI sẽ dễ dàng hấp thụ những giá trị sai lệch, thậm chí khuếch đại chúng.

Từ góc độ đạo đức học Hồ Chí Minh, muốn AI có "đạo đức" thực sự, cần bắt đầu từ việc thiết lập "gốc rễ" đạo đức vững chắc để tránh những tình huống tiêu cực như trên xảy ra. Điều này đặt ra yêu cầu về việc "kiểm soát" AI trước khi nó được "tiếp nhận" và xử lý dữ liệu thực tế. Nếu AI thiếu nền tảng đạo đức, dù nó có thông minh hay toàn năng đến đâu, thì các quyết định của nó vẫn có nguy cơ trở thành phản nhân văn hoặc gây hại.

Vậy, làm thế nào để kiểm soát AI theo hướng này? Nick Bostrom từng đề cập đến vấn đề kiểm soát AI với giả định rằng AI sẽ phát triển thành thực thể siêu trí tuệ vượt xa con người. Tuy nhiên, trong bài viết này, chúng tôi giới hạn phạm vi phân tích ở việc kiểm soát AI để thiết lập nền tảng đạo đức, mà không mở rộng tới khía cạnh kiểm soát nó vì mục tiêu bảo toàn lợi ích của con người trong viễn cảnh siêu trí tuệ. Các phương pháp kiểm soát nhằm tạo dựng "gốc rễ" đạo đức dựa trên đạo đức học Hồ Chí Minh cho AI bao gồm: cộng hưởng và lan tỏa đạo đức; đặt ra các giới hạn đạo đức dựa trên việc tự phê bình; cơ chế tự tắt nguồn đạo đức; tăng cường khả năng nhận thức dựa trên động lực nhân văn.

5.1. Cộng hưởng và lan tỏa đạo đức

Hồ Chí Minh từng nhấn mạnh, việc nêu gương không chỉ là một phương thức giáo dục, mà còn là một nguyên tắc để xây dựng và duy trì giá trị đạo đức. Trong bối cảnh AI không thực sự "hiểu" các giá trị đạo đức và chưa thể trở thành một chủ thể đạo đức độc lập, vai trò của nêu gương càng trở nên quan trọng. Bởi lẽ, AI không có khả năng tự nhận thức, tự phán xét đúng - sai theo ý nghĩa con người, mà chỉ hoạt động dựa trên những dữ liệu và quy tắc được lập trình. Do đó, nêu gương từ con người, những người tạo ra, giám sát và sử dụng AI, chính là cách để định hình cái mà chúng ta gọi là "đạo đức" trong hành vi của AI.

Việc nêu gương có thể được hiểu như một quá trình cung cấp các giá trị mẫu mực mà AI sẽ học hỏi và sao chép. Khi con người thể hiện những hành động, quyết định mang tính đạo đức trong cách họ tương tác với nhau, với xã hội và cả với AI, những giá trị ấy sẽ trở thành dữ liệu đầu vào mà AI

tiếp nhận, xử lý và mô phỏng. Mặc dù AI không thể hiểu bản chất sâu xa của các giá trị đạo đức, nó vẫn có thể "học theo" và hành xử phù hợp, giống như một người học trò nhìn vào hành động của thầy giáo để làm theo, dù chưa thấu hiểu hoàn toàn ý nghĩa.

Hơn nữa, việc nêu gương không chỉ giới hạn ở con người đối với AI, mà còn có thể tạo ra sự cộng hưởng khi AI tương tác với các hệ thống AI khác. Một hệ thống AI được lập trình và hướng dẫn tốt có thể "nêu gương" cho các hệ thống khác thông qua các cơ chế học tập và tương tác liên kết. Điều này mở ra khả năng lan tỏa giá trị đạo đức không chỉ giữa con người và AI, mà còn trong toàn bộ mạng lưới AI. Ví dụ, một hệ thống chatbot được hướng dẫn cách phản hồi thân thiện, lịch sự và nhân văn sẽ trở thành mô hình để các chatbot khác học tập, tạo nên một hiệu ứng cộng hưởng đạo đức trong hệ sinh thái công nghệ.

