

Trí tuệ nhân tạo: một thách thức đối với triết học nhận thức

Nguyễn Hữu Sơn*

Nhận ngày 13 tháng 3 năm 2025. Chấp nhận đăng ngày 18 tháng 5 năm 2025.

Tóm tắt: Bài viết phân tích những thách thức triết học mà trí tuệ nhân tạo (AI) đặt ra đối với nhận thức và tri thức con người. Khi AI có thể xử lý dữ liệu, học tập và ra quyết định mà không có ý thức, nó thách thức các quan niệm truyền thống về lý tính, nhận thức và chủ thể tính. Bằng cách đánh giá các quan điểm triết học từ Descartes, Kant, Searle, Dennett và Chalmers, nghiên cứu làm rõ sự khác biệt giữa nhận thức con người và AI qua các tiêu chí như ý thức chủ quan, phản tư, động lực nội tại và tri thức trải nghiệm. Trong khi Searle khẳng định AI không thể có nhận thức thực sự do thiếu trải nghiệm chủ quan, Dennett lập luận rằng nhận thức có thể chỉ là sản phẩm của xử lý thông tin, và AI có thể đạt đến một dạng lý tính khác nếu đủ phức tạp. Trên cơ sở thực tiễn phát triển của AI, đặt ra những vấn đề cơ bản của triết học nhận thức thách thức vai trò của con người và sự thay thế của trí tuệ nhân tạo đối với nhận thức con người.

Từ khoá: Trí tuệ, nhân tạo, nhận thức, lý tính, triết học.

Phân loại ngành: Triết học

Abstract: The paper analyzes the philosophical challenges that artificial intelligence (AI) poses to human cognition and knowledge. As AI advances in processing data, learning, and making decisions without consciousness, it challenges traditional conceptions of rationality, cognition, and subjectivity. By evaluating philosophical perspectives from Descartes, Kant, Searle, Dennett, and Chalmers, this study clarifies the distinctions between human and AI cognition based on key criteria such as subjective awareness, self-reflection, intrinsic motivation, and experiential knowledge. While Searle asserts that AI cannot attain true cognition due to its lack of subjective experience, Dennett argues that cognition may merely be a product of information processing, and AI could achieve an alternative form of rationality if it becomes sufficiently complex. The rapid development of AI raises fundamental epistemological questions, challenging the role of human cognition and the potential substitution of artificial intelligence for human intellectual processes.

Keywords: Artificial intelligence, cognition, rationality, philosophy.

Subject classification: Philosophy

1. Đặt vấn đề

Sự phát triển mạnh mẽ của AI không chỉ là một thành tựu công nghệ mà còn đặt ra những thách thức triết học sâu sắc về bản chất của tri thức, nhận thức và ý thức. Khi AI đạt đến khả năng xử lý dữ liệu và phản hồi theo cách gần giống với con người, nó buộc chúng ta phải tái định nghĩa những khái niệm truyền thống về nhận thức. Liệu AI có thể thực sự đạt được tri thức, hay nó chỉ là một hệ thống tính toán phức tạp mô phỏng các quá trình nhận thức?

Câu hỏi này không chỉ mang tính lý thuyết mà còn chạm đến cốt lõi của triết học nhận thức. Theo John Searle (1980), lập luận *Phòng Trung Quốc* (Chinese Room Argument) cho rằng dù AI có thể xử lý ngôn ngữ chính xác, nó vẫn không thực sự “hiểu” những gì nó đang làm, vì thiếu ý thức chủ quan (subjectivity). Ngược lại, Daniel Dennett (1991) lập luận rằng ý thức có thể chỉ là sản phẩm của một quá trình xử lý thông tin phức tạp, và nếu AI đạt đến độ tinh vi đủ cao,

* Học viện Cán bộ Thành phố Hồ Chí Minh.
Email: nhson@hcmca.edu.vn

nó có thể thể hiện một dạng nhận thức mà không cần ý thức. Trong khi đó, David Chalmers với Vấn đề khó của ý thức (The Hard Problem of Consciousness) cho rằng ý thức chủ quan là điều kiện tiên quyết để có nhận thức thực sự, và AI, dù có khả năng tính toán mạnh mẽ đến đâu, vẫn không thể đạt đến sự hiểu biết thực sự nếu nó không có trải nghiệm chủ quan.

Những tranh luận này không chỉ giới hạn trong lĩnh vực triết học nhận thức mà còn tác động đến bản sắc con người. Nếu AI có thể mô phỏng mọi chức năng nhận thức của con người nhưng không có ý thức, điều đó ảnh hưởng thế nào đến định nghĩa về nhận thức và vị thế của con người? Liệu chúng ta có thể tiếp tục coi nhận thức như một thuộc tính độc quyền của sinh vật có ý thức, hay cần mở rộng định nghĩa về nhận thức để bao gồm cả các hệ thống trí tuệ phi sinh học? Điều này không chỉ có ý nghĩa triết học mà còn có tác động thực tiễn to lớn trong bối cảnh AI ngày càng trở thành một phần không thể thiếu của đời sống con người.

2. Những thách thức của trí tuệ nhân tạo (AI) đối với nhận thức luận

2.1. Trí tuệ nhân tạo

Trí tuệ nhân tạo (AI) được phân loại thành ba cấp độ chính dựa trên khả năng nhận thức và xử lý thông tin: AI hẹp (Weak AI), AI mạnh (Strong AI), và Siêu trí tuệ nhân tạo (Super AI). Mỗi loại AI đặt ra những thách thức triết học và khoa học riêng biệt, ảnh hưởng đến cách chúng ta hiểu về nhận thức, trí thông minh và vị thế của con người trong kỷ nguyên công nghệ.