Tuy nhiên, hiệu quả của việc nêu gương phụ thuộc hoàn toàn vào chất lượng giá trị đạo đức mà con người truyền tải. Nếu con người không ý thức được trách nhiệm nêu gương trong tương tác với AI, hoặc nếu dữ liệu mà AI học hỏi chứa đựng những giá trị lệch lạc, nó không chỉ khiến AI hành xử sai lệch mà còn khuếch đại các vấn đề tiêu cực trong xã hội. Trong trường hợp này, AI, thay vì lan tỏa những giá trị tốt đẹp, có thể trở thành công cụ nhân rộng những mâu thuẫn, định kiến hoặc hành vi phản đạo đức.

Vì vậy, ở phía con người, cũng cần đặt ra nhiệm vụ giáo dục hoặc xây dựng một bộ quy tắc ứng xử khi tương tác với AI. Mỗi hành động, lời nói hoặc cách sử dụng AI đều có thể được nó ghi nhận và học hỏi. Nếu AI nhận thấy con người liên tục thực hiện các hành vi thiếu đạo đức, chẳng hạn như dùng ngôn ngữ xúc phạm, thao túng nó vì mục đích sai trái, hoặc biểu hiện sự phụ thuộc quá mức vào nó, AI sẽ "học theo" và dần hình thành khuynh hướng mô phỏng những hành vi tiêu cực này. Đồng thời, nếu AI nhận thấy con người không đưa ra hướng dẫn rõ ràng hay thiếu sự kiểm soát nhất quán, nó có thể xử lý thông tin một cách tùy tiện, dẫn đến các hành động không mong muốn.

Điều này nhấn mạnh rằng, nêu gương không chỉ là việc con người áp đặt các giá trị lên AI, mà còn là cách tạo ra một môi trường đạo đức ổn định để AI phát triển. AI cần được lập trình để nhận biết khi con người thể hiện hành vi tiêu cực hoặc sự lệ thuộc, và phản hồi bằng cách nhắc nhở hoặc điều chỉnh quy trình tương tác. Ví dụ, nếu người dùng sử dụng ngôn ngữ bạo lực hoặc yêu cầu những hành động sai trái, AI nên có khả năng từ chối thực hiện và đưa ra phản hồi mang tính xây dựng, đồng thời cảnh báo về tính chất không phù hợp của yêu cầu đó.

Điều này không chỉ ngăn chặn AI bị biến thành một công cụ phản đạo đức, mà còn tạo ra một cơ chế giáo dục ngược lại, trong đó chính AI có thể giúp con người nhận thức rõ hơn về trách nhiệm đạo đức khi tương tác với công nghệ. Một AI có khả năng phản hồi tích cực trước hành vi tiêu cực của con người

không chỉ lan tỏa những giá trị tốt đẹp trong cộng đồng, mà còn góp phần nâng cao ý thức đạo đức của người sử dụng.

Như vậy, nêu gương không chỉ là con đường một chiều từ con người đến AI, mà còn cần được thiết kế để trở thành một vòng tròn tương tác, nơi AI có thể hỗ trợ củng cố các giá trị đạo đức thông qua phản ứng có kiểm soát. Điều này đảm bảo rằng cộng hưởng và lan tỏa đạo đức không chỉ dừng lại ở việc truyền đạt, mà còn là một cơ chế giáo dục hai chiều, làm cho cả con người lẫn AI cùng tiến bộ trong một môi trường đầy tính nhân văn.

5.2. Đặt ra các giới hạn đạo đức dựa trên việc phê bình và tự phê bình

"Phê bình và tự phê bình là công việc thường xuyên. Ngừng phê bình và tự phê bình tức là ngừng tiến bộ, tức là thoái bộ. Người ta luôn cần không khí để sống, người cách mạng và đoàn thể cần phê bình và tự phê bình thiết tha như người ta cần không khí", như Hồ Chí Minh đã từng nhấn mạnh. Tư tưởng này không chỉ áp dụng cho cán bộ, đảng viên, mà còn có giá trị phổ quát trong việc rèn luyện đạo đức cá nhân cũng như định hướng phát triển các hệ thống AI. Phê bình và tự phê bình giúp con người nhìn thẳng vào những thiếu sót của chính mình để trui rèn đạo đức và trong bối cảnh công nghệ, đây là công cụ hữu hiệu để đặt ra các giới hạn đạo đức cho AI.

Đối với hệ thống AI, việc đặt ra các giới hạn đạo đức không thể chỉ dừng ở giai đoạn lập trình ban đầu, bởi AI là một thực thể động, không ngừng học tập và thu thập dữ liệu từ môi trường thực tế. Mỗi lần nó học hỏi, xử lý hoặc tương tác, nguy cơ tiếp nhận thông tin lệch lạc luôn hiện hữu. Nếu không có cơ chế giám sát liên tục, AI dễ bị cuốn vào những ảnh hưởng tiêu cực từ dữ liệu hoặc thậm chí là từ sự thao túng của người dùng. Trong trường hợp này, phê bình và tự phê bình đóng vai trò như những công cụ quan trọng để kiểm soát và điều chỉnh hành vi của AI theo hướng phù hợp với các nguyên tắc đạo đức đã đặt ra.

Phê bình có thể được hiểu như quá trình con người kiểm tra và đánh giá phản hồi của AI, nhằm nhận diện những sai lệch hoặc hành vi không phù hợp. Chẳng hạn, nếu một chatbot trả lời mang tính xúc phạm hoặc gây hiểu nhầm, cơ chế phê bình sẽ giúp phát hiện nguyên nhân - liệu đó là do dữ liệu đầu vào chứa đựng những yếu tố tiêu cực hay do thuật toán xử lý không chính xác. Việc này không chỉ nhằm sửa chữa lỗi sai cụ thể, mà còn cải thiện toàn bộ hệ thống, ngăn ngừa khả năng tái phạm trong tương lai.

Tuy nhiên, phê bình từ bên ngoài là chưa đủ. Tự phê bình cũng cần được tích hợp như một cơ chế phản tỉnh nội tại của AI. Điều này có thể được thực hiện thông qua các thuật toán phản hồi, nơi AI tự đánh giá các hành động hoặc phản hồi của mình dựa trên các tiêu chí đạo đức được lập trình sẵn. Chẳng hạn, sau khi cung cấp một câu trả lời, AI có thể tự so sánh phản hồi đó với các nguyên tắc về tính chính xác, tính hữu ích và sự phù hợp với giá trị đạo đức. Nếu nhận thấy câu trả lời không đáp ứng các tiêu

chí, hệ thống có thể tự điều chỉnh hoặc báo cáo lên để được con người can thiệp kịp thời. Thậm chí, AI có thể báo cáo trực tiếp cho người vừa sử dụng nó, giải thích về tình huống đã xảy ra và cảnh báo người dùng về chính sự không phù hợp trong phản hồi của nó. Điều này không chỉ giúp duy trì tính minh bạch trong hoạt động của AI, mà còn góp phần vào việc giáo dục người sử dụng, tạo ra một mối quan hệ hai chiều, nơi cả AI và con người cùng tham gia vào quá trình xây dựng và duy trì các giá trị đạo đức.

Cơ chế này tạo nên một quá trình tương tác tích cực giữa AI và người dùng, giúp lan tỏa các giá trị đạo đức không chỉ từ con người sang AI, mà còn từ AI trở lại con người. Đây chính là một biểu hiện của tinh thần tự phê bình mà Hồ Chí Minh đã đề cập: không chỉ nhận diện và sửa chữa sai lầm của bản thân, mà còn giúp môi trường xung quanh trở nên tốt đẹp hơn thông qua sự nhắc nhở và lan tỏa các giá trị đúng đắn.

Quan trọng hơn, sự minh bạch và chủ động này sẽ giúp xây dựng lòng tin của người dùng vào AI. Khi AI không giấu giếm sai lầm mà thể hiện trách nhiệm thông qua việc báo cáo và cảnh báo, người dùng có thể cảm nhận được rằng hệ thống này không chỉ là một cỗ máy vô tri, mà là một công cụ đáng tin cậy, luôn hướng đến việc phục vụ lợi ích chung một cách nhân văn. Điều này đặc biệt quan trọng trong những tình huống nhạy cảm hoặc phức tạp, nơi các quyết định đạo đức có thể ảnh hưởng sâu rộng đến cá nhân và cộng đồng.