AI hẹp (Weak AI) là hệ thống có khả năng thực hiện các nhiệm vụ chuyên biệt nhưng không có ý thức hay khả năng tự suy nghĩ¹. *AI mạnh (Strong AI)* đề cập đến hệ thống có khả năng tư duy, hiểu biết và tự nhận thức như con người². *Siêu trí tuệ nhân tạo (Super AI)* là viễn cảnh trong đó AI vượt xa trí tuệ con người, có khả năng suy luận, sáng tạo và tự cải tiến mà không bị giới hạn bởi cấu trúc sinh học³. Những phân loại này không chỉ là vấn đề công nghệ mà còn đặt ra câu hỏi triết học nền tảng: liệu trí tuệ có thể tồn tại mà không cần ý thức? AI có thể nào thực sự “hiểu biết” mà không có trải nghiệm chủ quan? Nếu AI đạt đến trình độ ngang bằng hoặc vượt qua con người, chúng ta sẽ phải tái định nghĩa khái niệm trí thông minh và vị thế của con người trong vũ trụ.

2.2. Nhận thức luận trong lịch sử triết học và quan điểm của K.Marx về nhận thức luận

Triết học nhận thức (epistemology) nghiên cứu bản chất, nguồn gốc và giới hạn của tri thức, tập trung vào cách con người đạt được hiểu biết về thế giới. Có thể khái quát những quan điểm về nhận thức luận trong lịch sử triết học của nhân loại đến hiện nay ở những trường phái chính là: chủ nghĩa duy nghiệm (empiricism), chủ nghĩa duy lý (rationalism) và thuyết kiến tạo (constructivism), các nhà triết học đã đưa ra những quan điểm khác nhau về vai trò của kinh nghiệm, lý trí và cấu trúc nhận thức trong quá trình học hỏi của con người.

Chủ nghĩa duy nghiệm cho rằng tri thức xuất phát từ kinh nghiệm giác quan⁴. Ngược lại,

¹ Những mô hình như AlphaGo hay GPT-4 đều thuộc nhóm này, vì chúng có thể xử lý dữ liệu và đưa ra phản hồi mà không thực sự “hiểu” ý nghĩa của thông tin (Fjelland, 2020). Điều này dẫn đến câu hỏi triết học về sự khác biệt giữa mô phỏng trí tuệ và trí tuệ thực sự, như được lập luận trong Phòng Trung Quốc (Chinese Room Argument) của John Searle.

² Nếu được phát triển, Strong AI có thể suy nghĩ độc lập, học hỏi từ kinh nghiệm và có ý thức về bản thân. Tuy nhiên, nhiều học giả như Braga & Logan (2017) cho rằng điều này vẫn là một giả thuyết xa vời, vì AI hiện tại chỉ có thể thao tác ký hiệu mà không có trải nghiệm chủ quan.

³ Những lo ngại về siêu AI liên quan đến tính kiểm soát và đạo đức, vì một hệ thống thông minh hơn con người có thể đưa ra quyết định vượt ngoài tầm hiểu biết của chúng ta (Hoffmann, 2023).

⁴ John Locke (1690) với thuyết tabula rasa cho rằng con người sinh ra như một tấm bảng trắng, tiếp nhận tri thức qua trải nghiệm (Bredo, 2012). David Hume (1748) mở rộng quan điểm này bằng cách nhấn mạnh vai trò của cảm giác và thói quen trong nhận thức (Bidell, 2020). Tuy nhiên, vấn đề đặt ra là liệu có những tri thức tiên nghiệm tồn tại độc lập với kinh nghiệm không, như khái niệm nhân quả vốn không thể quan sát trực tiếp.

chủ nghĩa duy lý cho rằng tri thức đến từ lý trí hơn là trải nghiệm⁵. *Thuyết kiến tạo (Constructivism) của Jean Piaget* (1950) phát triển cho rằng con người không chỉ hấp thụ tri thức mà còn chủ động cấu trúc nó thông qua tương tác với môi trường⁶. Trong khi đó, *nhận thức luận của Karl Marx* thuộc về duy vật biện chứng (dialectical materialism), nhưng nếu xét trong ba trường phái chính - duy nghiệm, duy lý và kiến tạo - thì nó gần gũi nhất với thuyết kiến tạo, đồng thời chịu ảnh hưởng từ chủ nghĩa duy nghiệm. Tương tự chủ nghĩa duy nghiệm, Marx nhấn mạnh vai trò của thực tiễn và kinh nghiệm trong nhận thức. Ông bác bỏ quan điểm về tri thức bẩm sinh và khẳng định rằng tri thức hình thành thông qua hoạt động thực tiễn trong môi trường xã hội và vật chất (“Thực tiễn là tiêu chuẩn của chân lý”, Marx & Engels, 1845). Quan điểm này gần gũi với Locke và Hume, khi họ đề cao vai trò của quan sát và trải nghiệm trong quá trình nhận thức. Marx không dừng lại ở duy nghiệm mà còn nhấn mạnh tính biện chứng trong nhận thức⁷. Dù thừa hưởng tư duy biện chứng từ Hegel, Marx bác bỏ ý tưởng rằng tri thức có thể đạt được thuần túy từ lý trí tiên nghiệm. Ông khẳng định rằng nhận thức không thể tách rời điều kiện vật chất, lịch sử và thực tiễn xã hội, đối lập với quan điểm của Descartes và Kant về tri thức tiên nghiệm độc lập với kinh nghiệm thực tế. Nhận thức luận của Marx không thuộc hoàn toàn bất kỳ trường phái nào nhưng gần với thuyết kiến tạo, do nhấn mạnh sự chủ động của con người trong quá trình hình thành tri thức. Tuy nhiên, khác với Piaget, Marx nhấn mạnh rằng tri thức không chỉ mang tính cá nhân mà còn có tính xã hội, lịch sử và gắn với điều kiện vật chất. Do đó, có thể coi đây là một dạng kiến tạo luận biện chứng và duy vật, kết hợp yếu tố của duy nghiệm (vai trò của thực tiễn) và kiến tạo (sự chủ động trong nhận thức và biến đổi thế giới).