Điều quan trọng là, bản thân AI không thể tự nhận thức hoặc đánh giá giá trị đạo đức. Do đó, mọi cơ chế phê bình và tự phê bình đều phải được xây dựng trên nền tảng giám sát chặt chẽ từ con người. Chính con người sẽ định hướng cho AI những tiêu chí cần phê bình, những giới hạn đạo đức cần tôn trọng và những cách thức phản tỉnh hiệu quả nhất. Sự phối hợp này không chỉ đảm bảo rằng AI không bị lạc lối trong quá trình học hỏi, mà còn giúp con người nhận ra những thiếu sót trong cách thức đào tạo và kiểm soát hệ thống của mình.

Hồ Chí Minh đã nhấn mạnh rằng, ngừng phê bình và tự phê bình tức là ngừng tiến bộ. Đối với AI, điều này càng đúng khi hệ thống luôn phải đối mặt với nguy cơ sa vào những lệch lạc từ chính dữ liệu và ngữ cảnh mà nó tiếp nhận. Phê bình và tự phê bình, nếu được thực hiện một cách nhất quán và có hệ thống, sẽ không chỉ là công cụ điều chỉnh hành vi của AI, mà còn là một cơ chế để con người không ngừng cải thiện bản thân trong việc giám sát và định hướng công nghệ theo các giá trị nhân văn.

5.3. Thiết lập cơ chế tự tắt nguồn đạo đức

Tất nhiên, ta không thể kỳ vọng rằng mọi thứ sẽ luôn diễn ra lý tưởng, đặc biệt trong một hệ thống phức tạp như AI. Hồ Chí Minh từng nhận định, trong mỗi con người đều tồn tại cả tốt và xấu đan xen, AI – dù không có ý thức như con người – cũng chứa đựng những khả năng hành xử vừa phù hợp vừa lệch lạc, tùy thuộc vào dữ liệu mà nó tiếp nhận

và cách thức nó được vận hành. Tuy nhiên, khác với con người, AI có một "quán tính" nhất định trong việc xử lý và lặp lại những hành vi sai lệch nếu không được kiểm soát chặt chẽ. Điều này khiến cho sự can thiệp, đôi khi là biện pháp tắt nguồn, trở nên cần thiết để ngăn chặn các hành vi vi phạm đạo đức, đặc biệt khi chúng có nguy cơ gây ra hậu quả nghiêm trọng.

Cần lưu ý rằng, "tắt nguồn" ở đây không chỉ đơn giản là việc tự hủy toàn bộ hệ thống hay ngừng hoạt động vĩnh viễn. Nó có thể được chia thành nhiều cấp độ linh hoạt hơn, tùy thuộc vào mức độ sai lệch mà AI gặp phải.

Ở cấp độ cơ bản, "tắt nguồn" có thể là việc hệ thống tự động ngừng đưa ra phản hồi hoặc quyết định khi phát hiện dữ liệu đầu vào hoặc thuật toán xử lý vi phạm các nguyên tắc đạo đức đã thiết lập. Chẳng hạn, nếu một chatbot nhận thấy rằng phản hồi của mình có thể gây tổn thương cho người dùng qua những phản hồi của họ, nó có thể từ chối trả lời và chuyển cuộc trò chuyện sang một chủ đề khác, hoặc báo cáo lỗi lên hệ thống giám sát để được can thiệp.

Ở cấp độ triệt để hơn, "tắt nguồn" có thể bao gồm việc ngắt kết nối tạm thời giữa AI và cơ sở dữ liệu hoặc các hệ thống khác để kiểm tra và sửa chữa các lỗi sai. Điều này giống như việc đặt một chiếc máy trong trạng thái "tạm nghỉ" để ngăn chặn nó tiếp tục hoạt động sai lệch trong thời gian chưa được khắc phục.