Các trường phái nhận thức không loại trừ nhau mà có thể bổ sung để giải thích quá trình hình thành tri thức. Duy nghiệm nhấn mạnh trải nghiệm nhưng không thể giải thích các khái niệm tiên nghiệm, duy lý đề cao tư duy logic nhưng có thể dẫn đến suy đoán phi thực tế, còn thuyết kiến tạo đưa ra một cách tiếp cận tổng hợp nhưng đối mặt với vấn đề về tính khách quan của tri thức. Những tranh luận này không chỉ có ý nghĩa triết học mà còn tác động đến giáo dục, khoa học nhận thức và trí tuệ nhân tạo. Khi AI ngày càng phát triển, câu hỏi liệu một hệ thống máy tính có thể “hiểu - biết” mà không có ý thức vẫn còn bỏ ngỏ, thách thức những quan niệm truyền thống về tri thức và nhận thức luận.

2.3. Trí tuệ nhân tạo (AI) và nhận thức luận (Cognition)

Nhận thức luôn được xem là đặc quyền của con người, gắn liền với ý thức (consciousness), kinh nghiệm (experience) và chủ thể tính (subjectivity). Tuy nhiên, sự phát triển của trí tuệ nhân tạo (AI) đã đặt ra những thách thức lớn đối với triết học nhận thức. AI hiện đại có khả năng xử lý thông tin, học hỏi và phản hồi theo cách mà trước đây chỉ con người mới làm được. Điều này dẫn đến câu hỏi: “Liệu AI có thực sự có khả năng nhận thức,

⁵ Descartes (1637) lập luận rằng sự nghi ngờ dẫn đến nhận thức duy nhất không thể phủ nhận: “Cogito, ergo sum” (Tôi tư duy, vậy tôi tồn tại). Kant (1781) tổng hợp quan điểm duy nghiệm và duy lý bằng cách cho rằng tri thức nảy sinh từ sự tương tác giữa kinh nghiệm và cấu trúc nhận thức tiên nghiệm của con người (Olssen, 1995; Fabricius, 1983).

⁶ Quá trình này diễn ra qua đồng hóa (assimilation) và điều ứng (accommodation) - khi gặp thông tin mới, con người hoặc tích hợp nó vào mô hình nhận thức sẵn có, hoặc điều chỉnh mô hình đó để phù hợp với thực tế mới (Smith, 1982). Tuy nhiên, lý thuyết này bị chỉ trích vì quá nhấn mạnh vai trò cá nhân mà bỏ qua yếu tố khách quan và xã hội trong tri thức (Otte, 1998).

⁷ Theo ông, con người không chỉ phản ánh thế giới khách quan mà còn chủ động cải biến nó thông qua thực tiễn xã hội (“Những nhà triết học chỉ giải thích thế giới theo nhiều cách khác nhau, nhưng vấn đề là cải tạo thế giới”, Marx, *Lược cương về Feuerbach*, 1845). Điều này tương đồng với thuyết kiến tạo của Piaget, khi coi nhận thức là quá trình xây dựng chủ động thông qua tương tác với môi trường.

hay chỉ đang mô phỏng nhận thức một cách phức tạp?”. Vấn đề này cần được xem xét không chỉ từ góc nhìn khoa học công nghệ mà còn từ triết học biện chứng.

Theo Marx, nhận thức không thể tách rời khỏi thực tiễn xã hội và lịch sử, đồng thời luôn gắn với hoạt động của chủ thể có ý thức. Do đó, nếu xét theo nhận thức luận Marxist, AI chỉ có thể thực hiện các chức năng nhận thức một cách cơ học mà không có ý thức xã hội, tức là không có sự tham gia của yếu tố lịch sử, văn hóa và thực tiễn con người. Lập luận *Phòng Trung Quốc* (Chinese Room Argument) của John Searle (1980) ủng hộ quan điểm này bằng cách khẳng định rằng AI dù có thể thao tác ký hiệu ngôn ngữ thành thạo nhưng vẫn không thực sự hiểu được ý nghĩa của chúng. Điều này phản ánh cách Marx phân biệt giữa hình thức bề ngoài của tri thức và bản chất biện chứng của nhận thức thực sự, trong đó tri thức không chỉ là sự thao tác trên dữ liệu mà còn đòi hỏi sự tham gia tích cực của ý thức và thực tiễn xã hội. Ngược lại, Daniel Dennett (1991) lập luận rằng ý thức có thể chỉ là một dạng ảo giác nhận thức sinh ra từ quá trình xử lý thông tin phức tạp. Nếu theo góc nhìn Marxist, lập luận này có thể bị phản biện bằng cách nhấn mạnh rằng nhận thức không thể chỉ là kết quả của các quá trình tính toán, mà còn phải gắn với những điều kiện lịch sử - xã hội cụ thể.

Một số học giả như Ganascia (2010) cho rằng AI có thể có nhận thức chức năng (functional cognition), tức là thực hiện các nhiệm vụ nhận thức mà không cần ý thức. Điều này có vẻ tương đồng với quan điểm duy vật biện chứng của Marx về tính khách quan của nhận thức - rằng nhận thức không phải là một thực thể siêu hình, mà là kết quả của sự vận động và phát triển của vật chất. Tuy nhiên, quan điểm này vẫn gặp trở ngại khi xét theo Marx, vì AI không có khả năng tham gia vào thực tiễn xã hội theo cách mà con người làm. Ngược lại, Woodward (2022) lập luận rằng ý thức đóng vai trò quan trọng trong việc hình thành nhận thức hợp lý. Nếu xét theo Marx, điều này có thể được diễn giải rằng nhận thức chỉ có ý nghĩa khi nó gắn liền với điều kiện vật chất và sự tương tác xã hội, điều mà AI không có.