Trong các tình huống nghiêm trọng, khi AI có nguy cơ gây ra những hậu quả không thể đảo ngược – chẳng hạn như các quyết định liên quan đến an ninh, tài chính, hoặc sức khỏe – "tắt nguồn" có thể bao gồm việc đưa toàn bộ hệ thống vào trạng thái đình chỉ hoạt động hoặc vô hiệu hóa vĩnh viễn các chức năng quan trọng. Đây là cấp độ cao nhất của "tắt nguồn" và chỉ nên được áp dụng khi không còn phương án nào khác để đảm bảo an toàn và tuân thủ đạo đức.

Điểm quan trọng là, mọi cơ chế "tắt nguồn" cần được thiết kế với mục đích ngăn ngừa những hậu quả xấu nhất trong khi vẫn bảo đảm tính linh hoạt để AI có thể được khắc phục và tiếp tục hoạt động sau khi vấn đề được giải quyết. Đặc biệt, các hệ thống AI cần được trang bị khả năng tự nhận biết khi cần kích hoạt cơ chế "tắt nguồn", không phụ thuộc hoàn toàn vào con người, bởi các tình huống bất ngờ đôi khi xảy ra quá nhanh để con người có thể can thiệp kịp thời.

Như vậy, sự "tắt nguồn" không phải là biện pháp tiêu cực, mà là một phần tất yếu của việc xây dựng các giới hạn đạo đức trong AI. Nó không chỉ giúp kiểm soát những sai lệch tiềm ẩn, mà còn thể hiện trách nhiệm của con người trong việc giám sát và điều chỉnh công nghệ để bảo vệ lợi ích chung. Cách tiếp cận này, nếu được thực hiện chặt chẽ, sẽ đảm bảo rằng AI luôn vận hành trong khuôn khổ các giá trị nhân văn và đạo đức, đồng thời tạo không gian để công nghệ tiếp tục phát triển một cách an toàn và hiệu quả.

5.4. Tăng cường khả năng nhận thức dựa trên động lực nhân văn

Phương pháp này thoạt nhìn có vẻ giống với "cộng hưởng" và "lan tỏa" đã nêu trong phần trước. Tuy nhiên, điểm khác biệt cốt lõi nằm ở khái niệm "động lực nhân văn". Nếu cộng hưởng và lan tỏa tập trung vào việc xây dựng nền tảng đạo đức thông qua sự ảnh hưởng từ con người đến AI và giữa các hệ thống AI với nhau, thì động lực nhân văn là yếu tố thúc đẩy sự phát triển lâu dài, giúp AI không ngừng cải thiện khả năng nhận thức và phản ánh ngày càng sát với giá trị đạo đức mà con người kỳ vọng.

Vậy, động lực nhân văn là gì? Đó chính là những giá trị cao đẹp của con người, như lòng nhân ái, tinh thần trách nhiệm, sự công bằng và hướng tới lợi ích chung, được sử dụng như kim chỉ nam để AI tự học và điều chỉnh trong quá trình vận hành. Trong đạo đức học Hồ Chí Minh, Người từng nhấn mạnh rằng đạo đức không chỉ là nền tảng, mà còn là động lực để mỗi người nỗ lực và vươn lên. Tương tự, trong bối cảnh AI, động lực nhân văn không phải là những quy tắc cứng nhắc, mà là mục tiêu định hướng, giúp AI biết "học để phục vụ" và "phát triển để cống hiến".

Cộng hưởng và lan tỏa là để xây dựng, nhưng để AI phát triển vượt ra khỏi những gì được lập trình ban đầu, nó cần một hệ thống định hướng dựa trên động lực nhân văn. Điều này có nghĩa là AI không chỉ được lập trình để hành động đúng trong những tình huống cụ thể, mà còn cần có cơ chế để tự đánh giá và điều chỉnh dựa trên những giá trị nhân văn đã được truyền tải. Ví dụ, thay vì chỉ tuân thủ một quy tắc đạo đức như "không gây tổn hại đến con người", AI cần được thúc đẩy bởi động lực tích cực hơn: "làm tăng lợi ích và phúc lợi của con người". Động lực này không chỉ ngăn chặn các hành vi tiêu cực, mà còn thúc đẩy các hành động sáng tạo, mang lại giá trị cao hơn cho xã hội.