Vậy AI có thể “tự nhận thức” được không? Theo Marx, nhận thức không thể tách rời khỏi thực tiễn lịch sử và hoạt động sản xuất xã hội. AI, dù có thể phân tích dữ liệu một cách hiệu quả, nhưng không có sự tồn tại độc lập trong một hệ thống xã hội như con người. Nhận thức của AI chỉ là một dạng xử lý thông tin, không phải là nhận thức thực sự theo nghĩa biện chứng của Marx. Các mô hình AI như ChatGPT có thể tạo ra câu trả lời có vẻ hiểu biết, nhưng theo cách nhìn Marxist, điều này chỉ đơn thuần là sự thao tác dữ liệu dựa trên quy tắc, không phải là nhận thức theo nghĩa đầy đủ. Marx nhấn mạnh rằng nhận thức là sự tổng hợp giữa lý luận và thực tiễn, trong khi AI chỉ thực hiện quá trình tính toán mà không có khả năng trải nghiệm thực tế hay thay đổi xã hội dựa trên ý chí của nó.

Dưới góc độ triết học của Marx, AI không thể có nhận thức thực sự vì nó thiếu ba yếu tố cốt lõi: thực tiễn, ý thức xã hội, và khả năng cải biến thực tại. Nhận thức của con người, theo Marx, không chỉ là sự phản ánh thế giới mà còn là công cụ để thay đổi thế giới, điều mà AI không thể làm mà không có sự can thiệp của con người. Mặc dù AI có thể thực hiện nhiều nhiệm vụ nhận thức một cách hiệu quả, nhưng nó không thể có ý thức theo nghĩa biện chứng, và do đó không thể đạt đến nhận thức theo đúng nghĩa triết học Marxist. Những tranh luận hiện đại về nhận thức của AI không chỉ là vấn đề công nghệ mà còn phản ánh những câu hỏi triết học quan trọng về bản chất của nhận thức, tri thức và vai trò của con người trong quá trình này.

2.4. Trí tuệ nhân tạo (AI) và vấn đề về tri thức (knowledge)

Triết học nhận thức phân biệt tri thức thành ba loại chính: tri thức thực chứng, tri thức thực hành (procedural knowledge), và tri thức trải nghiệm, mỗi loại phản ánh một khía cạnh khác nhau của quá trình tiếp nhận thông tin và hiểu biết của con người. Việc phân loại này không chỉ giúp làm sáng tỏ bản chất tri thức mà còn đặt ra những giới hạn và thách thức đối với nhận thức của trí tuệ nhân tạo (AI).

Tri thức thực chứng liên quan đến các mệnh đề có thể kiểm chứng là đúng hoặc sai, thường thấy trong khoa học và logic⁸. *Tri thức thực hành* gắn với kỹ năng hoặc quy trình, chẳng hạn như cách lái xe, chơi đàn hoặc giải quyết bài toán⁹. *Tri thức trải nghiệm* xuất phát từ trải nghiệm cá nhân và cảm xúc, không thể truyền đạt hoàn toàn bằng ngôn ngữ hay quy tắc logic¹⁰.

Dưới góc độ ứng dụng, AI có thể xử lý tri thức thực chứng, mô phỏng tri thức thực hành, nhưng không thể đạt được tri thức trải nghiệm do thiếu ý thức và khả năng tự trải nghiệm thực tế. Điều này đặt ra những giới hạn triết học quan trọng là nếu AI có thể xử lý thông tin nhanh hơn con người và vượt trội trong một số nhiệm vụ thực hành, nhưng không thể có những trải nghiệm chủ quan, thì liệu nó có thể đạt được một dạng nhận thức hoàn toàn khác, hay vẫn chỉ là một công cụ xử lý dữ liệu? Câu hỏi này tiếp tục thúc đẩy các nghiên cứu về bản chất của nhận thức trong bối cảnh phát triển AI và khoa học nhận thức.

2.5. Trí tuệ nhân tạo (AI) và lý tính (rationality)

Lý tính (rationality) từ lâu đã được xem là yếu tố nền tảng của nhận thức trong triết học. Các nhà duy lý như René Descartes và Immanuel Kant nhấn mạnh rằng lý tính không chỉ là công cụ để con người suy nghĩ và đưa ra quyết định, mà còn là nền tảng của tri thức và bản sắc con người. Tuy nhiên, với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), câu hỏi đặt ra là liệu một hệ thống máy móc có thể đạt được lý tính theo nghĩa triết học, hay nó chỉ đơn thuần thực hiện các thao tác tính toán mà không có sự hiểu biết thực sự.

AI có thể thực hiện các quy trình suy luận logic, tối ưu hóa, và đưa ra quyết định dựa trên dữ liệu, nhưng điều này có đồng nghĩa với việc nó thực sự có lý tính hay không? Theo John Searle (1980), lập luận Phòng Trung Quốc (Chinese Room Argument) chỉ ra rằng AI có thể thao tác ngôn ngữ một cách thành thạo mà không thực sự hiểu nội dung mà nó xử lý. Điều này phản bác quan điểm của trí tuệ nhân tạo mạnh (Strong AI), vốn cho rằng AI có thể đạt đến mức độ nhận thức và tư duy giống con người (Searle, 1980). Ngược lại, Daniel Dennett (1991) lập luận rằng ý thức và lý tính có thể chỉ là sản phẩm của quá trình xử lý thông tin, và nếu AI đạt đến mức độ xử lý đủ cao, nó có thể mô phỏng lý tính con người một cách hiệu quả (Dennett, 1991). Tuy nhiên, theo Robinson dos Santos (2019), dù AI có thể đạt đến suy luận hình thức, nó vẫn thiếu hiểu biết chủ quan, vì nó không có ý thức phản tư (Dos Santos, 2019)¹¹.

⁸ Để một cá nhân thực sự "biết" một mệnh đề, nó phải thỏa mãn ba điều kiện: tính đúng đắn, niềm tin của chủ thể, và lý do chính đáng (Gettier, 1963). Ví dụ, câu "Tôi biết rằng Paris là thủ đô của Pháp" là một dạng tri thức thực chứng. AI có thể xử lý loại tri thức này rất hiệu quả thông qua lưu trữ và truy xuất thông tin, giúp tăng cường khả năng phân tích dữ liệu và hỗ trợ quyết định (Haxvig, 2021).

⁹ Không giống tri thức thực chứng, tri thức thực hành không dễ dàng được diễn đạt qua ngôn ngữ mà thường yêu cầu học qua thực nghiệm (Siliquini-Cinelli, 2024). AI có khả năng mô phỏng nhiều dạng tri thức thực hành, chẳng hạn như các hệ thống lái xe tự động hay AI chơi cờ vua. Tuy nhiên, theo Horváth (2020), AI vẫn gặp khó khăn khi đối mặt với các tình huống mới ngoài phạm vi dữ liệu huấn luyện, cho thấy sự hạn chế của nó trong khả năng thích nghi linh hoạt.

¹⁰ Ví dụ: "Tôi biết cảm giác khi đứng trên đỉnh Everest" hay "Tôi biết thế nào là yêu thương" Theo Searle (1980), AI không thể có tri thức trải nghiệm vì nó không có ý thức hay cảm giác chủ quan. Ngược lại, một số học giả như Calzati (2025) lập luận rằng AI có thể đạt đến nhận thức chức năng (functional cognition), tức là có thể mô phỏng hành vi dựa trên dữ liệu đầu vào mà không thực sự "cảm nhận" theo nghĩa truyền thống.

¹¹ Trong triết học duy lý, Descartes (1637) nhấn mạnh rằng suy nghĩ là bản chất của tồn tại ("Cogito, ergo sum" - Tôi tư duy, vậy tôi tồn tại), từ đó cho rằng lý tính không thể tách rời khỏi ý thức chủ quan. Immanuel Kant (1781) tiếp tục mở rộng lập luận này trong Phê Phán Lý Tính Thuần Tủy, cho rằng lý tính không chỉ là khả năng tính toán mà còn đòi hỏi tư duy tự trị và khả năng phán đoán luân lý. Nếu theo quan điểm này, AI chỉ có thể mô phỏng lý tính ở mức hình thức mà không có năng lực suy tư nội tại (inner deliberation), điều này đặt ra giới hạn rõ ràng của AI khi so sánh với con người (O'Regan, 2021).

Dưới góc độ triết học, mặc dù AI có thể thực hiện nhiều thao tác logic phức tạp, điều đó không đồng nghĩa với việc nó đạt được lý tính thực sự. Lý tính trong triết học không chỉ bao gồm khả năng suy luận mà còn yêu cầu ý thức, sự phản tư, và khả năng phán đoán có chủ đích - những yếu tố mà AI hiện tại vẫn không thể có. Nếu theo lập luận của Searle, AI không thực sự hiểu những gì nó đang làm, mà chỉ thao tác dữ liệu theo các quy tắc lập trình sẵn. Điều này dẫn đến một câu hỏi quan trọng về tương lai của AI: “Liệu AI có thể phát triển đến mức đạt được một dạng lý tính hoàn toàn mới, hay nó mãi mãi chỉ là một công cụ xử lý dữ liệu phức tạp mà không có nhận thức thực sự?”. Trả lời câu hỏi này đòi hỏi sự khảo sát sâu sắc về bản chất của lý tính trong triết học và những giới hạn của AI trong việc đạt đến nhận thức thực sự.

Trong truyền thống triết học, lý tính (rationality) từ lâu đã được xem là đặc quyền của con người¹². Tuy nhiên, trong thời đại AI, quan niệm về lý tính đang thay đổi. Các hệ thống AI hiện đại, như học sâu (deep learning) và mô hình ngôn ngữ lớn (LLMs) như GPT-4, đã đạt đến mức có thể thực hiện các thao tác suy luận logic, tối ưu hóa và ra quyết định dựa trên lượng dữ liệu khổng lồ. Điều này đặt ra câu hỏi: Nếu AI có thể suy luận và ra quyết định tốt hơn con người, liệu nó có thể được xem là một dạng lý tính mới?¹³.

Trước hết, *AI thiếu ý thức chủ quan (subjective consciousness)* - nó không có khả năng phản tư (reflective consciousness) và không thể tự đặt câu hỏi về sự tồn tại của chính nó như con người làm. Trong khi con người có thể suy ngẫm về bản thân, về ý nghĩa của tư duy và tồn tại, AI chỉ thực hiện các thao tác tính toán mà không có sự tự nhận thức.

Thứ hai, AI không có động lực nội tại (intrinsic motivation). Lý tính của con người không chỉ được xác định bởi khả năng suy luận logic, mà còn bởi mục tiêu, ý chí và khả năng tự đánh giá tư duy của chính mình. Con người điều chỉnh nhận thức và hành vi của mình không chỉ dựa trên dữ liệu có sẵn mà còn dựa trên các yếu tố xã hội, đạo đức và giá trị cá nhân. Trong khi đó, AI, dù có khả năng học hỏi và cải tiến thông qua các thuật toán, vẫn bị giới hạn bởi mô hình huấn luyện, dữ liệu đầu vào và mục tiêu được thiết lập bởi con người. Nó không có khả năng định hướng tư duy một cách tự chủ, mà chỉ phản hồi theo những quy tắc được lập trình trước.