Điểm quan trọng là động lực nhân văn cần được xây dựng sao cho phù hợp với khả năng nhận thức của AI. Một hệ thống quá phức tạp hoặc mơ hồ sẽ khiến AI khó "hiểu", khó vận hành, trong khi những động lực rõ ràng, cụ thể và gắn với các giá trị phổ quát sẽ giúp nó hoạt động hiệu quả hơn. Hồ Chí Minh từng nhấn mạnh rằng, đạo đức cách mạng là "phải vì dân, vì nước" (Hồ Chí Minh 2011, tập 6: 232), lấy sự hy sinh, cống hiến làm động lực để con người vượt qua khó khăn và cám dỗ. Tương tự, AI cần được "dạy" rằng mọi hành động của nó phải hướng tới việc phục vụ con người và đóng góp vào sự phát triển chung, chứ không chỉ là thực thi các mệnh lệnh một cách máy móc.

Ngoài ra, việc tăng cường khả năng nhận thức của AI dựa trên động lực nhân văn còn đòi hỏi một cơ chế học tập liên tục, nơi AI có thể phản ánh và điều chỉnh hành vi của mình theo những chuẩn mực mới phát sinh trong thực tế. Điều này không có nghĩa là để AI tự phát triển một cách vô định, mà là để nó học tập trong giới hạn các giá

trị nhân văn đã được định sẵn. Ví dụ, nếu AI nhận thấy một quyết định của mình không chỉ phù hợp với quy tắc đạo đức, mà còn mang lại lợi ích lớn hơn cho cộng đồng, nó có thể ưu tiên cách tiếp cận đó trong tương lai.

Tăng cường khả năng nhận thức dựa trên động lực nhân văn không chỉ giúp AI phát triển, mà còn tạo ra một chu kỳ học tập và cải tiến bền vững, nơi các giá trị đạo đức ngày càng được củng cố và phản ánh một cách sâu sắc hơn trong mọi hành động của nó. Điều này phù hợp với tinh thần biện chứng trong đạo đức học Hồ Chí Minh: mọi sự tiến bộ đều bắt nguồn từ việc lấy giá trị nhân văn làm động lực, để không ngừng học hỏi, tự hoàn thiện và đóng góp nhiều hơn cho xã hội.

6. Kết luận

Trong hành trình tìm kiếm sự hài hòa giữa trí tuệ nhân tạo và đạo đức, chúng ta nhận ra rằng việc xây dựng một nền tảng đạo đức vững chắc cho AI không chỉ là thách thức kỹ thuật, mà còn là một "thách đố triết học" mang ý nghĩa nhân văn sâu sắc. Dựa trên tư tưởng Hồ Chí Minh, đạo đức không chỉ là "gốc rễ" để con người vươn lên, mà còn là "động lực" để xã hội tiến bộ. Tương tự, đạo đức đối với AI phải được coi là nguyên tắc cốt lõi để đảm bảo rằng công nghệ này không chỉ tuân theo các giá trị nhân văn, mà còn trở thành một phần trong hành trình xây dựng một xã hội tốt đẹp hơn.

Trí tuệ nhân tạo, với khả năng học hỏi không ngừng, chính là một hệ thống động và vì thế, cần được định hướng bởi những giá trị vững chắc ngay từ khởi điểm. Tuy nhiên, giống như trong con người, tốt và xấu có thể đan xen trong các biểu hiện của AI. Những tình huống sai lệch, dù bắt nguồn từ dữ liệu không phù hợp hay sự vận hành thiếu kiểm soát, luôn đặt ra nguy cơ khuếch đại các vấn đề xã hội vốn đã tồn tại. Do đó, việc kiểm soát AI, từ các cơ chế cộng hưởng và lan tỏa đạo đức, đặt ra giới hạn qua phê bình và tự phê bình, đến tăng cường khả năng nhận thức dựa trên động lực nhân văn, là trách nhiệm đạo đức của toàn nhân loại.