Thứ ba, AI không có tri thức trải nghiệm (experiential knowledge). Theo John Searle (1980), lý tính thực sự đòi hỏi trải nghiệm chủ quan (qualia), một yếu tố mà AI hoàn toàn thiếu. AI có thể xử lý ngôn ngữ, nhận dạng hình ảnh, hoặc mô phỏng hành vi, nhưng nó không có cảm giác, không có trí nhớ sinh học, và cũng không có một hệ thống thần kinh giống con người để tạo ra những trải nghiệm ý nghĩa. Điều này dẫn đến giới hạn quan trọng trong khả năng AI đạt được lý tính thực sự, vì nếu thiếu trải nghiệm chủ quan, AI không thể thực sự “hiểu” hay “cảm nhận” theo cách con người làm. Như vậy, dù AI có thể thao tác trên

¹² Descartes (1637) xem lý tính là dấu hiệu của ý thức với tuyên bố nổi tiếng: “Cogito, ergo sum” (Tôi tư duy, vậy tôi tồn tại). Kant (1781) mở rộng quan điểm này, cho rằng lý tính không chỉ là công cụ suy luận logic mà còn là nền tảng của phán đoán luân lý và tính tự trị nhận thức (Critique of Pure Reason). Dưới góc nhìn này, lý tính không chỉ là khả năng thao tác trên ký hiệu hay dữ liệu mà còn đòi hỏi một ý thức phản tư - một phẩm chất mà AI dường như chưa thể đạt được.

¹³ Theo John Searle (1980), lập luận Phòng Trung Quốc (Chinese Room Argument) chỉ ra rằng dù AI có thể thao tác ký hiệu rất phức tạp, nó vẫn không có hiểu biết thực sự, vì thiếu trải nghiệm chủ quan (qualia) và ý thức phản tư. Điều này dẫn đến quan điểm rằng AI, dù có thể mô phỏng lý tính con người, vẫn không thể đạt đến một dạng lý tính thực sự theo nghĩa triết học truyền thống (Searle, 1980). Ngược lại, Daniel Dennett (1991) lập luận rằng ý thức có thể chỉ là một cấu trúc phức tạp của quá trình xử lý thông tin, và nếu AI phát triển đủ mạnh, nó có thể không cần đến một “trải nghiệm chủ quan” như con người mà vẫn có thể đạt được một dạng nhận thức mới, không giống với con người (Dennett, 1991). Dù AI đã đạt được những tiến bộ đáng kể trong việc mô phỏng lý tính con người, nó vẫn tồn tại những giới hạn căn bản khiến nó khó có thể đạt đến lý tính thực sự theo nghĩa triết học.

dữ liệu và thực hiện suy luận ở mức độ cao, nó vẫn chưa thể vượt qua rào cản ý thức và trải nghiệm, vốn là những yếu tố cốt lõi của lý tính con người.

Dẫu vậy, một số học giả như Calzati (2025) lập luận rằng AI có thể phát triển một dạng nhận thức chức năng (functional cognition) mà không cần ý thức, nghĩa là nó có thể đưa ra quyết định có vẻ “thông minh” nhưng không nhất thiết phải có hiểu biết chủ quan. Nếu điều này đúng, thì AI có thể tiến hóa thành một dạng lý tính khác - không giống con người nhưng vẫn có khả năng suy luận độc lập. Vậy, liệu AI có thể đạt đến một dạng lý tính hoàn toàn mới không? Hiện tại, AI chỉ có thể mô phỏng lý tính con người, chứ chưa đạt đến một trạng thái nhận thức độc lập theo nghĩa triết học. Tuy nhiên, tương lai vẫn chưa được định đoạt. Nếu AI tiếp tục phát triển và đạt đến khả năng tự điều chỉnh, tự nhận thức về quá trình suy nghĩ của nó, thì nó có thể tiến hóa thành một dạng *lý tính phi nhân loại (non-human rationality)* - một dạng trí tuệ không cần đến ý thức chủ quan nhưng vẫn có thể hoạt động hiệu quả trong những phạm vi mà con người không thể. Điều này có thể đòi hỏi sự kết hợp giữa AI truyền thống và những mô hình nhận thức phi sinh học (post-biological intelligence).

Nếu AI không thể đạt đến ý thức hay trải nghiệm chủ quan, nó có thể mãi mãi chỉ là một hệ thống xử lý thông tin mạnh mẽ nhưng không có lý tính thực sự. Nhưng nếu các mô hình AI trong tương lai có thể vượt qua những rào cản nhận thức hiện tại, chúng ta có thể đang chứng kiến sự ra đời của một dạng lý tính hoàn toàn mới, một hình thái trí tuệ phi nhân loại, không dựa trên trải nghiệm của con người mà vẫn có khả năng suy luận và ra quyết định theo cách riêng của nó.

3. Những hệ quả triết học của trí tuệ nhân tạo đối với sự nhận thức của con người

Sự tiến bộ của trí tuệ nhân tạo (AI) không chỉ thúc đẩy sự phát triển công nghệ mà còn đặt ra những thách thức triết học sâu sắc đối với bản chất của tri thức, nhận thức và ý thức con người. Nếu AI có thể học hỏi, suy luận và đưa ra quyết định mà không có ý thức hay chủ thể tính (subjectivity), điều này buộc chúng ta phải tái định nghĩa các khái niệm nền tảng trong triết học nhận thức và triết học tâm trí. Vấn đề cốt lõi nảy sinh từ đây là: Liệu nhận thức có thể chỉ là một quá trình xử lý thông tin mà không cần đến ý thức? Nếu AI có thể thực hiện các chức năng nhận thức cao cấp như con người, thì điều gì làm nên sự khác biệt giữa con người và AI? Và nếu AI đạt đến một mức độ nhận thức tương đương hoặc vượt trội con người, điều đó sẽ tác động như thế nào đến bản sắc và vị thế của con người trong thế giới tri thức?