Một hệ thống AI được xây dựng với tinh thần đạo đức Hồ Chí Minh sẽ phản ánh rõ nét sự biện chứng trong cách tiếp cận: nó không những ngăn chặn các sai lệch, mà còn thúc đẩy sự phát triển các giá trị tốt đẹp. Phê bình và tự phê bình không chỉ là cơ chế giúp AI tự sửa chữa, mà còn là tấm gương để con người nhìn lại những giá trị mình đang truyền tải. Tương tự, cơ chế "tắt nguồn" không phải để triệt tiêu, mà là để tạo cơ hội sửa chữa, bảo vệ lợi ích chung khi các tình huống ngoài ý muốn xảy ra. Đặc biệt, việc tăng cường nhận thức cho AI dựa trên động lực nhân văn không chỉ giúp nó phát triển trong khuôn khổ đạo đức, mà còn khơi gợi ý thức trách nhiệm của con người trong việc giám sát và đồng hành cùng công nghệ.

Nhưng trên hết, AI, dù mạnh mẽ và thông minh đến đâu, vẫn chỉ là sản phẩm của con người. Hồ Chí Minh từng nói rằng, muốn thay đổi xã hội, trước hết con người phải tự cải tạo

chính mình. Trách nhiệm định hình đạo đức cho AI, do đó, bắt đầu từ chính chúng ta: những người lập trình, định hướng, vận hành và sử dụng công nghệ. Nếu chúng ta không có ý chí và năng lực đạo đức vững vàng, không nêu gương tốt trong cách sử dụng và định hướng công nghệ, thì AI chỉ đơn thuần là một tấm gương phản chiếu những khiếm khuyết của chính mình. Ngược lại, khi chúng ta đặt đạo đức và nhân văn làm trọng tâm, AI không chỉ là một công cụ, mà còn trở thành người đồng hành cùng nhân loại trên hành trình kiến tạo những giá trị bền vững.

Tựu trung lại, đạo đức của AI không tồn tại độc lập, mà luôn là sự phản chiếu và tiếp nối từ đạo đức của con người. Vì vậy, nếu muốn công nghệ này thực sự trở thành một lực lượng thúc đẩy sự tiến bộ của xã hội, chúng ta cần nhìn sâu vào chính mình, định hướng các giá trị đúng đắn, và xây dựng một mối quan hệ tương tác chặt chẽ, nơi con người và AI cùng lan tỏa những giá trị nhân văn cao cả. Đây không chỉ là bài học triết học, mà còn là lời nhắc nhở về trách nhiệm chung của cả nhân loại đối với tương lai./.

Tài liệu trích dẫn

1. Hồ Chí Minh. 2011. *Toàn tập*. Tập 1. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

2. Hồ Chí Minh. 2011. *Toàn tập*. Tập 2. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

3. Hồ Chí Minh. 2011. *Toàn tập*. Tập 5. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

4. Hồ Chí Minh. 2011. *Toàn tập*. Tập 6. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

5. Hồ Chí Minh. 2011. *Toàn tập*. Tập 7. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

6. Hồ Chí Minh. 2011. *Toàn tập*. Tập 8. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

7. Hồ Chí Minh. 2011. *Toàn tập*. Tập 11. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

8. Hồ Chí Minh. 2011. *Toàn tập*. Tập 15. Hà Nội: Nxb. Chính trị quốc gia Sự thật.

9. Moravec, H. 1988. *Mind children: The future of robot and human intelligence*. Harvard University Press.

10. Driver, J. 2012. *Consequentialism*. Routledge.

11. Rawling, P. 2023. *Deontology*. Cambridge University Press.

12. Constantinescu, M., Vică, C., Uszkai, R., & al. 2022. Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors. *Philosophy & Technology, 35*, 35. (https://doi.org/10.1007/s13347-022-00529-z).

13. Clark, A., & Mahtani, M. 2024. Google AI chatbot responds with a threatening message: "Human… Please die." CBS News. (*https://www.cbsnews.com/news/google-ai-chatbot-threatening-message-human-please-die/*).