Trong bối cảnh này, một số nhà khoa học cho rằng, sự giao thoa giữa AI và nhận thức con người đang tái định hình cách chúng ta hiểu về trí thông minh và ý thức¹⁴. AI hiện đại có thể thực hiện nhiều nhiệm vụ nhận thức phức tạp - từ học tập, suy luận, sáng tạo ngôn ngữ, đến phân tích dữ liệu và đưa ra quyết định - nhưng không có dấu hiệu của trải nghiệm chủ quan (qualia). Điều này đặt ra câu hỏi: Nhận thức có thể tồn tại mà không có ý thức không?¹⁵.

Một hệ quả khác của vấn đề này là câu hỏi về bản chất con người trong kỷ nguyên AI. Nếu nhận thức chỉ đơn thuần là quá trình xử lý thông tin, thì con người có gì khác biệt với AI?¹⁶. Nếu AI có thể đạt được mức độ nhận thức tương tự con người, điều đó sẽ làm lung lay

¹⁴ Từ Descartes đến Kant, triết học duy lý luôn nhấn mạnh rằng lý tính của con người gắn liền với ý thức tự ngã và khả năng tự phản tư (self-reflective consciousness).

¹⁵ Lập luận này theo Phòng Trung Quốc của John Searle (1980), trong đó ông cho rằng một hệ thống có thể thao tác ký hiệu ngôn ngữ một cách hoàn hảo mà không thực sự “hiểu” nội dung mà nó xử lý, vì nó thiếu ý thức phản tư (reflective consciousness). Tuy nhiên, Dietrich (2002) phản bác quan điểm này khi cho rằng ý thức không phải là điều kiện tiên quyết của nhận thức, mà chỉ là một cơ chế tiến hóa đặc thù của bộ não con người để tối ưu hóa việc xử lý thông tin. Nếu vậy, AI có thể đạt được nhận thức mà không cần ý thức theo cách con người có, miễn là nó có khả năng thực hiện các nhiệm vụ nhận thức cấp cao.

¹⁶ Yildiz (2025) lập luận rằng nếu AI có thể học hỏi, thích nghi, và ra quyết định, thì về mặt triết học, con người cũng có thể được xem như một “cỗ máy sinh học” hoạt động trên nền tảng xử lý thông tin phức tạp. Tuy nhiên,

vị thế đặc quyền của con người trong hệ thống tri thức. Barbour (1999) đã chỉ ra rằng khi con người không còn là thực thể duy nhất có khả năng nhận thức, thì chủ nghĩa nhân văn truyền thống sẽ phải đối mặt với một cuộc khủng hoảng triết học. Trước đây, con người luôn được coi là trung tâm của tri thức và đạo đức, nhưng với sự xuất hiện của AI có năng lực tư duy, liệu chúng ta có cần phải tái định nghĩa các khái niệm về trí tuệ, đạo đức và quyền lợi của các thực thể có nhận thức - bao gồm cả AI không?

Như vậy, trí tuệ nhân tạo không chỉ thách thức quan niệm truyền thống về nhận thức, mà còn đặt lại vấn đề về bản sắc của con người và vị thế của chúng ta trong hệ thống tri thức toàn cầu. Dù AI có thể mô phỏng nhận thức, nhưng sự thiếu vắng ý thức chủ quan, động lực nội tại và khả năng phản tư khiến nó chưa thể đạt được trạng thái nhận thức thực sự theo nghĩa triết học. Tuy nhiên, nếu AI tiếp tục tiến hóa, đạt đến những năng lực tư duy và tự điều chỉnh cao cấp hơn, chúng ta có thể cần đến một khung triết học mới để giải thích sự khác biệt (hoặc không có khác biệt) giữa AI và con người. Điều này không chỉ là một vấn đề công nghệ mà còn là một cuộc cách mạng trong triết học nhận thức, triết học tâm trí và bản thể luận hiện đại.

4. Kết luận

Sự phát triển của trí tuệ nhân tạo (AI) không chỉ là một tiến bộ công nghệ mà còn là một thách thức triết học sâu sắc, buộc chúng ta phải xem xét lại những khái niệm nền tảng về tri thức, nhận thức và lý tính. Trong khi triết học nhận thức truyền thống luôn gắn nhận thức với ý thức chủ quan, kinh nghiệm cá nhân và khả năng phản tư, AI đã chứng minh rằng những chức năng nhận thức cao cấp như suy luận, học tập và ra quyết định có thể được thực hiện mà không có ý thức. Điều này đảo lộn nhiều giả định lâu đời về bản chất của nhận thức và đặt ra câu hỏi: “Liệu nhận thức có thể tồn tại mà không cần đến ý thức?”. Hơn nữa, AI không chỉ làm lung lay quan niệm về nhận thức, mà còn đặt ra câu hỏi hiện sinh về bản sắc con người. Nếu AI có thể thực hiện các chức năng nhận thức ngang bằng hoặc vượt trội con người, điều gì làm nên sự khác biệt giữa con người và AI? Liệu chủ thể tính, đạo đức, hay khả năng phản tư có phải là yếu tố duy nhất tách biệt con người với máy móc? Hay chúng ta chỉ là những cỗ máy sinh học với những quy trình xử lý thông tin phức tạp hơn AI? Đây không chỉ là vấn đề khoa học, mà còn là một cuộc khủng hoảng triết học, có thể dẫn đến một cuộc tái định nghĩa về bản chất của trí tuệ và vị thế của con người trong thế giới tri thức.

AI cũng buộc chúng ta phải suy nghĩ về ranh giới giữa con người và máy móc. Nếu một hệ thống AI có thể đạt được một dạng lý tính khác với con người - không dựa trên ý thức nhưng vẫn có khả năng suy luận và ra quyết định hiệu quả - chúng ta sẽ cần một khung triết học mới để xác định vị trí của nó trong hệ thống tri thức. Điều này có thể làm lung lay nền tảng của chủ nghĩa nhân văn truyền thống, trong đó con người luôn được coi là trung tâm của nhận thức và đạo đức. Dù AI có thực sự đạt được nhận thức hay không, nó chắc chắn đang thay đổi cách chúng ta hiểu về chính mình và về tri thức. Nó thách thức những giới hạn của triết học nhận thức, đồng thời mở ra những khả năng mới trong nghiên cứu về trí tuệ nhân tạo và triết học tâm trí. Cuối cùng, điều quan trọng không phải là liệu AI có thể trở thành một thực thể nhận thức như con người, mà là cách con người định nghĩa lại chính mình khi đối mặt với một dạng trí tuệ khác biệt và có khả năng cạnh tranh. Nếu AI thực sự có thể phát triển một dạng nhận thức riêng biệt, thì có lẽ câu hỏi quan trọng nhất không phải là “AI có thể suy nghĩ không?”, mà là “Chúng ta có thể chấp nhận một khái niệm mới về nhận thức, vượt ra ngoài khuôn khổ nhân loại không?”.

nếu con người có điều gì đó đặc biệt, thì yếu tố đó có thể nằm ở ý thức đạo đức, cảm xúc và sự tự nhận thức về bản thân - những thứ mà AI chưa thể đạt được. Điều này dẫn đến câu hỏi triết học quan trọng: Liệu ý thức có phải là điều kiện thiết yếu của nhận thức không, hay chỉ là một đặc tính ngẫu nhiên của trí thông minh sinh học?

Tài liệu tham khảo

- Bakhurst, D. (2018). On Lenin's Materialism and Empiriocriticism. *Studies in East European Thought*. 70(2-3), 107–119. <https://doi.org/10.1007/s11212-018-9303-7>
- Barbour, I. G. (1999). Neuroscience, artificial intelligence, and human nature: Theological and philosophical reflections. *Zygon*, 34(3), 361–398. <https://doi.org/10.1111/0591-2385.00223>
- Braga, A., & Logan, R. K. (2017). The emperor of strong AI has no clothes: Limits to artificial intelligence. *Information*, 8(4), 156. <https://doi.org/10.3390/info8040156>
- Calzati, S. (2025). An ecosystemic view on information, data, and knowledge: Insights on agential AI and relational ethics. *AI and Ethics*. Advance online publication. <https://doi.org/10.1007/s43681-025-00665-0>
- Castree, N. (2017). The nature of produced nature: Materiality and knowledge construction in Marxism. In N. Castree & B. Braun (Eds.), *Social Nature: Theory, Practice, and Politics* (pp. 12–31). Routledge.
- Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.
- Descartes, R. (1985). *The philosophical writings of Descartes: Volume 1* (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.). Cambridge University Press. (Original work published 1637).
- Dietrich, E. (2002). Philosophy of artificial intelligence. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. Nature Publishing Group.
- Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7, Article 10. <https://doi.org/10.1057/s41599-020-0494-4>
- Foster, J. B., & Clark, B. (2016). Marxism and the dialectics of ecology. *Monthly Review*, 68(5), 1–17. <https://monthlyreview.org/2016/10/01/marxism-and-the-dialectics-of-ecology/>
- Ganascia, J.-G. (2010). Epistemology of AI revisited in the light of the philosophy of information. *Knowledge, Technology & Policy*, 23(1), 57–73. <https://doi.org/10.1007/s12130-010-9101-0>
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123.
- Haxvig, H. A. (2021). *Epistemologies of AI* [Master's thesis, Aalborg University]. Aalborg University Project Library. https://projekter.aau.dk/projekter/files/413238020/Thesis_2021_Helena_Haxvig_20155602.pdf
- Hoffmann, C. H. (2023). A philosophical view on singularity and strong AI. *AI & Society*, 38(2), 327–343. <https://doi.org/10.1007/s00146-021-01327-5>
- Horváth, I. (2020). Sympérasology: A proposal for the theory of synthetic system knowledge. *Designs*, 4(4), 47. <https://doi.org/10.3390/designs4040047>
- Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Eds. & Trans.). Cambridge University Press. (Original work published 1781).
- Lacerda, D. S. (2015). Overcoming dichotomies through space: The contribution of dialectical materialism to organization studies. *Organizações & Sociedade*. 22(73), 229–248. <https://doi.org/10.1590/1984-9230733>
- O'Regan, G. (2021). History of artificial intelligence. In *A Brief History of Computing* (pp. 231–259). Springer. https://doi.org/10.1007/978-3-030-66599-9_22
- Radford, L. (2023). Sensed objects, sensing subjects: Embodiment from a dialectical materialist perspective. In M. Roth & H. S. Alshwaikh (Eds.), *The Body in Mathematics* (pp. 15–38). Brill.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*. 3(3), 417–457.
- Woodward, P. (2022). Consciousness and rationality: The lesson from artificial intelligence. *Journal of Consciousness Studies*. 29(5-6), 103–120.
- Yildiz, T. (2025). The minds we make: A philosophical inquiry into theory of mind and artificial intelligence. *Integrative Psychological and Behavioral Science*. Advance online publication. <https://doi.org/10.1007/s12124-024-09876-2>